Analyse en corpus de chaînes de coréférence : la coréférence non-stricte à l'épreuve de la linguistique outillée

Marine Delaborde

Thèse de doctorat en sciences du langage

préparée à

École doctorale N°622 : Sciences du langage

Unité de recherche : Lattice (UMR 8094)

dans le but d'obtenir le diplôme de doctorat délivré par

Université Sorbonne Nouvelle

Analyse en corpus de chaînes de coréférence

La coréférence non-stricte à l'épreuve de la linguistique outillée

---

pour une présentation et une soutenance publique par

Marine Delaborde

en décembre 2020

Sous la direction de Frédéric Landragin

Composition du jury :
Laure Gardelle, Professeur des universités (Université Grenoble Alpes), Rapportrice,
Emmanuel Schang, Maître de conférences HDR (Université d'Orléans), Rapporteur,
Jeanne-Marie Debaisieux, Professeur des universités (Université Sorbonne Nouvelle), Examinatrice,
Guy Achard-Bayle, Professeur des universités (Université de Lorraine), Examinateur,
Frédéric Landragin, Directeur de recherche (Lattice), Directeur.

Analyse en corpus de chaînes de coréférence

Résumé

Une chaîne de coréférence désigne l'ensemble des expressions linguistiques qui réfèrent à la même entité. La relation de coréférence entre les « maillons » d'une chaîne implique que le référent doit être strictement le même pour chaque expression qui la compose. Cependant, il arrive que le référent d'une expression soit difficile à identifier et que la relation de coréférence entre plusieurs expressions ne soit pas stricte de manière certaine. Pour un lecteur, ce manque de précision ne pose pas nécessairement de difficultés. En revanche, lors de l'annotation d'un corpus en coréférences, il est question d'indiquer clairement le référent de chaque expression. Les phénomènes de coréférence non stricte peuvent donc causer des difficultés d'annotation.

Cette thèse a débuté au sein du projet ANR Democrat, avec une tâche d'annotation qui a permis de faire émerger des difficultés d'annotation théoriques et techniques liées à la coréférence non stricte. Nous proposons donc de passer en revue les phénomènes linguistiques impliqués dans la coréférence non stricte, notamment le flou (co)référentiel ainsi que les cas typiques relevés en corpus. Dans un second temps, nous proposons une étude de l'annotation de ces phénomènes dans un sous-corpus de Democrat. Cette étude révèle une grande variabilité d'annotation de ces phénomènes dont nous tirons une classification. Pour éviter les difficultés d'annotation liées à ces phénomènes, nous proposons un cadre plus précis pour l'annotation de la coréférence floue. Cela implique des précisions à ajouter au manuel d'annotation ainsi qu'un schéma d'annotation adapté, prenant en compte la coréférence floue.

Mots clés : référence, coréférence, chaînes de coréférence, flou, annotation, corpus, schéma d'annotation.

Corpus analysis of coreference chains

Abstract

A coreference chain designates the set of linguistic expressions that refer to the same entity. The coreference relation between a chain's elements implies that the referent must be strictly the same for each expression that composes it. However, the referent of an expression is sometimes difficult to identify and the coreference relation between several expressions cannot therefore be strict without any doubt. For a reader, this lack of precision does not necessarily pose difficulties. Nevertheless, the coreference annotation task of a corpus consists in unequivocally identifying the referent of each expression. Non-strict coreference phenomena can therefore generate annotation difficulties.

This thesis began within the ANR Democrat project, with an annotation task that highlighted the emergence of theoretical and technical annotation difficulties related to non-strict coreference. We thus propose to review the linguistic phenomena involved in non-strict coreference, in particular the (co)referential fuzziness as well as the typical cases found in corpora. In a second step, we propose a study of the annotation of these phenomena in a Democrat sub-corpus. This study reveals a great variability in the annotation of these phenomena from which we derive a classification. To avoid the annotation difficulties related to these phenomena, we propose a more precise framework for the annotation of fuzzy coreference. This implies precisions to be added to the annotation manual as well as an adapted annotation scheme, taking into account the fuzzy coreference.

Keywords : reference, coreference, coreference chains, fuzzyness, annotation, corpus, annotation scheme.

Sommaire

Introduction

Problématique

Le langage nous donne la capacité de pouvoir parler de *quelque chose* :

> « La communication linguistique ayant souvent pour objet la réalité extra-linguistique, les locuteurs doivent pouvoir désigner les objets qui la constituent : c'est la *fonction référentielle* du langage » (Ducrot et Todorov 1972).

Le phénomène linguistique de la référence désigne le lien entre une expression référentielle et l'entité qu'elle désigne : son référent. Les expressions référentielles peuvent être, entre autres, des groupes nominaux, des noms propres ou encore des pronoms. La coréférence est le lien qui relie plusieurs expressions référentielles qui désignent le même référent, formant ainsi une chaîne de coréférence dont les expressions référentielles sont les maillons. La coréférence implique une relation d'égalité stricte entre les référents de plusieurs expressions référentielles. En pratique, il arrive parfois que le référent d'une expression soit difficile à identifier de manière précise et que le lien de coréférence entre deux expressions ne soit pas tout à fait strict. Pour un lecteur, cette subtilité — que l'on pourrait qualifier de « flou (co)référentiel » — ne posera pas nécessairement de problème de compréhension du texte. Cependant, pour une tâche d'annotation de texte en coréférence, le problème se pose : faut-il séparer deux chaînes qui coréfèrent de manière floue ou bien les regrouper ? Peut-on trouver une troisième alternative pour prendre en compte ce phénomène ?

Cadre et contexte de l'étude

Ce travail prend source au sein du projet ANR [1] Democrat [2], qui a permis de mettre en lumière des phénomènes de coréférence non stricte qui n'étaient pas pris en compte dans le manuel d'annotation du projet et n'ont donc pas été annotés en tant que tels par les annotateurs du projet. Le point de départ de cette étude relève de la linguistique théorique

1. Agence Nationale pour la Recherche.
2. DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique.

et de la linguistique de corpus afin de définir les notions liées à la problématique et d'observer ces phénomènes en corpus pour analyser leur fonctionnement. Ces observations en corpus permettent ensuite de fournir des recommandations pour le traitement de la coréférence floue en corpus, avec des visées de traitement automatique des langues. Le traitement automatique de la coréférence, comme les système de détection par exemple, impliquent de devoir catégoriser des phénomènes en corpus. Or, cela n'est pas toujours possible avec le phénomène de flou appliqué à la coréférence. Nous proposerons donc un cadre précis permettant de traiter ce phénomène à travers l'adaptation d'un manuel et d'un schéma d'annotation.

Comment prendre en compte les relations de coréférence non stricte ?

Dans un premier temps, il est pertinent de se pencher sur l'annotation de la coréférence. Pour cela, une première étape consiste à définir les notions que cette étape implique, afin d'établir le cadre et le positionnement de ce travail. La seconde étape concerne le travail réalisé au sein de l'ANR Democrat. Un travail d'annotation a été réalisé au cours de ce travail de thèse afin de prendre conscience des enjeux de l'annotation de la coréférence. Ce travail a permis de mettre en évidence des cas particuliers qui peuvent poser des difficultés d'annotation, tant d'un point de vue théorique que technique. Une fois l'annotation réalisée, comment prendre en compte les résultats obtenus ? La plateforme TXM a apporté des solutions et les collaborateurs du projet Democrat commencent à proposer des méthodologies pour l'analyse des chaînes.

Les observations faites dans la première partie de la thèse nous guident vers la prise en compte de relations de coréférence non stricte. La première question qui se pose dans ce contexte est le choix du référent. À ce sujet, la notion de saillance est importante. Elle permet par exemple de choisir un référent dans un cas d'ambiguïté référentielle, lorsque deux options qui ne sont pas reliées sémantiquement sont possibles. La coréférence proche relève aussi de la coréférence non stricte car les référents de plusieurs expressions sont proches sémantiquement, sans être exactement les mêmes, alors qu'ils sont pourtant bien identifiés dans chaque cas. Les référents évolutifs sont des référents qui subissent des modifications. Ces modifications apportent parfois un tel changement qu'il est utile de se poser la question de la création d'un nouveau référent. Ce cas de figure laisse pointer la question d'un non déterminisme référentiel qui est lié à une relation de (co)référence floue. La coréférence floue concerne des expressions qui pourraient être coréférentes car les référents semblent proches, mais de manière floue. Il ne s'agit ni d'ambiguïté, ni de coréférence proche. Ce type de relation apparaît principalement avec la référence à des

groupes pluriels et l'emploi de certains pronoms, principalement « on ». L'étude de ces cas particuliers en corpus nous amènent à penser que si un lecteur ne résout pas toujours précisément ces relations de coréférence au moment de la lecture d'un texte, cette information est bonne à prendre pour l'annotation de ce phénomène en corpus.

Notre dernière partie se consacre justement à la transition entre les observations en corpus et le traitement automatique de la coréférence. Certains travaux ont pris en compte la coréférence non stricte en corpus. Cependant, l'étude de l'annotation des cas typiques liés à la coréférence floue, comme le pronom « on », dans Democrat nous incite à effectuer d'autres recommandations. En ce qui concerne l'annotation de la coréférence, nous proposons des précisions à ajouter à un manuel d'annotation de la coréférence. Nous proposons aussi un schéma d'annotation qui puisse prendre en compte la coréférence floue. Pour finir, nous abordons les conséquences de la prise en compte d'un tel phénomène pour le traitement automatique de la coréférence.

Ce travail de thèse aborde le phénomène de la coréférence floue, dont certains aspects ont déjà été évoqués dans d'autres travaux, avec des terminologies variées. La prise en compte de ce phénomène en corpus constitue un nouveau défi pour lequel nous proposons un cadre pratique et théorique.

Première partie

Annoter les chaînes de coréférence

Afin de comprendre l'annotation de chaînes de coréférence, il est important de poser les bases conceptuelles pour établir un cadre théorique en définissant les phénomènes impliqués tout en précisant la portée du phénomène étudié. Il sera ensuite possible de se pencher sur les différentes manières d'étudier les chaînes de coréférence, qu'elles soient issues de l'étude de l'anaphore, des observations en corpus ou du traitement automatique des langues. Cela permet de positionner le travail de cette thèse en fonction de ces différentes approches en précisant par la même occasion la terminologie adoptée ainsi que les phénomènes (co)référentiels retenus ou non dans ce travail.

L'annotation de ce phénomène peut se faire de différentes manières qui dépendent de la nature ou du domaine dans lequel cette tâche se situe. Cette thèse s'inscrit dans une tâche collective de plus grande ampleur, le projet ANR Democrat qui sera présenté au chapitre 2 et à l'intérieur duquel le rôle de la thèse sera précisé. Avant de commencer l'annotation, plusieurs questions se posent et leur réponse impose certaines contraintes. L'annotation de la coréférence a ses propres caractéristiques et inconvénients ainsi certaines questions linguistiques sont souvent écartées. Un manuel d'annotation doit donc être mis en place dès le départ, comme cela a été le cas dans le projet Democrat. Il sert de point de départ pour la tâche d'annotation et permet de servir de modèle. Il implique toujours de faire des choix, justifiés par le cadre théorique mais aussi par les contraintes des outils utilisés. Différents outils d'annotation ont été utilisés dans le projet mais aussi pour cette thèse. L'exploitation des annotations reste elle aussi un problème ouvert que le projet a eu vocation d'aider à résoudre à travers l'outil TXM compte tenu du manque de méthodologie d'analyse des chaînes de coréférence constaté au début de ce travail [3].

3. Des travaux ont été réalisés depuis sur le sujet, notamment au sein du projet Democrat.

Chapitre 1

Les chaînes de coréférence

Qu'est-ce qu'une chaîne de coréférence et quels sont les phénomènes impliqués ? Comment étudier ce phénomène complexe qui peut porter sur une simple phrase comme s'étendre sur un roman complet ? Différentes approches ont été tentées dans plusieurs domaines, avec des objectifs variés. Ce chapitre est l'occasion de faire un point aussi bien conceptuel que terminologique car des termes différents sont parfois utilisés pour décrire le même phénomène et un même phénomène peut aussi porter des noms différents en fonction des auteurs. Référence, expression référentielle, anaphore, coréférence, chaînes de (co)référence, etc. Autant de notions proches et qui peuvent se recouper pour lesquelles il est nécessaire de faire le point. Cette étape est cruciale et permet le positionnement théorique de cette thèse ainsi que la précision de la terminologie adoptée et de la portée du phénomène de chaîne de coréférence avec les phénomènes retenus ou non.

1.1 Cadre théorique : les notions impliquées

1.1.1 La référence : une fonction fondamentale du langage

La notion de référence est à la base de notre sujet d'étude. Cette question a été largement abordée par de nombreux philosophes et linguistes. Bien qu'ayant d'autres fonctions fondamentales, pour de nombreux philosophes comme Platon et Aristote, le langage a d'abord été considéré comme étant essentiellement référentiel (Hottois 2002). L'un de leurs arguments en faveur de cette idée est que le sens trouve son origine dans la référence dans la mesure où « l'origine du sens linguistique est extralinguistique ». Par la suite, cette idée a été remise en question : pour une grande partie des auteurs du XX$^{\text{ème}}$ siècle, le sens n'est plus lié à la référence.

En linguistique, Saussure et al. (1922) s'attaque à ce sujet controversé en définissant la langue comme un ensemble de signes et le signe comme une entité composée de deux faces complémentaires : le « signifiant », qui est le concept, la représentation mentale du signe, et le « signifié », correspondant à sa forme matérielle. Un an plus tard, Odgen et

Richards (1923) ajoutent la notion de référence à ce concept avec leur idée de triangle sémiotique. Selon eux, le signe serait plutôt une entité ternaire composée d'une représentation physique (*Symbol*), d'un concept (*Thought*) et de la chose à laquelle tout cela réfère (*Referent*). Bien qu'il existe une « stabilité intersubjective » (Kleiber 1999) liée à des capacités perceptives relativement similaires pour chaque être humain, ils défendent l'idée que la communication humaine peut être problématique en raison du fait que les locuteurs interprètent les mots en fonction de leur propre vécu. C'est pourquoi la représentation mentale d'un mot peut varier d'un individu à un autre alors que le référent, la chose dont on parle, sera la même pour tout le monde. Cette différence d'interprétation est donc souvent la cause de malentendus car les mots peuvent prendre une connotation différente en fonction de chaque individu.

Comme le rappelle Charolles (2002, p. 7), les nombreux auteurs qui se sont intéressés au concept de la référence se sont souvent attachés à définir cette notion en la distinguant de la notion de *dénotation*. Lorsqu'un terme, comme une entrée dans le dictionnaire, fait allusion à une classe bien définie satisfaisant un certain nombre de caractéristiques, il s'agit de dénotation. La référence, quant à elle, implique une identification précise d'un référent par un locuteur pour un interlocuteur au moyen d'une expression référentielle. Milner (1976, p. 64) effectue aussi cette distinction en opposant la « référence virtuelle » (dénotation) à la « référence actuelle » (référence), ce qui rend bien compte de la dichotomie entre la théorie et la pratique. Un nom peut dénoter, mais c'est aussi le cas des adjectifs ou encore des verbes. Par exemple, le verbe « manger » dénote toute action consistant à avaler un aliment après l'avoir mâché. Comme le souligne Strawson (1950), il y a du sens qui peut être non référentiel. Le terme « chaussure » peut donc dénoter tout objet réalisé afin de recouvrir et protéger un pied [1], sans référer. En revanche, lorsqu'il est utilisé dans la phrase « J'ai trouvé une chaussure dans les escaliers », il s'agit de référence en raison des caractéristiques soulevées par Charolles (2002, p. 9) : la nature intentionnelle, projective et communicationnelle de cet emploi. En effet, à travers cette expression le locuteur montre sa volonté d'identifier précisément un référent particulier, qui est une entité extralinguistique sur laquelle il peut se mettre d'accord avec un interlocuteur.

Une autre distinction est souvent opérée entre la *dénotation* et la *connotation* (Mill 1889 ; Martinet 1967 ; Barthes 1970 ; Kerbrat-Orecchioni 1983). La dénotation correspond donc au sens littéral d'un terme, celui que l'on peut retrouver dans le dictionnaire. Tandis que la connotation correspond aux informations supplémentaires qui peuvent s'ajouter à ce sens : ce que le terme évoque indépendemment du sens dénotatif.

1. Si l'on se réfère à la définition du dictionnaire : https://www.cnrtl.fr/definition/Chaussure

La référence est donc un acte de langage consistant à désigner précisément et intentionnellement, par le biais d'une expression référentielle, un objet extralinguistique qui existe dans le monde, selon FREGE (1892), ou encore que l'on peut se représenter. Par exemple, il est possible de référer à une licorne (KARTTUNEN 1976) : cet animal n'existe pas matériellement mais il est aisé de se le représenter. Cependant, on peut aussi se demander s'il est possible de référer à une entité qui n'existe pas et que l'on ne peut pas non plus se représenter, comme un cercle carré ou « les nombres impairs multiples de deux » qui, selon CHAROLLES (2002), « posent beaucoup de problèmes ». Pour DUCROT et TODOROV (1972), la réalité à laquelle on réfère ne doit pas nécessairement correspondre à la réalité du monde qui nous entoure. La langue permet plutôt de construire un « univers de discours » imaginaire auquel elle peut se référer.

Pour KARTTUNEN (1976) il y a une distinction à opérer entre les référents, au sens où on l'a employé jusqu'ici, et les « référents de discours » (*discourse referents*). Les référents de discours sont les référents mentionnés, qui font partie de l'univers discursif des interlocuteurs. Il donne ces exemples[2] :

Exemple [1]
a. Bill a une voiture.
b. Elle est noire.
Lauri KARTTUNEN, *A theory of truth and semantic representation*, 1976, page 4.

Dans l'exemple [1], il y a bien un référent de discours pour la voiture. Ce qui n'est pas le cas dans l'exemple suivant :

Exemple [2]
a. Bill n'a pas de voiture.
b. *Elle est noire.
Lauri KARTTUNEN, *A theory of truth and semantic representation*, 1976, page 4.

Dans l'exemple [2], il se base sur l'impossibilité de reprise[3] pour démontrer qu'il n'y a pas de référent de discours correspondant à une voiture dans ce cadre-là. En effet, aucune voiture spécifique n'est introduite dans la phrase (a). Il avance donc que l'acceptabilité d'une phrase comme (b) présuppose l'existence de quelque chose qui n'est pas là. Cela ne dépend donc pas de l'existence ontologique du référent mais de son existence dans le discours.

À la suite des travaux de Karttunen, plusieurs linguistes comme WEBBER (1978) et PRINCE (1981) suivent cette approche basée sur l'analyse du discours pour traiter de l'anaphore[4]. WEBBER (1978) avance cinq arguments en faveur d'un « modèle de dis-

2. Exemples traduits.
3. Phénomène de coréférence, qui sera défini en 1.1.4.
4. Un phénomène linguistique lié à la référence défini en 1.1.3.

cours » dans ce but. Le premier argument étant que l'un des objectifs du discours est de permettre à un locuteur de communiquer à un interlocuteur sur un modèle qu'il peut avoir d'une situation. Le discours est donc le moyen de pouvoir partager ce modèle avec l'interlocuteur. Le second argument est que ce discours correspond à une collection structurée d'entités qui sont organisées selon leurs rôles et les relations qu'elles entretiennent. Le troisième argument de Webber correspond au fait que la fonction d'une expression anaphorique définie est de référer à une entité présente dans le modèle de discours du locuteur et que l'utilisation de cette expression par le locuteur permettra à l'interlocuteur d'accéder à une entité similaire dans son propre modèle de discours. Le quatrième argument est que le référent d'une expression anaphorique définie est présent dans le modèle de discours du locuteur et que ce dernier présume d'un référent équivalent dans le modèle de discours de l'interlocuteur. Le dernier argument correspond au fait que le choix du référent par l'interlocuteur se fait en partie en fonction de la manière dont le référent présent dans le modèle de discours du locuteur est décrit.

Comme le souligne aussi Cornish (2010) plus récemment, les approches traditionnelles du traitement de l'anaphore sont plutôt basées sur les termes du contexte textuel et ces approches doivent être révisées. Selon lui, la référence anaphorique ne peut correctement être décrite qu'en se basant sur l'interdépendance entre le texte et le discours tout comme le contexte. Nous nous baserons donc sur ces dernières approches reposant sur l'analyse du discours pour cette étude.

1.1.2 Les expressions référentielles : des représentations linguistiques

Les expressions référentielles sont des expressions linguistiques qui peuvent participer au phénomène de référence. Elles peuvent être de différents types : il peut s'agir d'un mot ou d'un groupe de mots (syntagme[5]) comprenant des noms, propres ou communs, ou encore des pronoms et même des déterminants possessifs, qui renvoient à la personne possédant la chose.

Les noms

De manière générale, les noms ont un pouvoir de référence. C'est le cas pour les noms communs comme pour les noms propres, comme le montre l'exemple suivant :

5. Selon le TLFi (Pierrel, Dendien et Bernard 2004) : « [Le plus souvent déterminé par un adj. spécifiant la nature du noyau du syntagme.] Groupe d'unités linguistiques significatives formant une unité dans une organisation hiérarchisée de la phrase. »

Exemple [3]
« [**Paul**] a pris [**l'ordinateur**] pour jouer à [**un jeu vidéo**] dans [**cette pièce**].[6] »

Dans l'exemple [3][7], il y a quatre expressions référentielles (avec quatre référents différents) : le nom propre « Paul », l'expression nominale définie « l'ordinateur », l'expression nominale indéfinie « un jeu vidéo » et l'expression nominale démonstrative « cette pièce ».

Une *expression nominale définie* est composée d'un article défini (*le*, *la*, *les*) suivi d'un nom. On parle souvent de « *description* définie » car, tout comme les syntagmes nominaux démonstratifs, possessifs et indéfinis et à la différence des noms propres et des pronoms, les syntagmes nominaux définis transmettent des informations sur la catégorie du référent (Charolles 2002, p. 75). Elles indiquent une unicité du référent et peuvent être « complètes » lorsqu'elles s'appliquent à une entité unique et possèdent une « autonomie référentielle ». Elles ont dans ce cas une « capacité identificatoire ». Elles peuvent aussi être « incomplètes » lorsqu'il y a un manque d'information concernant l'unicité du référent. Dans ce cas, le mode de référence est donc différent : la référence est limitée au contexte de la situation de communication. Selon Giurgea (2010), il existe aussi des expressions nominales sans nom apparent ou exprimé, avec une ellipse nominale. Dans l'exemple [3], il pourrait y avoir « le vieux », « le fixe », « le portable »[8] ou encore « le tout nouveau » à la place de « l'ordinateur ». Dans certains cas, il est nécessaire d'avoir une occurrence du nom dans le contexte précédent pour pouvoir le supprimer ensuite.

Une *expression nominale démonstrative* est une expression déictique : son interprétation dépend donc de la situation d'énonciation. La référence s'effectue alors grâce au contexte par le biais du démonstratif qui permet une rupture avec le contexte intellectuel de la référence (utile dans la référence des expressions définies) et implique une proximité avec le référent (Charolles 2002, p. 137).

Pour les *expressions nominales indéfinies*, l'établissement de la référence se fait à l'aide de l'énoncé entier et pas simplement de l'expression. Elles ne requièrent pas de « contact mental préalable » avec le référent et peuvent être spécifiques ou génériques (Charolles 2002, p. 181).

De même, les dates peuvent être des expressions référentielles comme dans l'exemple suivant :

6. Dans cette thèse, des crochets seront utilisés pour délimiter les expressions référentielles analysées. Un indice pourra aussi être utilisé avec ces crochets pour matérialiser le référent d'une expression et ainsi montrer les relations de coréférence ou non entre les expressions.
7. Exemple construit.
8. Cela peut aussi être interprété comme une substantivation de l'adjectif.

Exemple [4]
« [**Le 15 janvier 2019**] il est tombé assez de neige pour skier. »

Dans l'exemple [4], « Le 15 janvier 2019 » réfère à une date, qu'il est possible de reprendre par la suite avec un syntagme nominal comme « ce jour-là » par exemple. Les dates sont souvent introduites par un déterminant et peuvent fonctionner comme des syntagmes nominaux.

Un nom seul, même sans déterminant, peut aussi référer comme dans les trois exemples suivants :

Exemple [5]
« Il est actuellement en [**formation**] avec [**le chef de** [**département**]]. »

Dans l'exemple [5], « formation » et « département » ne possèdent pas de déterminant et n'ont pas nécessairement vocation à être référentiels. Cependant ils peuvent être référentiels dans la mesure où ils peuvent être repris. Ce phénomène de reprise est visible dans l'exemple suivant :

Exemple [6]
« [**Elle**]$_i$ était sans [**enfant**]$_j$. Pourtant [**elle**]$_i$ [**en**]$_j$ aurait voulu[9]. »

Dans l'exemple [6], le nom « enfant » peut référer sans déterminant. Il est même repris par la suite alors qu'il ne désigne pas, à l'origine, de référent établi dans le discours. Cela permet de prendre du recul à propos de l'exemple [2] de Karttunen dans lequel une reprise du même type serait possible : « Bill n'a pas de voiture mais il en aurait voulu une ». Cependant, la reprise n'implique pas nécessairement une relation de coréférence. En effet, si le premier nom fait référence au concept et que « une » réfère à une voiture spécifique, le « en » permet néanmoins de référer au concept, assurant une transition entre le générique et le spécifique. Ce pronom n'est pas présent en anglais, langue sur laquelle se base Karttunen (1976, p. 5) pour avancer son affirmation d'impossibilité de reprise. L'exemple suivant montre un autre type de nom sans déterminant, qui peut néanmoins être repris :

Exemple [7]
« Je m'habille souvent en [**rouge**]. »

Dans l'exemple [7], l'adjectif substantivé (Apothéloz 2002, p. 101) « rouge » réfère et peut même être repris par la suite : « cette couleur me donne bonne mine ».

9. Exemple et annotations issus du manuel d'annotation du corpus de Democrat.

Il peut arriver qu'une expression référentielle soit imbriquée dans une autre comme dans l'exemple suivant :

Exemple [8]
« **[Les chaussettes de [l'archiduchesse]]** sont-elles sèches ? »

Dans l'exemple [8], « les chaussettes de l'archiduchesse » est une expression référentielle en contenant une autre : « l'archiduchesse ».

Les noms propres

Les noms propres représentent une catégorie particulière et difficile à baliser. Leroy (2004) donne trois catégories de critères pour les caractériser, chacun de ces critères possédant immanquablement des exceptions. Premièrement, les critères factuels et sur la forme des mots portent notamment sur la marque graphique de la majuscule. Bien que répandu, ce critère ne suffit pas et d'autres viennent s'y ajouter : la non traduction et l'absence dans les dictionnaire. Deuxièmement, les critères morphosyntaxiques pointent l'absence de déterminant et de flexion. Pour finir, les critères sémantiques et pragmatiques font valoir la vacuité sémantique (ils ne possèdent pas de définition) et d'unicité référentielle (un seul référent est possible). C'est ce dernier critère qui retiendra particulièrement notre attention bien qu'un nom propre puisse référer à plusieurs entités différentes, comme dans les cas d'homonymie (Paris en France ou au Texas), ou qu'un nom commun puisse désigner un référent unique comme le soleil (Nouvel, Ehrmann et Rosset 2015).

Pour certains auteurs, le nom propre est « vide de sens » (Mill 1889, p. 100) : il peut référer mais il ne donne pas à lui seul d'indication descriptive sur son référent. Kripke, Jacob et Recanati (1982) ajoutent que le lien entre le nom propre et son référent repose sur une « convention particulière », il doit y avoir une « cérémonie de baptême » pour que celui qui le porte se nomme de la sorte. Il qualifie le nom propre de « désignateur rigide » en raison de la fixité du lien qui l'unit à son référent en dépit de ses évolutions. Cette théorie de vacuité de sens du nom propre a été critiquée par des auteurs selon lesquels le nom propre possède un sens (Russell 1905 ; Frege 1971). Searle (1972) avance qu'un nom propre signifie au moins un des attributs du référent. Un argument en faveur de cette théorie est la compréhension des phrases équatives dans lesquelles on peut remplacer un nom propre par un équivalent comme dans l'exemple suivant :

Exemple [9]
« L'[**Everest**] est le [**Chomolungma**]. »
John SEARLE, *Les Actes de Langage*, 1972.

Dans l'exemple [9], est une phrase équative dans laquelle le « Chomolungma » est le nom tibétain translittéré du mont Everest. Si les noms propres étaient vides de sens, il ne serait pas possible de comprendre ce type de phrase. Pour Charolles (2002, p. 57), la théorie de Searle confond la « valeur sémantique » du nom propre et les connaissances des locuteurs à propos du « porteur » du nom propre[10]. La théorie de Kleiber (1981) sur les noms propres est intéressante car elle avance que le sens du nom propre ne se situe pas dans les attributs du porteur mais dans la relation entre le nom propre et le porteur. L'information que porte un nom propre étant simplement qu'un porteur se nomme d'une certaine façon, le sens d'un nom propre est essentiellement « dénominatif ». Un nom propre indique donc que son référent n'est pas « n'importe quelle entité » (Kleiber 2004) mais une entité particulière, unique, spécifique (Jonasson 1994) et déjà catégorisée (Charolles 2002). Le nom propre finit même par « se charger de sens encyclopédique ».

Les attributs

Les adjectifs ne réfèrent pas, bien qu'ils puissent être repris par un pronom comme dans l'exemple suivant :

Exemple [10]
« Marie est [**brune**] et sa mère [**l**]'était aussi. »

Dans l'exemple [10], l'adjectif « brune » est repris par le pronom « l' » mais il ne désigne pas une entité. Les adjectifs désignent plutôt des attributs d'entités. En revanche, les attributs nominaux réfèrent, comme dans l'exemple suivant :

Exemple [11]
« **Paris** est [**la capitale de la France**]. »

Dans l'exemple [11], « la capitale de la France » réfère à la même entité que la ville de Paris.

10. L'entité particulière désignée par le nom propre.

Les pronoms

Les pronoms font partie de la classe des *mots grammaticaux*[11] : à la différence des *mots lexicaux*, leur nombre est limité et leur rôle syntaxique prime sur leur rôle sémantique. Toutefois, les pronoms ont une aptitude à référer tout autant que les noms. Ils ne donnent pas directement d'indication sur la catégorie de leur référent car ils ne possèdent pas de tête lexicale. Ils peuvent néanmoins référer de manière définie comme les expressions nominales définies. Les pronoms représentent une catégorie très hétérogène, l'exemple suivant en présente plusieurs types :

Exemple [12]
« [**Je**] suis contente que [**tu**] [**lui**] aies dit que cette robe était [**ma**] robe préférée. »

Dans le seul exemple [12], quatre pronoms différents ont un pouvoir de référence. Le pronom personnel de première personne « Je » réfère au locuteur. Le pronom personnel de deuxième personne « tu » réfère à son interlocuteur. Le pronom clitique objet « lui » désigne un troisième référent. Le syntagme nominal possessif « ma robe préférée » réfère à la robe mais le possessif « ma » indique une relation d'appartenance et fait référence au locuteur de manière indirecte.

Les pronoms relatifs ont la particularité de pouvoir introduire une nouvelle proposition : une subordonnée relative (déterminative ou explicative). De plus, ils possèdent une capacité de référence par leur statut de pronom, ils jouent un rôle de représentant comme dans l'exemple suivant :

Exemple [13]
« J'ai cuisiné le plat [**que**] tu préfères. »

Dans l'exemple [13], le pronom relatif « que » permet de relier la subordonnée à la principale et réfère indirectement au plat.

Un pronom clitique peut être réfléchi et renvoyer au sujet de la proposition, comme « me » dans l'exemple [14], ou non réfléchi, comme « le » dans l'exemple [15]. Un pronom non clitique peut aussi être réfléchi, comme « moi-même » dans l'exemple [16].

Exemple [14]
« Je [**me**] lave. »

Exemple [15]
« Je [**le**] lave. »

11. Appelés aussi parfois « mots outils ».

Exemple [16]
« Je me lave [**moi-même**]. »

Les pronoms indéfinis peuvent aussi référer comme dans l'exemple suivant :

Exemple [17]
« Les soldats sont partis lundi, [**quelques-uns**] sont revenus le lendemain. »

Le pronom indéfini « quelques-uns » de l'exemple [17] renvoie à une partie des soldats.

Les formes zéro

Il arrive que certaines expressions référentielles ne soient pas exprimées. Dans ce cas, on parle de *forme zéro* (notée ø) ou du *sujet zéro* d'un verbe par exemple. C'est le cas dans l'exemple suiant :

Exemple [18]
« [**Je**] restai deux ans à la maison et [**ø**] travaillai seul. »
Raymond RADIGUET, *Le Diable au corps*, 1923.

Dans l'exemple [18], le sujet du verbe « travaillai » n'est pas exprimé. La construction est pourtant complète : le complément n'est pas représenté par un pronom de forme pleine mais par une forme « vide » (Noailly 1996, p. 4).

Kerbrat-Orecchioni (2009) opère une distinction de modalité dans la référence des expressions référentielles. Certaines peuvent avoir une référence « absolue », comme les noms propres, dont le référent ne fait partie ni du texte ni de la situation d'énonciation. Cette modalité de référence repose sur la description de l'expression référentielle. D'autres expressions référentielles ont un mode de référence qui peut dépendre d'autres éléments. Elle parle pour ces expressions de référence « relative ». Dans ce cas, le contenu descriptif de l'expression référentielle ne suffit pas à identifier un référent de manière autonome. La référence peut donc être « relative au contexte linguistique » [12] si les informations permettant de construire la référence se trouvent dans le texte par exemple. Mais la référence peut être « relative à la situation d'énonciation » [13] si les informations supplémentaires nécessaires à la référence sont des expressions déictiques [14].

Les expressions référentielles peuvent donc référer de manière autonome ou bien reliées

12. ou « anaphorique »
13. ou « déictique »
14. Aussi appelées parfois « embrayeurs », « token-reflexives », « expressions sui-référentielles », « expressions indexicales » ou encore « démonstratifs » (Kleiber 1986). Ces expressions ne sont interprétables que dans les conditions particulières de l'énonciation. Par exemple : *là*, *je*, *demain*, etc.

par différentes relations impliquant le phénomène de référence. C'est le cas de l'anaphore ou la coréférence.

1.1.3 L'anaphore : une relation asymétrique

L'anaphore, du grec ancien *anaphora* (*ἀναφορά*) : *ana-* pouvant signifier « à nouveau » ou « en arrière » + *phorein* évoquant le fait de porter, implique donc l'idée de rappel. Nous ne nous intéresserons pas à l'anaphore *rhétorique*, qui est une figure de style jouant sur la répétition et consistant à commencer plusieurs phrases par le même mot ou groupe de mots. Nous nous intéressons plutôt à l'anaphore *grammaticale*, qui implique de se référer à un élément déjà présent dans le discours pour pouvoir en interpréter un autre : ici, la reprise est sémantique. Comme le souligne Corblin (1985a, p. 127), en anglais la relation est souvent nommée *anaphora* et le terme d'*anaphor* est utilisé pour désigner l'« anaphorique » : l'expression qui en reprend une autre. L'exemple suivant est issu d'un bloc que j'ai annoté pour le corpus Democrat[15]. Il contient une anaphore identifiée avec une simple requête sur le pronom « elle » en partant de l'hypothèse que ce type de pronom est souvent anaphorique :

> **Exemple [19]**
> « [**Marthe**]$_i$ se figura que je m'ennuyais de plus en plus. [**Elle**]$_i$ se sentait prête à tout pour me distraire. »
> Raymond RADIGUET, *Le Diable au corps*, 1923.

Dans l'exemple [19], le pronom « elle » est l'anaphorique du nom propre « Marthe ». Pour désigner le segment auquel on se réfère, il est courant d'utiliser le terme d'*antécédent* car il précède généralement l'anaphorique. Dans l'exemple [19], « Marthe » correspond donc à l'antécédent de « Elle ». Toutefois, Tesnière (1959) observe que cet antécédent peut parfois se trouver après l'anaphorique comme dans le cas du phénomène de *cataphore*. C'est pourquoi il préfère utiliser l'expression de « source sémantique » qui met l'accent sur le rôle de ce segment plutôt que sur sa localisation. Ducrot et Todorov (1972, p. 363) parlent quant à eux d'« interprétant de l'anaphorique » mais il est aussi parfois appelé « contrôleur de l'anaphorique », « (co)référent » ou encore « référé » (Reichler-Béguelin 1988). L'anaphore est donc une relation *asymétrique* et *non autonome* (Milner 1976, p. 65) dans la mesure où il est impossible d'interpréter l'anaphorique sans se reporter à l'antécédent. De plus, l'anaphore est *non transitive*[16] (Milner 1982, p. 33) car un anaphorique ne peut être anaphorisé à son tour. Toujours selon Milner, elle ne peut pas non plus être *réflexive* : un anaphorique ne peut pas s'anaphoriser lui-même.

15. Abordé dans la partie 2.1.
16. Sauf dans le cas de certaines successions de syntagmes nominaux.

Une anaphore est dite « fidèle » (Blanche-Benveniste et Chervel 2012) lorsque le nom utilisé est le même pour l'antécédent que pour l'anaphorique, même si le pronom varie. C'est le cas dans l'exemple suivant, qui est issu un bloc du corpus de Democrat et dont la requête permettant de l'identifier consiste à rechercher à l'aide d'une expression régulière le pronom « la » suivi d'un nom, suivis de plusieurs mots, suivis du pronom « cette » suivi du même premier nom :

Exemple [20]
« il aurait à descendre un bout de côte très raide, à traverser une immense prairie et à passer deux fois [**la rivière**]$_i$ à gué. Il lui avait même recommandé d'entrer dans [**cette rivière**]$_i$ avec précaution »
George SAND, *La Mare au Diable*, 1846.

Dans l'exemple [20], « la rivière » est repris par « cette rivière ». Le pronom varie mais pas le nom. Lorsque le nom varie, il s'agit par opposition d'une anaphore « infidèle », comme dans l'exemple suivant [17] :

Exemple [21]
« Le père prit la ligne de [**la fillette**]$_i$; mais son attention était ailleurs. À tout instant il se retournait pour regarder [**l'enfant**]$_i$. »
Marguerite AUDOUX, *Douce Lumière*, 1937.

Dans l'exemple [21], « la fillette » est repris par « l'enfant » à l'occasion d'une anaphore infidèle dont l'antécédent et l'anaphorique désignent cependant le même référent : le personnage Douce, une petite fille, endormie au moment où le père prend sa canne à pêche. Cette distinction entre les anaphores fidèles et infidèles est aussi l'occasion de mettre en avant le fait que des synonymes ne possèdent jamais complètement le même sens.

Il est aussi important de noter que si un anaphorique a besoin de son antécédent pour être interprété, ils ne désignent pas obligatoirement le même référent (Kleiber 1991, p. 6). Dans ce cas, Erku et Gundel (1987) parlent d'« anaphore indirecte ». Il peut alors s'agir de reprise uniquement lexicale, de relation partie-tout, d'un résumé de situation ou encore d'un passage au générique ou au collectif [18].

L'anaphore implique des termes (pronoms, démonstratifs, etc.) dont l'interprétation est « locale » (Corblin 1985b) car elle varie en fonction du contexte. Cette particularité est moins fréquente concernant la relation de coréférence.

17. L'exemple [21] est issu d'un bloc que j'ai annoté pour le corpus Democrat et dans lequel j'ai exploré la chaîne de coréférence (voir section 1.1.5) du personnage principal pour identifier une anaphore infidèle. Il est aisé d'en trouver sans avoir besoin de requête particulière.
18. Développé dans la partie 1.2.3

1.1.4 La coréférence : une relation symétrique

Les notions d'anaphore et de coréférence sont très proches et couvrent des domaines variés. Elles sont étudiées dans des domaines allant de la syntaxe (Guéron 1979 ; Gordon et Hendrick 1998 ; Adger 2003) au traitement automatique des langues (Poesio, Stuckardt et Versley 2016). Ces deux notions sont sont souvent confondues (Milner 1976, p. 65). Elles peuvent se recouper en effet mais l'une n'implique pas nécessairement l'autre.

Lorsque deux segments du discours désignent la même entité, elles sont reliées par une relation de coréférence (Corblin 1985b). En revanche, si une expression référentielle n'entretient aucun lien de coréférence avec une autre, on parlera de *singleton*. Ces notions sont illustrées dans l'exemple suivant :

> **Exemple [22]**
> « La cuisine de l'auberge n'était éclairée que par [**une lanterne de fer suspendue au plafond**]$_i$. Le squelette de [**ce luminaire**]$_i$ dessinait une large étoile d'ombre tremblotante sur tout l'intérieur de la pièce, et rejetait sa pâle clarté vers les solives enfumées du plafond. »
> George SAND, *Pauline*, 1881.

Dans l'exemple [22], « l'auberge » est un singleton car ce syntagme ne possède pas de lien de coréférence dans cet extrait. En revanche, « une lanterne de fer » et « ce luminaire » sont coréférents car ils désignent le même référent. Il s'agit aussi d'une anaphore grâce à la présence du démonstratif « ce » qui renvoie au syntagme nominal précédent. Les notions de coréférence et d'anaphore sont proches et peuvent se recouper. Dans l'exemple [19] abordé plus haut, « Marthe » et « elle » sont ainsi des expressions référentielles coréférentes. Nous avons vu que les anaphores peuvent être coréférentielles [19] ou non (1.1.3), et parallèlement une coréférence peut être anaphorique ou non : selon Kleiber (1988, p. 3), « toute expression coréférentielle d'une expression antérieure n'est pas nécessairement une expression anaphorique ». Cette remarque est illustrée par l'exemple suivant :

19. Point terminologique *coréférent* / *coréférentiel* : les emplois peuvent varier selon les auteurs et ces mots sont souvent synonymes. Nous y voyons néanmoins une différence subtile et nous utiliserons le terme « coréférentiel » pour désigner la qualité d'un élément à être impliqué dans une relation de coréférence et le terme « coréférent » pour désigner la nature de la relation entre deux éléments. Une anaphore *coréférentielle* implique une relation de coréférence dans laquelle une expression est *coréférente* à une autre.

Exemple [23]

« La terre tourne autour **[du soleil]**$_i$. **[Le soleil]**$_i$ est en effet le centre de l'univers. »

Georges KLEIBER, *Peut-on définir une catégorie générale de l'anaphore ?*, (1988, p.3)

Dans l'exemple [23], la première occurrence « du soleil » n'est pas nécessaire pour interpréter la seconde : « le soleil ». Cette relation de coréférence est par conséquent non anaphorique. La relation de coréférence est donc symétrique (Milner 1976, p. 65) : si A coréfère à B alors B coréfère à A. Milner souligne aussi que cette relation est transitive [20] : si A coréfère à B et que B coréfère à C alors A coréfère à C.

Les figures [1.1] et [1.2] schématisent le fonctionnement de la référence de la coréférence par rapport à celle de l'anaphore pour deux expressions référentielles (ER).

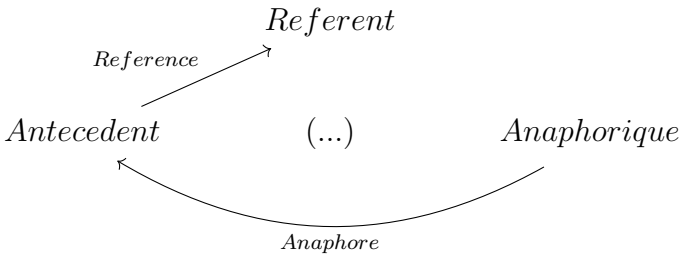


Figure 1.1 – La référence de l'anaphore.

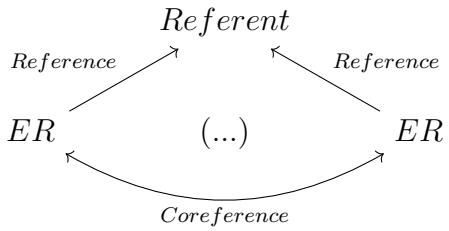


Figure 1.2 – La référence de la coréférence.

Comme pour la référence [21], Milner (1976) distingue la « coréférence virtuelle » de la « coréférence actuelle ». La coréférence actuelle est la relation entre deux groupes nominaux qui ont une référence actuelle. Il doit y avoir une « identité absolue » entre

20. À la différence de l'anaphore qui est : non autonome, asymétrique, non transitive et non réflexive (Voir partie 1.1.3).

21. Voir partie 1.1.1 sur la référence virtuelle et la référence actuelle.

les deux référents mais pas nécessairement entre les deux groupes nominaux. La coréférence virtuelle est quant à elle la relation entre deux unités lexicales différentes possédant exactement les mêmes propriétés lexicales. Elle est donc « impossible entre deux unités nominales quelles qu'elles soient » (Milner 1976, p. 72).

La coréférence et l'anaphore sont des phénomènes discursifs qui contribuent à la cohésion du discours en permettant par exemple d'éviter les répétitions. Elles peuvent aussi aider à apporter de nouvelles informations à propos d'un référent ou à rendre un élément saillant dans le discours comme dans l'exemple suivant :

Exemple [24]
« Elle dit la haute peine de [**Jacob**]$_i$ lorsque [**ses**]$_i$ fils [**lui**]$_i$ apportent la robe déchirée de [**son**]$_i$ enfant préféré. »
Marguerite AUDOUX, *Douce Lumière*, 1937.

Avec l'utilisation seule du nom propre :
« Elle dit la haute peine de [**Jacob**]$_i$ lorsque les fils de [**Jacob**]$_i$ apportent à [**Jacob**]$_i$ la robe déchirée de l'enfant préféré de [**Jacob**]$_i$. »

Dans l'exemple [24], la substitution des éléments anaphoriques coréférents par le nom propre « Jacob » rend compte de l'utilité fondamentale de ces deux notions dans le discours.

L'anaphore et la coréférence peuvent toutes deux être impliquées dans des « chaînes » d'expressions référant à une entité particulière tout au long d'un texte.

1.1.5 Les chaînes : une représentation de l'évolution d'un référent dans le discours

Il est maintenant établi que la coréférence est la relation entre plusieurs expressions référentielles qui désignent le même référent. L'ensemble de ces expressions forment ce que l'on appelle une *chaîne de coréférence*[22]. L'étude des chaînes de coréférence permet de suivre le « devenir discursif » (Schnedecker 2019, p. 2) d'un référent au fil d'un texte par exemple. Il s'agit donc de prendre en compte le texte dans sa « linéarité ». L'exemple suivant en donne un court aperçu :

22. Voir partie 1.2.2 sur la terminologie adoptée.

Exemple [25]

« Elle pensait à **[Noël]**$_i$. S'**[il]**$_i$ était là, **[il]**$_i$ saurait bien la défendre. Mais la ferme des Barray était peu éloignée du village, et **[Noël]**$_i$ n'avait rien à faire sur la route qui conduisait au Verger, distant de plus d'un kilomètre. »
Marguerite AUDOUX, *Douce Lumière*, 1937.

Dans l'exemple [25], « Noël », « il », « il » et « Noël » forment une chaîne de coréférence qui a pour référent le personnage de Noël Barray. Ce référent est repris dans différents passages dans le roman par des expressions aussi variées que « le jeune garçon », « ce Noël Barray », « son camarade », « qui », « son », « lui » et même « Je » lors de passages contenant du discours direct [23]. Cette chaîne de coréférence est relativement longue car elle concerne un personnage important du roman. Il en existe cependant des courtes comme dans l'exemple suivant :

Exemple [26]

« Dès le début de notre amour, Marthe m'avait donné **[une clef de son appartement]**$_c$, afin que je n'eusse pas à l'attendre dans le jardin, si, par hasard, elle était en ville. Je pouvais me servir moins innocemment de **[cette clef]**$_c$. »
Raymond RADIGUET, *Le Diable au corps*, 1923.

Dans l'exemple [26], « une clef de son appartement » et « cette clef » forment aussi une chaîne de coréférence [24] qui s'arrête là et ne continue pas plus loin dans ce bloc de texte. Ce bloc de texte correspond à un début de roman (les 10 000 premiers mots) pour lequel j'ai réalisé une annotation pour le corpus Democrat [25]. Cette chaîne de coréférence est l'une des nombreuses chaînes courtes issues de ce bloc. La longueur des chaînes est donc variable, cette longueur est généralement calculée en comptabilisant le nombre d'expressions référentielles (*maillons*) qu'elles contiennent. La figure [1.3] schématise différents exemples de longueur de chaîne (9 maillons pour la chaîne 1 contre 3 pour la chaîne 2) :

Indépendemment de la longueur, la répartition des maillons dans un texte peut varier d'une chaîne à l'autre. Cette idée est résumée dans la figure [1.4] qui représente trois chaînes d'une longueur de trois maillons dont la répartition est différente :

23. Une chaîne de coréférence assure la continuité référentielle dans un texte et elle et elle ne s'interrompt pas lors des passages au discours direct.
24. Bien que le nombre de maillon minimal d'une chaîne fasse débat (voir partie 1.2.2).
25. Voir section 2.1.

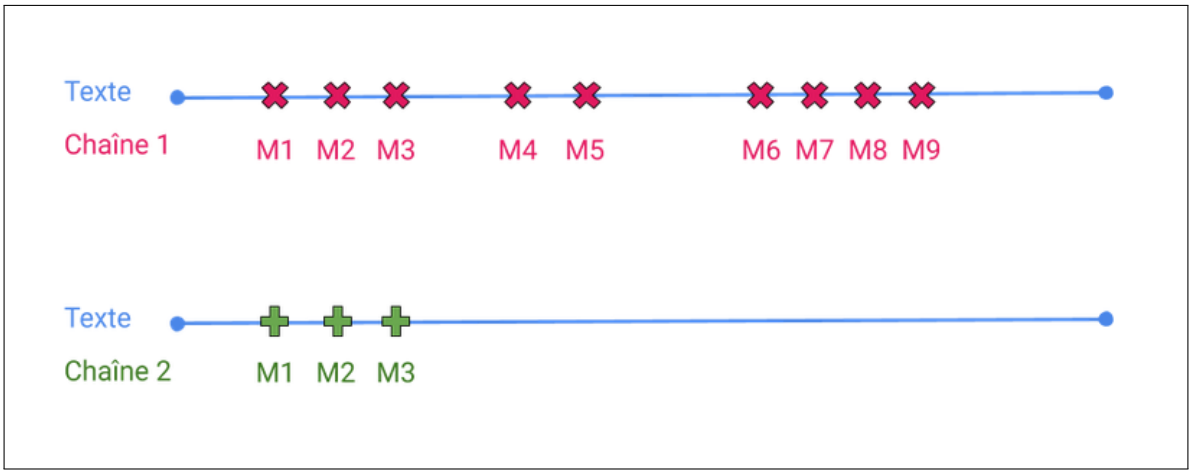


Figure 1.3 – Des chaînes de coréférence de longueurs différentes.

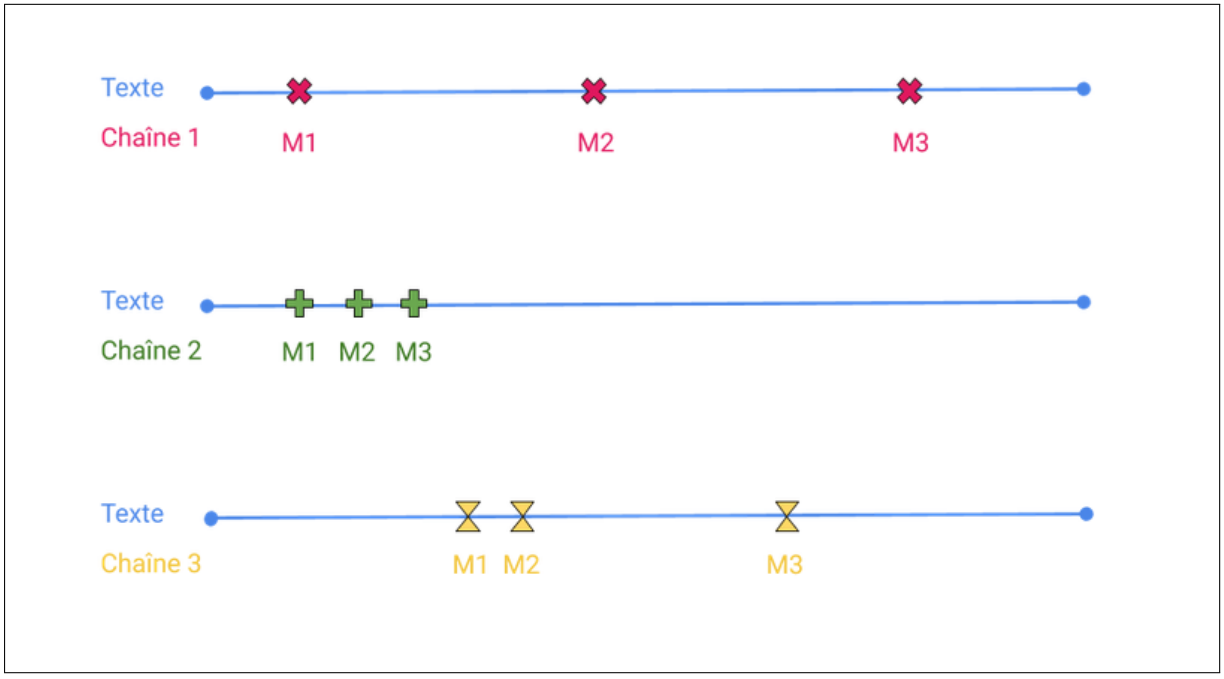


Figure 1.4 – Des chaînes d'une même longueur avec une répartition de maillons variable.

Pour étudier cette caractéristique dans les chaînes de coréférence, il est courant de calculer la « distance inter-maillonnaire » (Ariel 1990). Comme son nom l'indique, cette mesure permet de calculer la distance entre les maillons d'une chaîne de coréférence. Comme le souligne Schnedecker (2019), différentes questions se posent quant aux modalités de calcul de cette mesure. La distance est parfois calculée en nombre de syllabes (Lust 1981), d'autres fois en nombre de mots (Kibrik 2011) ou de syntagmes nominaux (Boudreau et Kittredge 2005) ou encore en nombre de phrases (Givón 1983). Une question se pose aussi à propos du point de départ pour le décompte des maillons : à partir du premier terme du maillon ou de sa tête syntaxique ?

D'autres mesures existent pour étudier les chaînes de coréférence comme le « coefficient de stabilité » (Perret 2000, p. 17) obtenu en « divisant, pour un référent donné (un personnage), le nombre total d'anaphores nominales par le nombre de désignations différentes ». Cette mesure permet d'étudier la « stabilité désignationnelle » : plus le coefficient est élevé, moins il y a de manières de désigner le référent par rapport au nombre d'anaphores. Un coefficient de stabilité élevé indique donc une chaîne de coréférence avec une grande stabilité référentielle.

Boudreau et Kittredge (2005) s'intéressent à une autre mesure : la « portée des chaînes ». Elle permet de calculer le pourcentage de texte que couvre une chaîne de coréférence. De son côté, Givón (1983) calcule la « persistance » des chaînes de coréférence en prenant en compte le nombre de maillons avant et après un certain point référentiel.

Les différentes notions abordées et définies dans cette section sont regroupées dans le tableau suivant :

Notions	**Caractéristiques**
Référence	Phénomène linguistique permettant de désigner un référent dans le discours.
Expression référentielle	Expression linguistique servant de support à la référence.
Anaphore	Relation asymétrique entre deux expressions référentielles, qui ne sont pas nécessairement coréférentes, dans laquelle l'interprétation de l'une dépend de l'autre.
Coréférence	Relation symétrique entre deux expressions référentielles qui désignent la même entité.
Chaîne de coréférence	Ensemble des expressions référentielles coréférentes désignant une même entité.

Tableau 1.1 – Notions impliquées dans les chaînes de coréférence.

La définition et l'état de l'art concernant les notions impliquées dans le phénomène des chaînes de coréférence étant établies, il est temps de délimiter précisément notre champ d'étude et notre positionnement pour ce travail.

1.2 Positionnement

Il existe diverses approches pour l'étude des chaînes de coréférence. Elles impliquent souvent des méthodologies différentes ainsi qu'une terminologie particulière. En fonction des études, les phénomènes pris en compte ne sont pas toujours les mêmes. Dans cette

partie, nous délimiterons notre objet d'étude et préciserons la terminologie adoptée ainsi que l'approche suivie dans cette thèse.

1.2.1 Différentes approches

Pour traiter les phénomènes se rapprochant de la coréférence, différentes approches sont adoptées. Elles varient en fonction des périodes, du domaine d'étude, du courant linguistique ou encore de l'objectif recherché.

Approches linguistiques issues des travaux sur l'anaphore

Dans le cadre de la Grammaire Générative Transformationnelle, des auteurs comme Lees et Klima (1963) [26] ont abordé la question de la coréférence à travers la relation entre les syntagmes nominaux et les pronoms. Ces travaux ont aboutit à des propositions de règles syntaxiques de coréférence (ou plutôt de non-coréférence). À la suite de cela, des travaux comme ceux de Guéron (1979) et Fauconnier (1974) ont abordé la question de la coréférence d'un point de vue sémantique et syntaxique [27].

La Théorie du Gouvernement et du Liage de Chomsky (1987) s'intéresse aux relations entre les expressions nominales et leur interprétation à travers la théorie du liage (Haegeman 1991). Dans cette théorie, l'étude de l'anaphore est limitée aux expressions nominales au sein d'une même phrase et l'antécédent est nécessairement un segment de texte (Gardelle 2012). La notion d'anaphore repose sur des règles grammaticales issues de courants génératives. Les aspects pragmatiques ne sont donc pas pris en compte dans cette approche.

D'autres études linguistiques non génératives considèrent l'anaphore comme une relation qui n'est pas simplement syntaxique. L'anaphore est tout de même considérée comme un phénomène *textuel* (Kleiber 1993) où l'antécédent doit être une expression « mentionnée dans le texte ». Comme nous l'avons abordé dans la partie 1.1.3, d'autres auteurs comme Cornish et al. (1999) proposent une autre définition de l'antécédent qui ferait partie d'une *représentation mentale* dont l'antécédent serait néanmoins un segment de texte. Qualifiées d'« approche mémorielle » par Kleiber (2001), d'autres approches (Charaudeau et Maingueneau 2002 ; Allan 2009) considèrent que l'aspect textuel n'est plus exclusif. Ces approches, plus pragmatiques, envisagent que l'antécédent doit faire partie de la représentation mentale et l'anaphore peut être soit textuelle, soit relative à la situation d'énonciation. Gardelle (2012, p. 38) conclut que la relation d'anaphore

26. Lees et Klima (1963) conçoivent les pronoms comme étant des substituts de syntagmes nominaux, ce qui sera remis en question dans d'autres travaux par la suite.
27. Une synthèse de ces travaux a été réalisée par Charlent (1983).

« prototypique » reste textuelle par nature mais qu'elle peut être saturée par différentes sources extralinguistiques.

Dans le domaine de la linguistique théorique, il s'agit de formaliser des théories sur la langue, principalement à destination de linguistes. Certains travaux sont basés sur des « énoncés fabriqués » (Schnedecker 2019, p. 10) mais avec le nombre croissant de corpus créés à l'heure actuelle, de plus en plus de travaux en linguistique comprennent l'utilité de partir de données attestées pour pouvoir étudier ce phénomène linguistique.

Approches issues de la linguistique de corpus

En linguistique de corpus, l'objectif est donc plutôt de partir d'énoncés attestés en interrogeant des corpus pour pouvoir en tirer des analyses (Tutin 2002 ; Condamines 2005 ; Baumer 2012 ; Schnedecker 2014 ; Schnedecker et Landragin 2014). Ces corpus sont des bases de données linguistiques créées avec un objectif précis. Ils comportent donc des caractéristiques particulières en fonction de cet objectif. Lorsqu'il s'agit de corpus textuels, la nature des textes peut varier d'un corpus à un autre : il peut s'agir de textes littéraires, journalistiques ou même de tweets ou d'extraits de blogs par exemple. Les textes peuvent être des textes rédigés mais ils peuvent aussi être de l'oral transcrit. La langue est également un critère : certains corpus sont monolingues et d'autres multilingues [28] afin de pouvoir comparer des phénomènes dans des langues différentes. Enfin, la plupart des corpus comportent des annotations, manuelles ou automatiques, qui varient elles aussi en fonction de l'objectif du corpus. Dans le projet Democrat [29], il a été décidé d'annoter les expressions référentielles (manuellement) mais aussi (automatiquement) les parties du discours [30] et les lemmes [31] à l'aide de l'outil TreeTagger (H. Schmid 1994) sur la plateforme TXM (Heiden, Magué et Pincemin 2010a). Les corpus ont des applications en linguistique descriptive afin d'aider à créer des modèles linguistiques issus des données, mais ils sont aussi utiles au traitement automatique des langues afin de servir d'entraînement à des systèmes de détection de coréférences par exemple.

Approches issues du traitement automatique des langues

Le traitement automatique des langues [32] (TAL) est un domaine à la croisée des domaines des sciences du langage, de l'informatique et de l'intelligence artificielle. Le TAL

28. Qu'ils soient parallèles, comparables ou non (Yapomo 2013).
29. Voir partie consacrée : 2.1
30. Classes grammaticales - Documentation du jeu de balises : https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html
31. Le lemme est la forme canonique neutre d'un mot : celle que l'on retrouve dans le dictionnaire.
32. Ou encore Traitement Automatique des langues naturelles (TALN) et *Natural Language Processing* (NLP) en anglais.

vise à traiter, modéliser voire reproduire le langage naturel au moyen d'outils informatiques. Ce domaine possède de ce fait des champs d'applications très variés comme le traitement du signal, avec la tâche de synthèse vocale par exemple, ou encore la syntaxe avec, à titre d'illustration, les tâches de lemmatisation ou d'étiquetage morphosyntaxique[33]. En traitement automatique des langues, l'objectif général est de créer des systèmes qui fonctionnent pour la langue dans sa globalité. Dans cette optique, en extraction d'information, le travail s'effectue principalement sur la base de grands corpus textuels (ou oraux) numériques dont on cherche à extraire des connaissances comme les *entités nommées* par exemple. Il s'agit d'un cas particulier d'expressions référentielles qui correspondent généralement à des personnes, des produits, des organisations, des lieux ou des événements. Ces étiquettes varient en fonction des projets ou campagnes d'évaluation et de leurs objectifs et applications. La reconnaissance des entités nommées est une tâche répandue en fouille de texte (*text mining*), elle est à la base de nombreuses autres tâches et permet notamment l'anonymisation de corpus ou encore la construction de bases de connaissances et plus généralement « l'accès à l'information » (Nouvel, Ehrmann et Rosset 2015). La reconnaissance d'entités nommées et l'*entity linking* (relier les mentions d'une même entité nommée entre elles) sont très proches de la tâche de détection des coréférences. On parle d'ailleurs plutôt de détection des relations de coréférence plutôt que de l'identification des chaînes en traitement automatique des langues.

Le traitement automatique des langues a commencé à s'intéresser à la coréférence à travers la tâche de *détection d'anaphores* dès les années 1960. Poesio, Stuckardt et Versley (2016) identifient cinq phases chronologiques représentatives de différents paradigmes de recherche liés à cette tâche. Ils appellent la première phase (1964-1981) « Very Early Work ». Cette période regroupe des études basées sur les anaphores nominales et pronominales à l'aide de systèmes à base de règles avec des techniques de recherche de correspondance de motifs. La seconde phase (1982-1987) est appelée « Early Work » et acquiert plus de robusticité en étant moins spécifique à un domaine en particulier. Ils appellent la troisième phase (1988-1994) « The Consolidation Phase » car les travaux de cette période sont de plus en plus robustes à l'aide de critères et de restrictions plus précis à propos de l'antécédent et des approches prenant en compte la syntaxe. La quatrième phase (1995-2001) est appelée « The Ressource-Driven, or Robustification Phase » car les systèmes commencent à prendre en compte des ressources telles que les annotations morphosyntaxiques et les chercheurs utilisent des systèmes statistiques d'apprentissage supervisé dont les résultats peuvent être évalués. C'est à cette période que la tâche de détection de coréférences apparaît lors des conférences MUC (MUC-6 1995) initiées et fiancées par la DARPA[34]. Leur dernière phase, la phase actuelle (2002-), se nomme

33. Donner la classe grammaticale pour chaque mot.

34. Defense Advanced Research Projects Agency (DARPA) : « Agence pour les projets de recherche avancée de défense ». Il s'agit d'une agence du département de la Défense des États-Unis qui a pour

« Recent Developments » et est largement influencée par l'existence de corpus annotés en expressions référentielles qui permettent de pousser plus loin l'utilisation de l'apprentissage automatique utilisant de plus en plus de nouvelles méthodes statistiques (Ng 2010), supervisées ou non, ainsi que des nouvelles métriques d'évaluation des systèmes et de nouvelles campagnes d'évaluation prenant le relais de MUC.

Les travaux sur la coréférence en traitement automatique peuvent adopter différentes approches pour la détection de coréférences (Recasens 2010). Dans les méthodes d'apprentissage statistique supervisé, on retrouve différents modèles. Le *mention-pair model* proposé dès 1995 par McCarthy et Lehnert (1995) et Aone et Bennett (1995) consiste à classifier les expressions référentielles (mentions) par paires coréférentes ou non. Ce modèle est populaire et obtient des résultats corrects en dépit de son incapacité à rendre compte de la transitivité [35] inhérente à la notion de coréférence. L'*entity-mention model* cherche plutôt à déterminer si une mention est coréférente à un autre ensemble de mention, qui forment un « cluster » correspondant à un référent. Cette approche appliquée par X. Luo et al. (2004) n'a pas réellement donné de meilleurs résultats [36] que le mention-pair model (Ng 2010). Les adaptations de ce modèle par Daumé III et Marcu (2005) et Culotta, Wick et McCallum (2007) ont permis par la suite des améliorations en permettant la prise en compte de variables [37] (*features*) au niveau des clusters. Le *ranking model* (Connolly, Burger et Day 1997) apporte un nouveau progrès avec le calcul de l'antécédent le plus probable pour une mention donnée. Les dernières avancées dans ce domaine sont principalement basées sur l'utilisation du *deep learning* et des réseaux neuronaux artificiels (*artificial neural networks*) (Wiseman et al. 2015 ; Lee, He et Zettlemoyer 2018 ; Tourille 2018 ; H. Luo et Glass 2018 ; Alfaro, Ruiz Costa-Jussà et Rodriguez Fonollosa 2019 ; Grobol 2019 ; Wilkens et al. 2020).

Plusieurs travaux en linguistique portent sur les entités du discours et donnent des indications sur les critères qui peuvent aider les systèmes de traitement automatique à prédire si une expression référentielle sera reprise (coréférente) ou non (singleton). Ces critères peuvent relever de la morphologie et de la syntaxe (Prince 1981 ; Wang, McCready et Asher 2006), des rôles discursifs (Chafe 1976 ; Hobbs 1979 ; Walker 1998 ; D. I. Beaver 2004) et de la sémantique (Karttunen 1976 ; Kamp 1981 ; Heim 1982 ; Roberts 1997). Recasens, De Marneffe et Potts (2013) ont pris en compte ces caractéristiques pour les intégrer à un modèle capable de distinguer les singletons des maillons coréférents. D'autres travaux se sont aussi penchés sur la détection des chaînes en fonction du type de texte (Boudreau 2004).

objectif la recherche et le développement des nouvelles technologies militaires.

35. Voir partie 1.1.4
36. En revanche, ce fut le cas pour Yangy et al. (2004).
37. Des paramètres ou informations supplémentaires à prendre en compte.

Dans ce domaine, l'étude de la coréférence se fait de manière globale, prenant en compte les chaînes de coréférence à partir de deux maillons [38]. Tout comme la reconnaissance d'entités nommées, la détection de la coréférence peut avoir des applications en traduction automatique (Mitkov 1996), notamment pour la traduction des pronoms. Cette tâche est utile en recherche et extraction d'informations (MUC-6 1995) mais aussi pour les résumés automatiques (Azzam, Humphreys et Gaizauskas 1999) ou les systèmes de question-réponse.

1.2.2 Terminologie adoptée : des précisions sur les phénomènes étudiés

Les chaînes de (co)référence(s)

En anglais, pour désigner les expressions référentielles désignant le même référent, on parle de « coreference chain » (Azzam, Humphreys et Gaizauskas 1999), « coreferential chain » (Mitkov 1999), « cohesive chain » (Tseng 2008) ou simplement de « coreferences » (Bagga 1998). En français, elles sont nommées « chaîne de référence » [39] par de nombreux linguistes francophones depuis Chastain (1975) et par la suite Corblin (1985a), Charolles (1988) et Schnedecker (1992). Ce phénomène est aussi appelé « chaînes coréférentielles » (Schang, Antoine et Lefeuvre-Halftermeyer 2017) mais il est souvent appelé « chaîne de coréférence(s) » (Boudreau et Kittredge 2005 ; Muzerelle, Schang, Antoine, Eshkol, Maurel, Boyer et al. 2012 ; Godbert et Benoit 2017 ; Landragin et Oberle 2018) en particulier dans le domaine du traitement automatique des langues. Schnedecker (2019) met en lumière l'importance de la notion de « chaîne de référence » contrastivement aux notions d'anaphore et de coréférence auxquelles elle ne se superpose pas. D'un point de vue linguistique, elle précise que cette notion est « un moyen d'accès privilégié à ce qu'on nomme la référence discursive ».

La progression des chaînes de coréférence permet de représenter visuellement la présence d'un référent tout au long d'un texte. Cette représentation des chaînes de coréférence est visible sur les diagrammes [40] des figures suivantes :

38. Voir partie 1.2.2 à propos du nombre de maillons d'une chaîne.
39. Le terme de « chaîne » vient de Vendler (2019).
40. Obtenus à l'aide de l'outil TXM (Quignard et al. 2018)

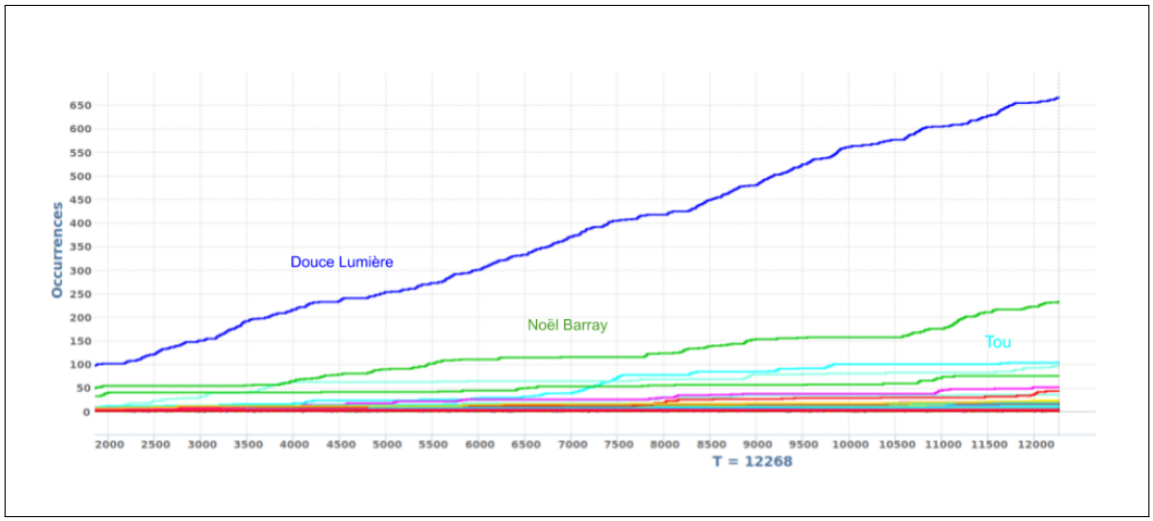


Figure 1.5 – Diagramme de progression des chaînes de coréférence dans les 12268 premiers mots de *Douce Lumière* de Marguerite AUDOUX (1937).

Dans la figure [1.5], une chaîne est plus longue que les autres. Son référent est le personnage principal du roman : Douce Lumière. Cette chaîne possède une progression forte et constante au cours de l'extrait car son référent est mentionné régulièrement. Les autres chaînes qui se démarquent sont celles des personnages secondaires (Noël Barray et Tou).

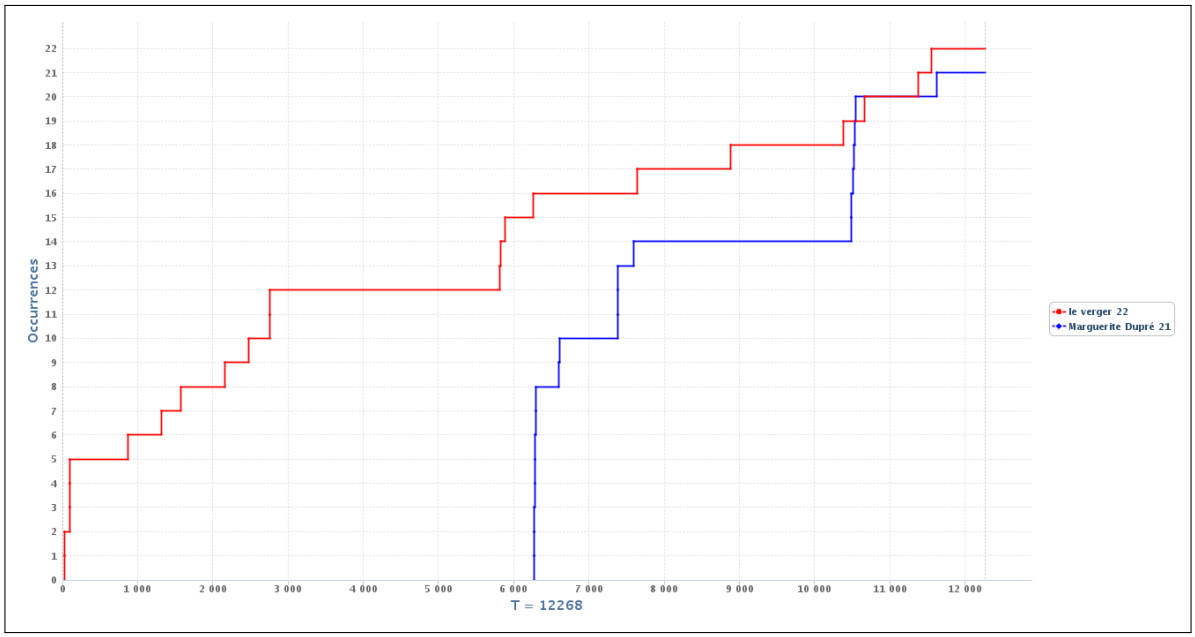


Figure 1.6 – Diagramme de progression de deux chaînes de coréférence différentes dans les 12268 premiers mots de *Douce Lumière* de Marguerite AUDOUX (1937).

Le diagramme de la figure [1.6] présente plus précisément deux chaînes de coréférence de 21 et 22 maillons. Ces deux chaînes possèdent une progression tout-à-fait différente au

cours du texte. La chaîne référant au verger débute dès le début du texte et est reprise jusqu'à la fin avec un palier entre le cap des 3000 et des 6000 mots. La chaîne référant au personnage de Marguerite Dupré débute quant à elle à la moitié du texte et forme deux groupes de mentions avec elle aussi un palier entre ces deux groupes. Cela revient à dire que l'auteur mentionne ce référent dans deux passages de cet extrait.

Ces figures montrent la variabilité de la composition des chaînes de coréférence. Comme le schématisait la figure [1.4], les chaînes sont de taille variable et la disposition de leurs maillons dans le texte peut être plus ou moins éparse. Avec ce type de fonctionnalité, l'outil TXM permet d'opérationnaliser l'étude des chaînes de coréférence.

Les maillons

Les expressions référentielles formant les *maillons* de ces chaînes sont les *mentions* coréférentes d'un même référent. Le terme de *mention* étant plutôt employé dans les études dans le domaine du traitement automatique des langues.

Le nombre minimum de maillons d'une chaîne est un sujet qui divise. Pour certains auteurs (Schnedecker 1992 ; Corblin 1995), une « chaîne de référence » comporte au moins 3 maillons, les notions de coréférence et d'anaphore étant utilisées pour décrire les relations entre seulement deux expressions référentielles. De manière générale en traitement automatique des langues, et comme pour Désoyer et al. (2015) par exemple, une chaîne de coréférence à deux maillons est tout à fait acceptable, pour ne pas dire essentielle. La prise en compte des chaînes à deux maillons permet de regrouper sous un même terme commode deux phénomènes parfois considérés comme similaires, parfois désignés par des termes séparés. L'une des fonctionnalités essentielles des chaînes étant de pouvoir suivre le devenir d'une entité dans le discours, il est intéressant de se pencher sur ces petites chaînes car les deux maillons peuvent être de nature différente ou non, proches ou non, etc. Ces informations sont intéressantes à prendre en compte, tout comme la position des singletons (que certains pourraient voir comme des chaînes à un maillon), afin de les mettre en parallèle avec les autres dans une étude globale de la coréférence. Le graphique suivant montre la fréquence des chaînes (et des singletons) en fonction du nombre de maillons, dans un bloc de texte de Democrat :

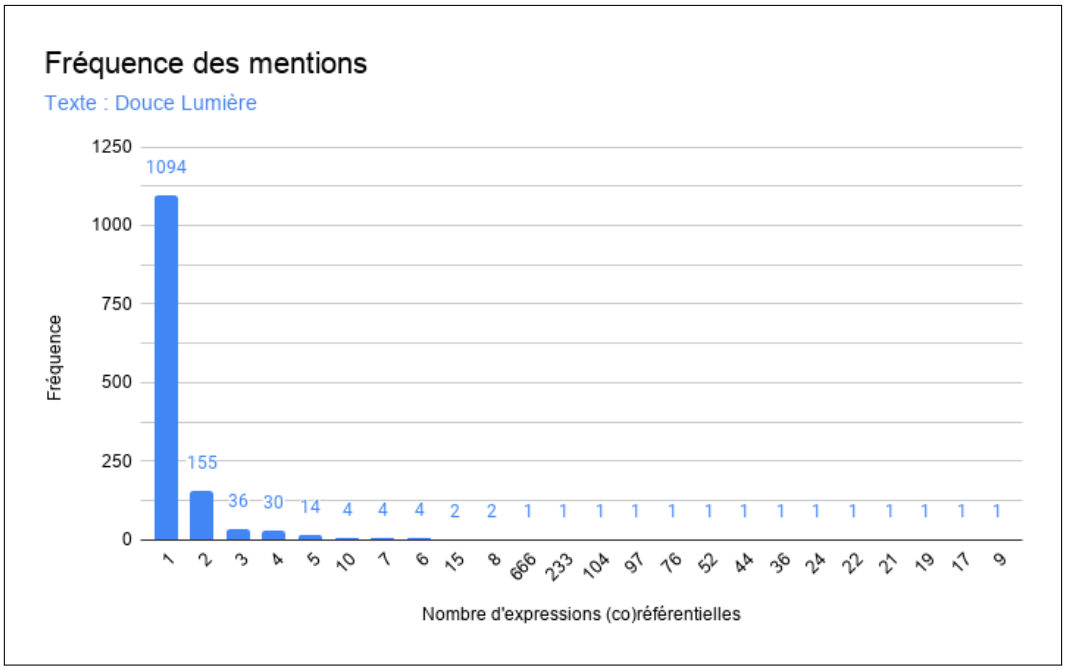


Figure 1.7 – Fréquence des expressions référentielles et mentions coréférentes dans les 12268 premiers mots de *Douce Lumière* de Marguerite AUDOUX (1937).

De manière générale et comme dans le graphique de la figure [1.7], pour un bloc de texte de 12268 mots, les chaînes comportant deux maillons sont plus nombreuses que les chaînes à trois maillons, elles-mêmes plus nombreuses que les chaînes à quatre maillons et ainsi de suite. Les mentions les plus nombreuses étant celles qui ne sont pas reprises (les singletons). Lorsqu'il est question d'extraire les relations de coréférence d'un texte de manière automatique, les chaînes de deux maillons sont donc importantes à prendre en compte ainsi que les singletons car ce sont des maillons potentiels de chaînes plus longues (quitte à les écarter par la suite si besoin). L'appellation « chaînes de coréférence » étant plutôt répandue dans les travaux de détection automatique de la coréférence prenant en compte les chaînes à partir de deux maillons, c'est cette dénomination que nous choisissons pour la rédaction de cette thèse.

1.2.3 Portée du phénomène et limites de la thèse

Afin d'achever le cadre théorique de cette thèse, il est nécessaire d'établir la portée de la notion de chaînes de coréférence avec les phénomènes (co)référentiels que nous retiendrons ou non.

Phénomènes non référentiels

Ce travail porte sur les chaînes de coréférence, nous nous intéresserons donc aux phénomènes coréférentiels[41]. Pour qu'il y ait *coréférence*, il doit y avoir *référence*. Nous laisserons donc de côté les adjectifs[42] car ils ne désignent pas des entités mais plutôt des attributs d'entité. De même, nous ne nous intéresserons pas aux noms dont l'emploi n'est pas référentiel. Par exemple, dans le cas des expressions figées comprenant des parties du corps humain comme dans l'exemple [27], il n'y a de référence à aucun pouce.

Exemple [27]
« J'ai mangé sur [**le pouce**] avant de venir. »

Les noms communs épithètes, qui sont apposés directement derrière un autre nom comme « gâteau » dans « mamie gâteau », ne réfèrent pas non plus. Par ailleurs, il n'y a pas de référence dans le cas des noms sans déterminants et introduits par une préposition à l'intérieur d'un nom composé comme « armoire à glace ». D'autant plus dans cet exemple qui désigne métaphoriquement un homme fort et non une armoire ni une glace.

Certains noms propres peuvent aussi être non référentiels, dans le cas des noms propres épithètes à valeur identifiante comme dans « l'avenue Montaigne » ou « le prix Pulitzer » pour lesquelles il n'y a pas de référence directe à l'écrivain ou au journaliste bien qu'il y ait un lien dans l'origine de l'appellation.

Il arrive aussi que des pronoms ne soient pas référentiels. C'est le cas des pronoms présents dans les expressions figées, notamment verbales, comme dans l'exemple [28] où le pronom « la » ne réfère pas.

Exemple [28]
« Il est en vacances alors il se [**la**] coule douce »

Les pronoms impersonnels ne sont pas référentiels non plus car l'actant du verbe n'est pas identifiable. Dans l'exemple [29], le pronom « il », explétif, ne réfère pas. Les emplois météorologiques sont particulièrement propices à ce type de phénomène à l'instar du célèbre « Il pleut. » où, bien sûr, personne ne « pleut ».

Exemple [29]
« [**Il**] y avait du monde au marché ce matin. »

Les pronoms négatifs ne désignent aucun élément : ils ne peuvent donc pas référer.

41. Des phénomènes impliqués dans une ou plusieurs relations de coréférence.
42. Voir la section consacrée : Les attributs

Dans l'exemple [30], le pronom « aucun » ne réfère pas. En revanche, le syntagme nominal « ses amis » réfère à une entité particulière tout comme le possessif « ses ».

Exemple [30]
« [**Aucun**] de ses amis n'est venu à son anniversaire. »

Dans la section sur les expressions référentielles (1.1.2), nous avons évoqué les expressions nominales et pronominales car elles représentent les catégories évoquées le plus couramment dans la littérature. Certains auteurs parlent pourtant d'une « référence verbale » (Asnes 2004). Bien que passionnant, cet aspect de la référence ne sera pas étudié dans cette thèse.

Phénomènes non coréférentiels

Les anaphores non coréférentielles comme les *anaphores indirectes* (1.1.3) seront aussi laissées de côté. Parmi ces anaphores, il y a les *anaphores associatives*. L'anaphorique est interprétable grâce à l'intégration de certaines relations de contiguïté dans le discours comme les relations méronymiques, locatives, fonctionnelles et actancielles (Kleiber 2001). Pourtant, l'anaphorique n'est pas coréférent à son antécédent. Ce qu'illustre l'exemple suivant :

Exemple [31]
« Les nouveaux amis l'accompagnèrent jusqu'à [**la grille du potager**]$_i$ par où il était venu. À pleines mains, il empoigna [**les gros barreaux rouillés**]$_j$, se hissa au faîte, enjamba les lances pointues avec adresse, se laissa tomber de haut, et tout courant, s'en fut parmi les grands sapins. »
Marguerite AUDOUX, *Douce Lumière*, 1937.

Dans l'exemple [31], il y a un cas d'anaphore associative : « les gros barreaux rouillés » sont interprétables en se référant à « la grille du potager » bien qu'il ne s'agisse pas exactement du même référent. Il est question d'une relation partie-tout entre ces deux référents, dont le deuxième n'est interprétable que grâce au premier.

L'*anaphore lexicale*[43] (Milner 1976, p. 72) fait aussi partie de ces anaphores indirectes. Cette relation implique des référents différents : l'anaphorique reprend uniquement une unité lexicale comme dans l'exemple [32] où le pronom « en » reprend « trois parts de gâteau » sans désigner le même référent.

43. Ou nominale.

Exemple [32]
« Paul a pris [**trois parts de gâteau**]$_i$ alors j'[**en**]$_j$ ai mangé seulement deux. »

Nous ne prenons pas non plus en compte les anaphores qui ne possèdent pas d'expression référentielle comme antécédent, comme les *anaphores résomptives* car il ne s'agit pas de coréférence. Comme dans l'exemple [33], l'antécédent d'une anaphore résomptive ne correspond pas à une expression référentielle mais à un élément du discours qui peut être évoqué, plus ou moins précisément, par une ou plusieurs phrases, consécutives ou non. La situation développée tout au long d'un ou plusieurs paragraphes peut ainsi être reprise par l'expression « toute cette agitation » par exemple. Avec le même processus, tout un chapitre de thèse peut lui aussi être repris par « cette question épineuse ». Comme le souligne Kleiber (1994), pour analyser une anaphore non coréférentielle, « il faut en plus de l'antécédent trouver le mécanisme qui permet de passer du référent de l'antécédent à celui de l'expression anaphorique ».

Exemple [33]
« Aux courses désordonnées s'étaient tout de suite ajoutés les jeux hardis et violents. Douce, légère et souple, suivait avec intérêt tous les mouvements de son camarade. Et derrière lui, elle faisait des culbutes savantes, franchissait des obstacles, grimpait jusqu'au faîte des arbres pour se nicher entre les feuilles ou se balancer entre les branches. Puis Noël se lassa de [**tout cela**]. »
Marguerite AUDOUX, *Douce Lumière*, 1937.

Dans l'exemple [33], « tout cela » est anaphorique et renvoie aux choses dont s'est lassé Noël. Certaines de ces choses sont exprimées par des expressions référentielles comme les « courses désordonnées » ou « les jeux hardis et violents » mais elles ne sont pas directement consécutives. Cependant, le problème le plus important est que certaines de ces choses sont exprimées à travers des syntagmes verbaux comme « elle faisait des culbutes savantes ». Or, il ne s'agit pas d'expressions référentielles, il ne peut donc pas y avoir coréférence. Dans ce cas, si l'expression « tout cela » s'avérait être reprise par la suite par une autre expression coréférente, elle pourrait faire partie d'une chaîne de coréférence dont elle serait le premier maillon. Si elle n'est pas reprise, elle sera plutôt considérée comme un singleton.

Les cas d'anaphore collective (exemple [34]) ou générique (exemple [35]) (Kleiber 1991) posent aussi des problèmes de coréférence comme le démontrent les exemples suivants :

Exemple [34]
« A Strasbourg, **ils** roulent comme des fous »
Georges KLEIBER, *Anaphore-deixis : où en sommes-nous ?*, 1991.

Dans cet exemple, le « ils » implique une interprétation anaphorique mais il est impossible de faire coréférer directement ce pronom à la ville de Starbourg.

Exemple [35]
« Paul a acheté **une Toyota**, car **elles|ces voitures** sont robustes »
WEBBER Bonnie Lynn, *« So What Can We talk About Now ? »*, 1983.
cité par : Georges KLEIBER, *Anaphore-deixis : où en sommes-nous ?*, 1991.

La situation est similaire dans cet exemple, il n'y a pas de lien de coréférence entre la Toyota de Paul et les Toyotas, la relation est simplement anaphorique.

Phénomènes coréférentiels

Plusieurs niveaux de contribution à une chaîne de coréférence ont déjà été pris en compte dans différents projets. C'est le cas avec la distinction entre les maillons forts et les maillons faibles (Landragin 2011b, p. 8) en opposant les expressions référentielles aux indices qui évoquent les référents mais sans réellement référer nécessairement, comme les marques d'accord ou les sujets non exprimés de verbes. On retrouve ce dernier phénomène dans les exemples [18] et [36].

Exemple [36]
« **[Ces braves]**$_i$ arrivèrent enfin et **[ø]**$_i$ fendirent la foule. »
Raymond RADIGUET, *Le Diable au corps*, 1923.

Dans l'exemple [36], le sujet du verbe « fendre » n'est pas exprimé, ce *sujet zéro* est donc un « élément référentiel linguistiquement non marqué » (Landragin 2011b, p. 8) qui évoque « Ces braves ». Il y a donc une « anaphore zéro » (Levinson 1987) car l'anaphorique n'est pas exprimé mais il est nécessaire de prendre en compte l'antécédent pour interpréter le sujet du verbe. Comme le remarque Noailly (1997, p. 99), il s'agit d'une place vide « fonctionnant comme facteur de cohésion discursive au même titre qu'une représentation pronominale ordinaire ». Dans l'exemple [18] présenté plus haut, le pronom « Je » est aussi anaphorisé par une forme vide. Noailly (1997, p. 107) souligne que ce vide n'est pas à confondre avec l'ellipse discursive car il ne s'agit pas d'une « simple facilité de parole ». Dans les cas d'anaphore zéro, bien que l'anaphorique ne soit pas présent dans le texte, le référent est bien le même que celui de son antécédent. La notion de « maillon faible » sera donc potentiellement utilisée au cours de cette thèse si des cas

propices à s'y conformer se présentent.

Pour conclure, une chaîne de coréférence est une notion linguistique qui caractérise l'ensemble des expressions référentielles désignant le même référent dans le discours. Ces expressions référentielles sont donc coréférentes et peuvent même avoir une interprétation anaphorique, bien que la coréférence n'implique pas nécessairement l'anaphore, et inversement. Une chaîne de coréférence permet le suivi de l'évolution du référent dans le discours, même lorsqu'elle ne contient que deux maillons.

Pour qu'il y ait coréférence, le référent doit être exactement le même pour plusieurs expressions référentielles. Nous ne prendrons pas en compte les cas d'anaphores sans coréférence. Cependant, nous nous intéresserons aux cas dans lesquels il peut y avoir coréférence mais où cette relation n'est pas complètement stricte de manière certaine, voire même floue.

Chapitre 2

Annoter un corpus en coréférence : une première expérience

Maintenant que la notion de chaîne de coréférence ainsi que les phénomènes linguistiques associés ont été définis, nous nous attacherons dans ce chapitre à traiter du problème de l'annotation. L'annotation des chaînes de coréférence est l'une des tâches réalisées dans le projet Democrat. Nous aborderons les principes et inconvénients de cette tâche ainsi que le traitement des cas particuliers et les questions linguistiques écartées au cours de ce processus. L'annotation réalisée selon le manuel d'annotation du projet Democrat a entraîné quelques difficultés qui seront aussi discutées. Une fois ce travail d'annotation réalisé, il faut pouvoir exploiter ces informations. Pour ce faire, l'outil TXM a été choisi pour Democrat, ce qui a permis de grandes avancées. Il reste cependant un manque de méthodologie pour l'analyse des chaînes de coréférence.

2.1 Le projet Democrat

Le projet Democrat a pour thème principal les chaînes de coréférence. Cette section présentera donc le projet et précisera le rôle de la thèse dans ce cadre précis.

2.1.1 Présentation

Le projet Democrat [1] a été financé par l'Agence Nationale de la Recherche pour quatre ans de 2016 à 2020. Il est le fruit de la collaboration entre des chercheurs issus de quatre laboratoires français situés respectivement à Paris, Strasbourg et Lyon : Lattice [2], LiLPa [3], ICAR [4] et IHRIM [5]. L'acronyme Democrat signifie : DEscription et MOdélisation des

1. http://www.lattice.cnrs.fr/democrat/
2. http://www.lattice.cnrs.fr/
3. http://lilpa.unistra.fr/
4. http://icar.univ-lyon2.fr
5. http://ihrim.ens-lyon.fr/

Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique.

À travers l'étude des chaînes de coréférence, ce projet aspire à développer les recherches sur la langue et la structuration textuelle. À travers la publication d'un corpus d'envergure annoté en coréférence, le projet avait pour but dès le départ d'aider à créer un modèle intégré et discursif de la référence et plus précisément de la construction des chaînes de coréférence. Ce corpus a aussi été pensé pour être un corpus de référence et d'apprentissage pour les campagnes d'évaluation sur la coréférence. L'objectif de ce projet est en outre de fournir un outil d'annotation et de manipulation des données annotées via la plateforme TXM (Heiden, Magué et Pincemin 2010a) ainsi qu'un système de détection automatique des coréférences. Ainsi, le projet Democrat a pour but de faire le pont entre les domaines de la linguistique et du traitement automatique des langues en passant par la linguistique de corpus outillée.

Corpus et modélisation

Ces objectifs ont été motivés par l'absence de travaux de cette ampleur sur la coréférence en français écrit. Salmon-Alt (2002) formulait déjà cette remarque concernant les corpus français annotés en relations anaphoriques, estimant que les ressources disponibles à cette époque étaient « insuffisantes, tant au niveau quantitatif qu'au niveau qualitatif ». Pour introduire le projet ANANAS[6], elle récapitulait les corpus disponibles en 2002 :

Auteurs	Nombre de Mots	Nombre d'expressions
(Bruneseaux et Romary 1997)	30 000	3 359
(Popescu Belis 1999)	10 000	638
(Clouzot, Antoniadis et Tutin 2000)	95 000	1 316
(Tutin et al. 2000)	1 000 000	?
(Salmon-Alt 2001)	11 000	1 344
(Trouilleux 2001)	45 000	886

Tableau 2.1 – Ressources disponibles pour l'anaphore en français en 2002 - issu de (Salmon-Alt 2002)

Le corpus de Salmon-Alt (2001) est le seul dans lequel la coréférence n'est pas annotée. Les annotations concernent un certain type d'anaphores associatives. Les autres corpus du tableau sont annotés en coréférence bien que pour certains, la relation d'ana-

6. Annotation Anaphorique pour l'Analyse Sémantique de Corpus

phore associative soit aussi annotée en plus. Le corpus de Tutin et al. (2000) est de taille conséquente, cependant les descriptions définies [7] ne sont pas annotées et le corpus n'est pas en accès libre pour la recherche. Une autre initiative a vu le jour à travers le corpus DEDE [8] (Gardent et Manuélian 2005) qui reste modeste (48 360 mots) et dans lequel les pronoms ne sont pas annotés.

Par la suite, des travaux sur la coréférence en français (Schang, Boyer-Pelletier et al. 2011) ont donné lieu à un corpus annoté en coréférences et anaphores : le corpus ANCOR [9] (Muzerelle, Lefeuvre, Antoine et al. 2013 ; Muzerelle, Lefeuvre, Schang et al. 2014), qui contient 488 000 mots. Il s'agit d'oral transcrit et les données correspondent à de la parole conversationnelle, donc spontanée, issue de trois corpus préexistants dont ESLO [10] (Eshkol-Taravella et al. 2011) et deux corpus provenant de Parole Publique (Nicolas et al. 2002) : OTG [11] et Accueil UBS [12]. Plus récemment, d'autres corpus de grande taille et de langues différentes ont vu le jour, notamment grâce à des initiatives comme Ogrodniczuk et Ng (2016) et Ogrodniczuk et Ng (2017).

Tous ces corpus provenant de projets différents laissaient une place vide qu'est venu occuper le corpus Democrat (Landragin 2019) pour traiter de la coréférence en français écrit avec une masse de données suffisante pour servir à la modélisation linguistique et au traitement automatique des langues. Ce corpus d'environ 580 000 mots regroupe 58 blocs de texte en français de 10 000 mots chacun. Ces textes ont été annotés en coréférence, donnant environ 200 000 mentions annotées. Ils peuvent être narratifs ou non et de genres textuels différents (roman, fable, biographie, article de presse, article encyclopédique, texte juridique, etc.). Ils peuvent aussi dater de périodes différentes, allant du 11ème au 21ème siècle, ce qui permet des études sur le plan diachronique.

Principalement à cause de l'absence de corpus tels que Democrat jusqu'à présent, il n'existe pas encore de modélisation formelle ni de réelle typologie des chaînes de coréférence. Les études sur ce sujet se sont beaucoup basées jusqu'ici sur des maillons isolés ou des maillons-clés. Il existe néanmoins les études de Givón (1983) et Givón (1989) qui traitent des *topiques continus* et *discontinus* mais cette modélisation ne prend pas en compte « les états de langue anciens », dont certains utilisaient principalement les répétitions (Glikman, Guillot-Barbance et Obry 2014), ni les particularités de divers genres textuels (Schnedecker 2014). La publication du corpus Democrat et les différentes publications du projet serviront de point de départ à une modélisation plus formelle et potentiellement implémentable dans TXM.

7. Phénomène défini en 1.1.2.
8. Un corpus annoté pour le traitement des DEscriptions DEfinies :
9. ANaphore et Coréférence dans les Corpus ORaux
10. Enquête SocioLinguistique à Orléans
11. Office du Tourisme de Grenoble
12. Université Bretagne-Sud

L'annotation : quel(s) outil(s) pour Democrat ?

Avant le projet Democrat, il n'existait pas d'outil permettant de visualiser, explorer et analyser des corrélations au sein des chaînes de coréférence, le tout diffusé largement et maintenu régulièrement. Différents outils existaient déjà mais les fonctionnalités n'étaient pas toutes présentes dans chacun d'entre eux.

Le logiciel CADIXE [13] (Bessières, Nazarenko et Nédellec 2001) permet d'annoter des entités mais pas les relations anaphoriques. MMAX2 (Müller et Strube 2006), dont l'interface est peu ergonomique pour l'annotation, permet bien d'annoter les relations mais ne possède pas de fonctionnalité de représentation et d'analyse des chaînes. Le logiciel GLOZZ (Widlöcher et Mathet 2009) permet quant à lui d'annoter les entités et les relations. Il permet aussi une représentation visuelle des chaînes mais nécessite de passer par GLOZZQL [14], un langage de requête peu intuitif permettant des recherches dans le corpus. GLOZZ fonctionne convenablement sur des textes courts. Cependant, les textes du projet Democrat (environ 10 000 mots par bloc de texte) nécessitaient un outil plus robuste.

Au cours du projet Democrat, Oberle (2018) a développé SACR, un outil permettant l'annotation simple et ergonomique de la coréférence (via drag-and-drop). Il ne permet pas de visualiser et analyser les chaînes. Un autre outil du même auteur, intitulé CRViewer [15], est en cours de développement pour cette fonctionnalité, ce qui donne une idée des efforts déployés actuellement pour ce champ d'études

L'outil ANALEC (Landragin, Poibeau et Victorri 2012) a été développé au Lattice avant le projet Democrat. Il a principalement été utilisé par des chercheurs du laboratoire, notamment au cours du projet MC4. Il permet l'annotation des entités et des relations mais aussi l'analyse et la représentation des chaînes. Cet outil n'est pas utilisé à grande échelle et n'est plus maintenu depuis plusieurs années.

TXM (Heiden, Magué et Pincemin 2010a) est une importante plateforme de manipulation de corpus largement diffusée et utilisée notamment dans de nombreux projets. Le projet Democrat a donc travaillé avec l'équipe de TXM afin de permettre l'utilisation des fonctionnalités d'ANALEC de manière robuste. TXM comporte maintenant l'extension « URS » [16] qui implémente une partie d'ANALEC et permet donc l'annotation des entités et des relations ainsi que l'analyse et la visualisation des chaînes.

13. Catégorisation Automatique de Documents pour l'Extraction de Réseaux d'Interactions GEniques
14. QL = Query language (langage de requête)
15. Il est disponible sous la forme d'un prototype sur le site web de l'auteur : https://boberle.com/projects/coreference-analysis-with-crviewer/
16. Le schéma URS (Unité - Relation - Schéma) est développé dans la section 2.2.1.

Les systèmes de Traitement Automatique des Langues (TAL)

La tâche de détection automatique de la coréférence est en constante évolution et de nouveaux systèmes sortent régulièrement. De nombreux systèmes utilisent des méthodes d'apprentissage profond, avec ou sans l'aide de ressources externes (Lee, He, Lewis et al. 2017) qui nécessiteraient un pré-traitement des données comme de l'étiquetage morphosyntaxique par exemple. Le volet TAL du projet vise à présenter deux systèmes de détection de la coréférence s'inspirant de ces systèmes appliqués au français : COFR (Wilkens et al. 2020) et DECOFR (Grobol 2020).

2.1.2 Rôles de la thèse dans le projet Democrat

Ce travail a été financé par le projet Democrat suite à un appel à candidature pour un contrat doctoral pour une thèse intitulée « Analyse en corpus de chaînes de coréférence ». Les objectifs ont été précisés au cours de la thèse jusqu'à l'ajout du sous-titre « La (co)référence floue à l'épreuve de la linguistique outillée ».

L'objectif de cette thèse est de fournir une analyse linguistique des chaînes de coréférence et plus particulièrement du flou référentiel au sein de celles-ci. Afin de prendre la mesure de ce phénomène, la première étape a été l'annotation manuelle de blocs de textes pour le corpus Democrat. Cette phase d'annotation développée dans la partie 2.3.2 a été l'occasion de nombreux questionnements à propos des expressions référentielles à annoter comme coréférentes ou non, malgré un manuel d'annotation déjà bien fourni. C'est principalement cette phase qui a permis de préciser l'orientation de la thèse et de faire émerger le sujet de la (co)référence floue qui sera développé dans les parties 2 et 3.

Au cours de cette thèse, l'étude des chaînes de coréférence dans le contexte du projet Democrat a permis de mettre en lumière des problèmes de référence et de coréférence, souvent déjà traités dans la littérature comme les référents évolutifs par exemple. Cette thèse est donc l'occasion de faire le point sur la coréférence non stricte et floue, notamment en mettant l'accent sur des cas typiques comme les groupes pluriels ou les pronoms problématiques référentiellement comme « on » et « ce ».

Cette thèse a aussi été l'occasion de travailler sur la coréférence floue et le pronom « on » au niveau psycholinguistique avec Lucie Rousier-Vercruyssen, post-doctorante en 2019 sur le projet Democrat. Ce travail a été réalisé en collaboration et a abouti à une expérience développée dans la section **??** à propos de l'interprétation du référent du pronom « on » dans différents contextes.

L'analyse de ces phénomènes a permis de faire une revue critique du manuel d'annotation du projet Democrat. Pour ensuite proposer un schéma d'annotation adapté (présenté

dans la section 6.1.3) qui puisse prendre en compte le phénomène de la coréférence floue qui pose bien des difficultés aux annotateurs.

Le problème de l'accord inter-annotateurs a été soulevé au cours du projet Democrat. Afin d'apporter de l'aide dans cette tâche d'évaluation à Marine Le Mené, post-doctorante en 2018 sur le projet Democrat, j'ai aussi annoté manuellement deux textes de 2 000 mots environ. Cette tâche est développée dans la section 2.3.2. Cette thèse a abordé l'intégralité des aspects attendus lors de la conception d'un corpus de référence.

2.2 Questions initiales et contraintes d'annotation

Dans cette section nous nous pencherons sur la question de l'annotation. Nous verrons quels en sont les intérêts, les enjeux et les principes. L'annotation de la coréférence est un cas particulier car elle implique l'annotation d'unités mais aussi de relations. Nous nous pencherons sur le modèle « URS » qui fournit une solution pour cette tâche. Pour l'annotation de la coréférence, il est aussi nécessaire de poser un cadre et de définir quels phénomènes annoter et comment.

2.2.1 Annoter en coréférence : principes et inconvénients

Annoter les coréférences est une tâche complexe d'annotation. Nous présenterons dans cette section les principes de l'annotation et les particularités de cette tâche pour la coréférence.

Intérêts et enjeux de l'annotation

Selon la définition de Leech (2005) : « *Corpus annotation is the practice of adding interpretative linguistic information to a corpus.* »[17]. Fort, Nazarenko et Claire (2011) élargissent cette définition en caractérisant l'annotation d'ajout d'une « interprétation sous la forme d'une note dans un flux de données ». En effet, l'information ajoutée n'est pas nécessairement linguistique. Le terme d'annotation désigne aussi souvent l'information annotée.

Certains auteurs sont réfractaires au principe même d'annotation. Sinclair (2004), redoute que les chercheurs analysent leurs données uniquement à travers le prisme de leurs annotations. Il pense que l'ajout de cette information linguistique représente une « activité périlleuse » qui entraîne une perte d'« intégrité » du texte. Hunston (2002) note

17. Traduction : *L'annotation de corpus consiste à ajouter une information linguistique à un corpus.*

aussi que le choix des catégories annotées est « typiquement déterminé avant toute analyse du corpus ». Pour utiliser l'annotation de manière consciencieuse, il est effectivement nécessaire de rappeler qu'annoter un corpus nécessite de faire des choix. En effet, comme le rappelle Habert (2000), « Quelle que soit sa richesse, une annotation est cependant toujours orientée par une tâche, même si cela est implicite ». Apposer des catégories sur un corpus correspond donc à une interprétation à visée linguistique[18] ou « taliste »[19]. Une annotation est d'autant plus utile si elle a été « déterminée pour être spécifique à une application particulière » (Leech 2005). En prenant en compte ces dernières remarques, les annotations peuvent représenter un enrichissement du texte très utile fournissant ainsi des *ressources* pour plusieurs domaines.

Annoter : quoi et comment ?

Il est possible de créer ces ressources selon différentes méthodes d'annotation. Eshkol-Taravella (2015) distingue trois types d'annotation correspondant à des applications distinctes. Le premier type d'annotation consiste à ajouter des remarques sur le texte. Le second correspond à l'annotation du document (ou du corpus) pour le décrire et le caractériser à travers les métadonnées. Le troisième correspond aux annotations linguistiques des données contenues dans le corpus. C'est ce dernier type d'annotation qui nous intéressera pour l'annotation de la coréférence dans ce travail de thèse.

L'annotation peut se faire manuellement ou encore automatiquement, à l'aide de programmes. L'annotation manuelle représente un coût humain et temporel avec un danger d'incohérences ou de désaccord entre les annotateurs lorsqu'il y en a plusieurs. En revanche, elle permet aussi d'accéder à une expertise que des programmes ne possèdent pas (encore). L'annotation manuelle est souvent la première étape qui permet de créer un corpus annoté qui servira de modèle d'entraînement pour l'apprentissage d'étiqueteurs automatiques. Lorsque les annotations d'un corpus sont vérifiées, il peut aussi servir à l'évaluation de systèmes automatiques, on parle alors de « corpus de référence » (Poudat et Landragin 2017), *gold standard* en anglais. Le danger de l'annotation automatique réside plutôt dans la circularité des résultats produits. C'est pourquoi les outils sont en permanente évolution. Une troisième solution d'annotation est possible pour trouver un compromis entre la précision de l'humain et la performance de la machine : l'annotation semi-automatique. Cette méthode implique l'intervention d'un utilisateur afin de valider les décisions prises par le système.

L'annotation manuelle peut se faire à partir de textes bruts ou déjà partiellement

18. « La linguistique de corpus peut ainsi être objective, mais non objectiviste, puisque tout corpus dépend étroitement du point de vue qui a présidé à sa constitution. » (Rastier 2005)

19. Relatif au Traitement Automatique des Langues (exemple : annoter un corpus qui puisse servir d'entraînement à des systèmes de recherche et extraction d'informations.).

annotés (en parties du discours, relations syntaxiques, etc.). Cette pré-annotation représente souvent un gain de temps et de qualité mais peut aussi introduire des biais (Fort et Sagot 2010).

Pour l'annotation, certains projets ont trouvé des astuces afin de mettre des annotateurs à contribution de manière « séduisante ». Amazon a lancé en 2005 un service de micro-travail nommé *Amazon Mechanical Turk*. Il s'agit d'une plateforme web de production participative (*crowdsourcing*) qui permet de faire effectuer des tâches à des humains contre rémunération. Ces tâches peuvent être par exemple de l'annotation de texte ou d'images. Ces tâches étant souvent très peu rémunérées, ces méthodes posent des problèmes éthiques (Fort, Adda et Cohen 2011) et sont même souvent qualifiées d'« esclavage » (Whitla 2009).

Les jeux ayant un but (*Games With A Purpose - GWAP*) proposent une méthode d'annotation ludique. Par exemple, Zombilingo (Fort, B. Guillaume et Stern 2014) permet l'annotation en syntaxe de dépendance. Le but est d'identifier des relations de dépendance syntaxique et plus l'utilisateur annote de phrases correctement, plus il gagne de points. Pour l'annotation des anaphores, le corpus Phrase Detective (Chamberlain, Poesio et Kruschwitz 2008) a été annoté selon ce principe. La méthodologie comporte une phase de formation ainsi que des instructions détaillées pour l'annotateur. L'évaluation de ces annotations se fait selon un corpus de référence et l'accord inter-annotateurs est élevé (Chamberlain, Kruschwitz et Poesio 2009).

Le jeu d'étiquettes (ou de catégories) annotées est défini avant la tâche d'annotation. Ce jeu d'étiquettes et leur description sont répertoriés dans ce que l'on appelle un manuel ou un guide d'annotation. Le rôle de ce dernier est de définir précisément la méthode d'annotation ainsi que les éléments à annoter en fonction de leurs propriétés. Il propose généralement des exemples pour chaque catégorie afin que des annotateurs différents puissent annoter exactement de la même manière.

Pour vérifier que des annotateurs différents annotent bien de la même manière, il est possible de calculer l'accord inter-annotateurs afin de comparer les annotations. Les données annotées sont « fiables » (*reliable*) si les annotateurs s'accordent sur les catégories auxquelles doivent appartenir les unités (Craggs et Wood 2005). Artstein et Poesio (2008) ajoutent que la fiabilité est un pré-requis à la « validité » (*validity*) du schéma d'annotation [20] à propos d'un phénomène. Un bon accord inter-annotateurs n'implique donc pas que le schéma soit valide mais simplement que les annotateurs sont en adéquation et annotent selon les mêmes principes (Mathet et Widlöcher 2016).

20. Le schéma d'annotation contient l'ensemble des annotations possibles.

L'annotation de la coréférence

Il existe deux méthodes fondamentales pour annoter manuellement le résultat de la résolution de la coréférence. La première consiste à relier les mentions entre elles (ou à l'antécédent pour les chaînes anaphoriques, comme c'est le cas dans le corpus ANCOR (Muzerelle, Lefeuvre, Antoine et al. 2013) par exemple). La seconde revient à identifier le référent de chaque mention puis à créer automatiquement les chaînes. Cette deuxième option, plus rapide (Landragin, Potier et Bothua 2017), a été choisie dans le projet Democrat par exemple. Pour cette tâche, la première étape a pour but l'annotation des expressions référentielles en en identifiant le référent. Chaque mention d'un référent est associée à celui-ci comme dans la figure suivante :

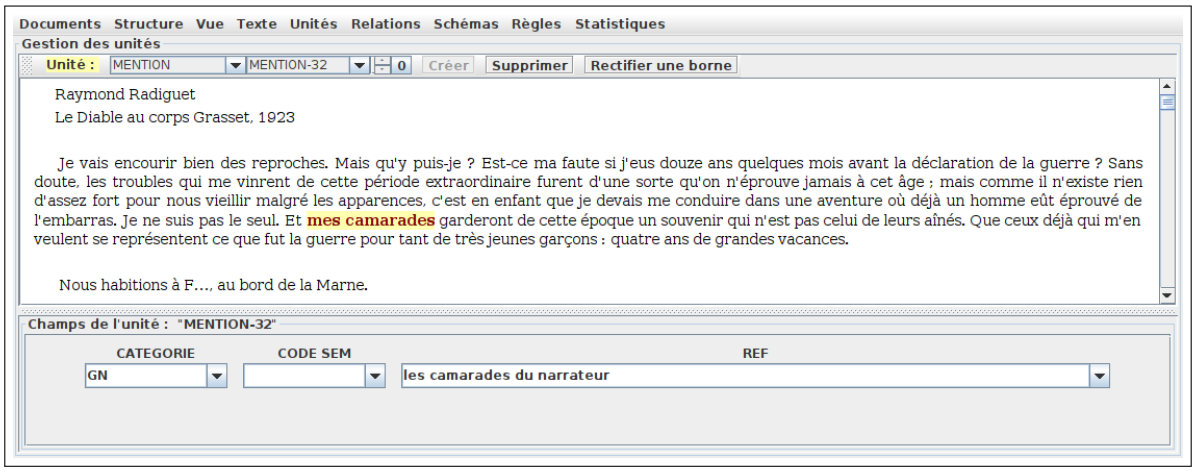


Figure 2.1 – Une mention associée à son référent dans Analec.

Dans la figure [2.1], le syntagme nominal « mes camarades » est une mention qui réfère aux camarades du narrateur. Il est généralement recommandé aux annotateurs de choisir un nom explicite et singularisant pour le référent au moment de l'annotation. La « catégorie » est remplie dans cette figure, cependant cette annotation n'est pas publiée dans le corpus Democrat courant 2019 car elle n'a pas systématiquement fait l'objet d'une vérification.

Pour les syntagmes nominaux, il est souvent d'usage d'annoter le groupe complet en incluant déterminants et adjectifs. Par exemple, on annotera « cette période extraordinaire » et pas simplement « période » ou « cette période ». C'est au schéma d'annotation de préciser quelles expressions seront annotées ou non. La seconde étape revient à établir les relations entre les mentions coréférentes. Il s'agit là encore du rôle du manuel d'annotation de préciser quelles sont les relations à annoter ou non.

Différents schémas ont été pensés pour l'annotation de la coréférence, sans qu'il n'y ait de règles universelles. Nous discuterons de certains d'entre eux, comme le schéma UCREL, qui a été développé pour le *Lancaster Anaphoric Treebank* (Fligelstone 1992 ;

Garside 1997) et permettait l'annotation des relations anaphoriques. Le schéma MUCCS développé pour MUC-7 (Hirschman et Chinchor 1997) se concentre sur les relations de coréférence entre les syntagmes nominaux et a été élaboré pour de l'anglais écrit. Il ne peut donc pas aider à traiter les cas problématiques rencontrés dans les corpus oraux comme les disfluences ou les références à la situation visuelle, ni les cas particuliers liés à d'autres langues comme les clitiques en français. Bien que ce schéma ait été critiqué (Van Deemter et Kibble 1999, p. 92) car il recommandait d'inclure dans les chaînes de coréférence des mentions non référentielles, il représente un premier effort de solution standardisée. Avec une volonté de standardisation également, la *Discourse Resource Initiative* (DRI) a ensuite proposé des recommandations pour l'annotation des anaphores, coréférentes ou non, à travers le schéma DRAMA, développé par Passonneau (1997). Ce schéma s'inspire, entre autres, de MUCCS en étant moins centré sur certains domaines et en se reposant davantage sur des critères linguistiques (Poesio, Stuckardt et Versley 2016). En plus des expressions référentielles classiques, ce schéma comprend l'annotation de syntagmes verbaux et adjectivaux comme antécédents de certaines anaphores. Il contient également des recommandations pour les entités « marquables » dans les dialogues. L'innovation du schéma proposé par Bruneseaux et Romary (1997) est l'annotation d'informations déictiques comme des références au contexte visuel par le pointage.

Les formats des schémas présentés précédemment varient, allant du langage SGML au XML TEI. De plus, ils possèdent chacun une idée différente des relations à annoter : la coréférence exclusivement pour MUCCS, les anaphores associatives en plus pour DRAMA, ainsi que pour Bruneseaux et Romary (1997), qui y ajoutent la référence au contexte visuel et la déixis du discours[21]. Le schéma de Lancaster prend aussi en compte les anaphores elliptiques. Face à la diversité des schémas d'annotation de la coréférence et de l'anaphore, le projet MATE (Mengel et al. 2000) a proposé une synthèse des schémas existants aboutissant à des recommandations pour l'annotation des coréférences (Poesio, Bruneseaux et Romary 1999) se matérialisant par un méta-schéma (Poesio 2000). Ce méta-schéma regroupe donc les propriétés des schémas présentés précédemment, qui peuvent être instanciées en fonction des besoins de l'utilisateur (cadre théorique, langue étudiée, etc.). Il est composé d'un noyau, utilisable pour la coréférence des syntagmes nominaux comme MUCCS, et de trois extensions. La première extension prend en compte la référence au contexte visuel. La seconde permet d'annoter les relations complexes comme les anaphores associatives. La troisième inclut un plus grand nombre d'antécédents possibles comme les clitiques ou la déixis.

Le méta-schéma proposé par MATE a été affiné et utilisé par la suite pour le corpus

21. Lorsque la référence est liée au contexte. Par exemple, si l'antécédent est un événement ou une action.

GNOME (Poesio 2004). Il a aussi aiguillé l'annotation de l'anaphore et de la coréférence pour des corpus comme ARRAU (Poesio et Artstein 2008), LiveMemories (Rodriguez et al. 2010), AnCora (Taulé, Marti et Recasens 2008) ou encore le Prague Dependency Treebank (Nedoluzhko et al. 2016). Les solutions proposées par MATE et GNOME sont particulièrement utiles pour l'annotation des relations anaphoriques, mais moins pour la coréférence. Le corpus Ontonotes (Pradhan et al. 2011), le Polish Coreference Corpus (Ogrodniczuk, Glowinska et al. 2014), ANCOR (Muzerelle, Lefeuvre, Antoine et al. 2013) et DEMOCRAT ne l'ont pas appliqué par exemple.

Le corpus Democrat utilise le format XML-TEI-URS (Grobol, Landragin et Heiden 2018), inspiré du méta-modèle URS [22] de GLOZZ et implémenté dans le format XML-TEI conçu par Grobol, Landragin et Heiden (2018) puis intégré à TXM (Heiden, Magué et Pincemin 2010b). L'extension de TXM se nomme « URS » en référence au modèle choisi pour l'annotation dans le projet. L'acronyme URS correspond aux initiales des mots : *Unité*, *Relation* et *Schéma*. Ce sont les trois types d'éléments contenus dans ce modèle. Il a initialement été développé pour le logiciel GLOZZ puis implémenté par la suite dans Analec et par extension dans TXM.

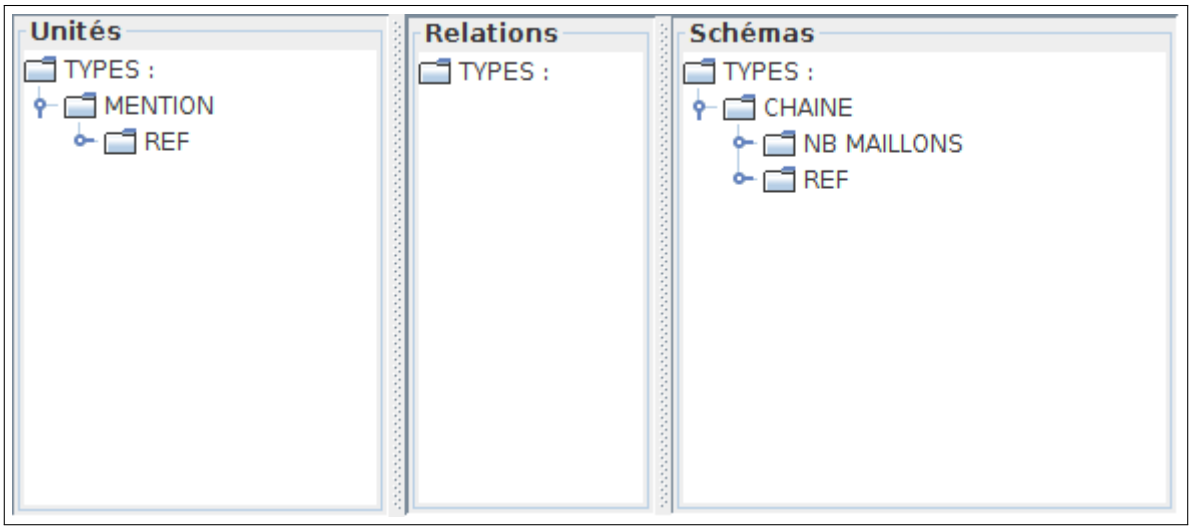


Figure 2.2 – Le schéma URS de Democrat dans TXM.

Une unité est une séquence contiguë de mots. Pour les chaînes de coréférence, cela correspond aux maillons. Une unité peut avoir plusieurs traits. Dans la première version de Democrat, il n'y en a qu'un seul : le champ « REF » qui est rempli manuellement par les annotateurs et qui correspond au nom du référent. D'autres traits qui peuvent être annotés automatiquement avant ou après l'annotation manuelle ont été envisagés dans le projet, comme les parties du discours. Ils sont cependant absents de la première version du

22. URS est un méta-modèle qui s'applique à toute une famille de schémas d'annotation (sauf ceux qui nécessitent des champs multivalués ou des champs conditionnels).

corpus. Le deuxième type d'élément correspond aux relations binaires entre les éléments. Elles n'ont pas été construites dans la version actuelle du corpus Democrat car elles n'étaient pas nécessaires. Il est toutefois possible de les créer automatiquement. Le dernier élément correspond aux schémas. Il correspondent aux chaînes de coréférence car il s'agit d'ensembles d'éléments URS. Pour Democrat, les traits ont été remplis automatiquement lors de la création des chaînes : le nom du référent (repris directement du champ REF des mentions) et le nombre de maillons (décompté au passage). L'idée d'annoter manuellement certains traits des chaînes comme le genre du référent, le nombre ou encore le type, a été considérée au début du projet puis laissée de côté.

La pré-annotation automatique peut constituer un gain d'efficacité dans la tâche d'annotation manuelle. Pour les chaînes de coréférence, appliquer un système de détection automatique de la coréférence sur le texte à annoter permettrait à l'annotateur de n'avoir simplement qu'à vérifier l'annotation produite. Des tests ont été effectués en vue de l'annotation du corpus Democrat par Landragin, Potier et Bothua (2017) avec le chunker SEM (Tellier, Y. Dupont et Courmet 2012). Le seul système disponible pour la détection de la coréférence en français n'étant pas suffisamment performant (Todirascu 2001), il a donc été décidé d'utiliser l'outil performant qui s'en rapproche le plus et qui permet de détecter automatiquement les *chunks* dans un texte. Un chunk correspond à la plus petite suite d'unités linguistiques possible formant un groupe avec une tête forte, sans être discontinue ou récursive (Abney 1991). Compte tenu des réflexions abordées dans la section 1.2.3, tous les chunks nominaux ne réfèrent pas nécessairement. Il est néanmoins préférable de partir d'une pré-annotation trop large plutôt que trop limitée comme cela aurait été le cas avec des entités nommées par exemple [23]. Il faut aussi prendre en considération le fait qu'une pré-annotation peut influencer l'annotateur (Fort, Ehrmann et Nazarenko 2009). En effet, des chunks nominaux non référentiels peuvent induire l'annotateur en erreur, comme avec les « il » impersonnels qu'il pourrait être tenté d'annoter par habitude. Le deuxième désavantage des chunks est la non détection de l'enchâssement. Plusieurs expressions référentielles peuvent être enchâssées, ce qui n'est pas le cas des chunks par définition.

Dans les figures [2.3] et [2.4], la pré-annotation a été réalisée avec SEM et fournit un document de travail dans lequel les groupes susceptibles d'être des expressions référentielles sont déjà annotés avec leur partie du discours. Dans la figure [2.3], le chunk « une aventure » est pré-annoté :

23. Comme nous l'avons déjà précisé en 1.2.1, les entités nommées recouvrent une partie limitée des expressions référentielles.

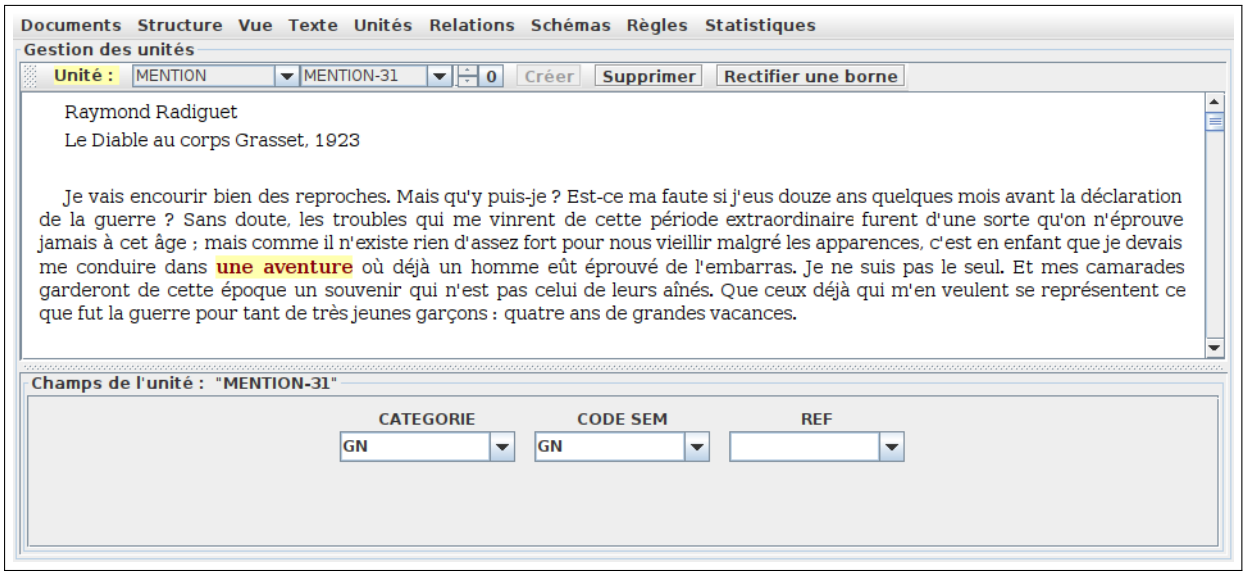


Figure 2.3 – Une pré-annotation en chunks par SEM dans Analec.

Dans la figure [2.4], le champ REF n'est pas rempli car cette figure présente l'étape précédant l'annotation manuelle. C'est le rôle de l'annotateur humain de renseigner le référent de chaque expression référentielle. Dans cette figure, nous retrouvons le problème de l'enchâssement lié aux chunks. L'expression référentielle « la déclaration de la guerre » en contient une autre : « la guerre ». Pour annoter ces expressions dans Democrat, l'annotateur doit donc modifier les bornes manuellement. De plus, tout outil automatique, même performant, peut produire des erreurs. À la ligne 4 par exemple, la mention « où » n'est pas pré-sélectionnée. Ces cas particuliers peuvent prendre plus de temps à corriger que le temps que prendrait une annotation manuelle à partir de zéro. C'est pourquoi la pré-annotation en chunks, après avoir été laissée au libre choix de l'annotateur, n'a finalement pas été retenue pour l'annotation dans Democrat.

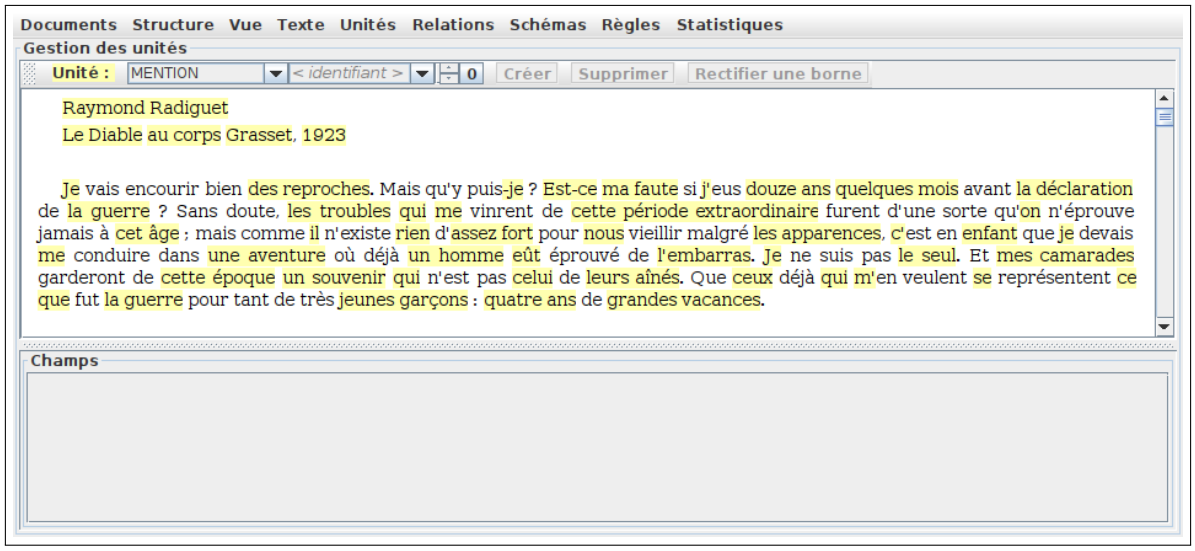


Figure 2.4 – Pré-annotation avec SEM pour l'annotation avec Analec. Les chunks ne sont pas tous des expressions référentielles.

2.2.2 L'annotation des cas particuliers

Chaque projet implique de spécifier un cadre théorique et technique bien défini. Cela nécessite de faire des choix à propos des phénomènes linguistiques à prendre en compte ou non, ainsi que de la modalité de traitement de ces phénomènes. Comme nous l'avons abordé précédemment, certains projets ont décidé d'annoter des relations anaphoriques non coréférentes. Dans le projet Democrat comme dans cette thèse, le cadre s'arrête aux relations de coréférence. Même dans ce cadre, l'annotation de certains éléments méritent une réflexion.

Les expressions référentielles qui ne sont pas reprises, les *singletons*, sont annotées ou non en fonction des projets. MUC6 (Grishman et Sundheim 1995) et Ontonotes (Pradhan et al. 2011) ne les annotent pas par exemple. En revanche, dans le projet Democrat il a été décidé de les annoter en raison de leur utilité pour le traitement automatique mais aussi car cette tâche constitue un gain de temps et d'efficacité (Landragin, Potier et Bothua 2017). En effet, il n'est pas évident de savoir si une expression référentielle sera reprise au moment où elle est annotée. Si elle est reprise dix pages plus loin, est-ce que l'annotateur se souviendra qu'elle a déjà été introduite ? Le simple fait de se poser la question est chronophage au niveau de l'annotation.

Comme pour le corpus ANCOR, dans le projet Democrat, il a été décidé d'annoter la totalité des syntagmes ainsi que les structures coordonnées car chacun de ces éléments a le potentiel de faire l'objet d'une reprise référentielle comme dans l'exemple [39] où « Noël et Tou » est repris mais « Noël » seul l'est aussi. En revanche, si le groupe est

exclusif : s'il y a un « ou » à la place du « et » par exemple, le groupe entier ne sera pas annoté dans Democrat. De plus, si une expansion du nom est mise en facteur comme dans l'exemple [37], cette expansion ne peut pas être inclue dans la mention du premier syntagme nominal, bien que l'expansion s'applique aussi à ce syntagme. Si les antécédents sont au contraire dispersés comme dans l'exemple [38], ils ne peuvent pas être annotés comme un référent unique pour des raisons techniques compte tenu de la linéarité du texte car il faudrait pouvoir exclure le verbe « retrouva » de la mention.

Exemple [37]
« [[**L'achat**]$_i$ et [**la démolition d'**[**une maison**]$_j$]$_k$]$_l$ »
Exemple issu du manuel d'annotation du projet DEMOCRAT.

Exemple [38]
« [**Pierre**]$_i$ retrouva [[**sa**]$_i$ **femme**]$_j$ au restaurant. [**Ils**]$_k$ dînèrent jusqu'à tard dans [**la nuit**]$_l$ ».
Exemple issu du manuel d'annotation du projet DEMOCRAT.

Exemple [39]
« [[**Noël**]$_j$ **et** [**Tou**]$_k$]$_i$ menaient au contraire un vrai tapage, et le rire éclatant de Douce ne s'arrêtait guère. Toute crainte s'était éloignée d'elle. Qui donc pourrait lui faire du mal aux côtés [[**du garçon**]$_j$ **et** [**du chien**]$_k$]$_i$? Les quelques gamins qui s'approchaient encore pour dire des injures et jeter des pierres n'y revenaient pas, ayant appris qu'il ne faisait pas bon se mesurer avec [**ce Noël Barray**]$_j$ »
Raymond RADIGUET, *Le Diable au corps*, 1923.

Cet exemple soulève aussi le problème des prépositions. Dans Democrat, elles ne sont inclues dans les mentions que dans les cas d'amalgame comme dans « [la patte [du chien]] », sinon elles ne le sont pas comme dans « [la patte de [la chienne]] ». La deuxième mention de Noël et de Tou inclut par conséquent « du » car ce dernier a amalgamé le déterminant.

Les déterminants n'étant pas référentiels, ils ne sont pas annotés dans le corpus Democrat, à l'exception des possessifs car ils encodent le possesseur. Le référent du déterminant possessif est donc le possesseur comme dans l'exemple [12] : « [Je]$_j$ suis contente que [tu]$_t$ [lui]$_l$ aies dit que cette robe était [ma]$_j$ robe préférée. ». En ce qui concerne les relatives, dans le corpus Democrat, elles sont annotées par le biais du pronom relatif. Elles ne font pas partie de l'expression référentielle nominale, le pronom relatif est annoté seul comme dans l'exemple [13] : « J'ai cuisiné [le plat]$_p$ [que]$_p$ tu préfères ».

Dans le cas des anaphores résomptives comme dans l'exemple [33], l'antécédent n'est

pas annoté dans Democrat. En revanche, l'anaphorique est annoté car il peut être coréférent à d'autres expressions référentielles. Il sera donc considéré comme étant le premier maillon de la chaîne, sinon il fera partie des singletons.

L'annotation des sujets zéro pose aussi des problèmes techniques d'annotation. Comment annoter l'absence d'une expression ? Faut-il rajouter un nouvel élément au texte comme dans l'exemple [36] et l'annoter ? Dans Democrat, lorsque le sujet d'un verbe n'est pas présent, dans une coordination ou à l'impératif, il est annoté sur le verbe dont il est le sujet comme dans l'exemple [40]. Cette solution permet de ne pas apporter de modification au texte bien que ce ne soit pas directement le verbe qui réfère mais son sujet non exprimé.

Exemple [40]
« [**Ces braves**]$_i$ arrivèrent enfin et [**fendirent**]$_i$ la foule. »
Raymond RADIGUET, *Le Diable au corps*, 1923.

Pour le corpus ANCOR, Muzerelle, Schang, Antoine, Eshkol, Maurel, Boyer et al. (2012) ont décidé d'annoter les formes explétives de « il » [24] avec une catégorie particulière car ces usages non référentiels peuvent induire les systèmes en erreur et ils veulent pouvoir les identifier. Ils décident pourtant de ne pas annoter le pronom « ça » ainsi que ses dérivés. Dans Democrat, les « il » explétifs ne sont pas annotés alors que le pronom « ça » l'est.

Comme cela a été noté précédemment, le phénomène de l'anaphore associative ne relève pas de la relation de coréférence. C'est pourquoi ce phénomène n'est pas annoté dans le projet Democrat.

Les appositions ne sont pas inclues dans la mention. Pour simplifier le choix de l'annotateur, dans le projet Democrat le critère pour décider si une expression est une apposition est la présence de virgules. Elles ne sont donc pas annotées sauf s'il y a une coordination ou si elles imposent une discontinuité dans la mention.

Comme nous l'avons abordé avec l'exemple [11], les attributs nominaux peuvent référer. Cependant, dans Democrat ils ne sont pas annotés si deux chaînes qui ont été construites de manière indépendante se trouvent reliées par une construction attributive. Dans l'exemple [41], il y a donc deux chaînes de coréférence pour l'étoile du matin et l'étoile du soir alors qu'il s'agit objectivement du même astre. Cette distinction est opérée car ces constructions font exister deux référents différents dans l'esprit du locuteur dans un premier temps.

24. Comme dans « Il pleut ».

Exemple [41]
« [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]**$_m$]$_n$. »
Exemple issu du manuel d'annotation du projet DEMOCRAT.

La référence générique soulève aussi des questions de coréférence et donc d'annotation. Pour l'annotation du corpus Democrat, il était recommandé d'annoter les coréférences génériques dans une seule chaîne, même dans le cas d'un passage au singulier/pluriel ou de changement de déterminant comme dans l'exemple [42], car le référent est toujours générique. En revanche, une nouvelle chaîne doit être créée lorsque le référent générique n'est pas coréférent avec un référent spécifique ou particulier comme dans l'exemple [43].

Exemple [42]
« [**Les chats**]$_i$ miaulent. [**Un chat**]$_i$, [**ça**]$_i$ miaule. [**Le chat**]$_i$, quand [**il**]$_i$ a faim, miaule. Etc. »
Exemple issu du manuel d'annotation du projet DEMOCRAT.

Exemple [43]
« [**Il**]$_i$ a pris [**un café**]$_j$. [**Un café**]$_k$, [**ça**]$_k$ réchauffe toujours. »
Exemple issu du manuel d'annotation du projet DEMOCRAT.

Dans les cas où il y a une ambiguïté[25] entre deux référents potentiels pour une seule expression référentielle, il a été recommandé aux annotateurs de Democrat de créer un nouveau référent « A ou B ». Dans les cas de flou[26], la recommandation était de distinguer au mieux les référents potentiels et de ne pas nécessairement produire une seule chaîne de coréférence dans ce cas-là. L'objectif étant d'éviter de faire une seule chaîne contenant tous les « on » du texte. En revanche, pour les référents évolutifs[27], il était recommandé de n'annoter qu'une seule chaîne.

2.3 L'annotation réalisée : difficultés rencontrées

Cette thèse prend le manuel d'annotation du corpus Democrat comme point de départ pour l'établissement de son cadre théorique et technique. Nous aborderons les qualités et faiblesses du manuel d'annotation du projet Democrat ainsi que les travaux d'annotation effectués pour cette thèse.

25. Phénomène développé en 3.2.1.
26. Phénomène développé en 4.1.
27. Phénomène développé en 3.4.

2.3.1 Revue critique du manuel d'annotation

L'évolution du manuel

Le manuel Democrat a connu différentes évolutions entre mai 2016 et février 2018 [28]. Cette stratégie a permis aux annotateurs de faire remonter des remarques concernant ce manuel et de proposer éventuellement des corrections. Une première version est restée relativement stable au début du projet lors de l'annotation des textes littéraires. L'annotation de textes de genres variés comme les textes de loi ont fait remonter des imprécisions du manuel qui a fait l'objet d'une vague de corrections pour aboutir à une deuxième version courant 2017. Elle a ensuite continué d'évoluer jusqu'en février 2018.

Pour l'annotation, les annotateurs ont donc pu avoir à se référer schématiquement à deux versions principales différentes. D'un point de vue technique, la première version détaille l'utilisation d'Analec (Landragin, Poibeau et Victorri 2012) pour l'annotation. Il était le seul outil du projet à ce moment là. Depuis, il y a eu la migration d'Analec à TXM (Heiden, Magué et Pincemin 2010a) et la création de SACR (Oberle 2018), dont l'utilisation est mentionnée dans la deuxième version du manuel. La première version mentionnait aussi l'utilisation du chunker pour la pré-annotation dont l'utilisation était facultative mais pour laquelle il fallait vérifier la catégorie SEM. Dans la dernière version, il n'en est pas fait mention, plus personne n'utilisant SEM. En ce qui concerne la méthode d'annotation, la première version du manuel en proposait deux. D'une part, la méthode « systématique » qui consiste à annoter toutes les expressions référentielles, y compris les singletons. D'autre part, la méthode « pragmatique » consistant à n'annoter que les mentions reprises dans la suite du texte. Une distinction qui n'est plus proposée dans la dernière version du manuel, car c'est la méthode systématique qui a été unanimement adoptée.

Le choix des phénomènes à annoter a évolué avec le manuel lui aussi. En ce qui concerne les expressions référentielles, la première version recommandait de laisser de côté les impératifs dans le cas des sujets zéro alors que la dernière version les inclut. En cas d'ambiguïté, elle imposait le choix du référent le plus probable (sauf dans le cas d'une ambiguïté trop importante avec « l'un », « l'autre » ou « chacun » par exemple) alors que la dernière version propose directement la création d'un nouveau référent.

La première version du manuel d'annotation mentionnait l'annotation facultative de la structure textuelle du texte avec des éléments comme les titres et sous-titres, les liens hypertextes, les chapitres ou encore les notes de bas de page.

Ces différences de méthodologie ont pu être déstabilisantes pour certains annotateurs

28. https://www.lattice.cnrs.fr/democrat/files/ANR-15-CE38-0008-DEMOCRAT_livrable_methodo.pdf

qui ont annoté des textes à différents moments du projet. Cependant, les évolutions du manuel ont été bénéfiques car elles ont permis de préciser la méthodologie d'annotation du corpus Democrat. Le manuel a pris en importance au fil du temps en précisant de plus en plus de cas particuliers comme les termes d'adresse, les dates, les usages de mots *en mention*, les titres, les listes, les parenthèses, le cas de « c'est » ou encore des déterminants complexes. La version finale du manuel contient aussi beaucoup d'exemples présentés avec une notation claire.

2.3.2 Travail effectué et exemples rencontrés

Une des premières étapes de cette thèse et contribution au projet Democrat a été l'annotation de quatre blocs de textes de 10 000 mots en suivant les recommandations du manuel d'annotation développé par les membres du projet. Ce travail représente une contribution importante au projet car la plupart des annotateurs n'en a annoté qu'un ou deux. Il s'agit de quatre textes du 19ème au 21ème siècle, trois d'entre-eux sont des romans de type narratifs et le quatrième bloc correspond à des articles de presse du début de l'année 2003 parus dans l'Est Républicain, de type non narratif. Chaque bloc correspond aux 10 000 premiers mots du texte complet et est détaillé dans le tableau suivant :

Titre	Auteur	Source	Date	Type textuel	Genre textuel
Pauline	G. Sand	Wikisource	1881	Narratif	Roman
Le Diable au corps	R. Radiguet	Wikisource	1923	Narratif	Roman
Douce Lumière	M. Audoux	Wikisource	1937	Narratif	Roman
Est Républicain	-	Ortolang	2003	Non narratif	Articles de presse

Tableau 2.2 – Textes annotés en coréférence pour le corpus Democrat.

L'annotation a été réalisée à l'aide de l'outil Analec pour tous les textes à l'exception de « Pauline » qui a été annoté avec TXM lorsque la version implémentant Analec a été disponible. J'ai pu tester l'annotation à l'aide de la pré-annotation en chunk avec SEM (Tellier, Y. Dupont et Courmet 2012) pour « Le Diable au corps ». Cependant, comme d'autres, pour un gain de temps et d'efficacité j'ai préféré partir d'un texte brut pour l'annotation (Landragin, Potier et Bothua 2017).

Le travail d'annotation de ces blocs a permis de mettre en lumière des cas problématiques d'expressions référentielles pour lesquelles il est difficile d'identifier clairement le référent ou de savoir s'il y a coréférence avec un autre maillon potentiel d'une chaîne de coréférence. C'est le cas notamment dans l'exemple [44] rencontré dès les premiers

paragraphes du premier texte annoté :

Exemple [44]

« Jusqu'à douze ans, je ne me vois aucune amourette, sauf pour une petite fille, nommée Carmen, à qui je fis tenir, par un gamin plus jeune que moi, [**une lettre**]$_i$ dans [**laquelle**]$_i$ je lui exprimais mon amour. Je m'autorisai de cet amour pour solliciter un rendez-vous. [**Ma lettre**]$_i$ lui avait été remise le matin avant qu'elle se rendît en classe. J'avais distingué la seule fillette qui me ressemblât, parce qu'elle était propre, et allait à l'école accompagnée d'une petite, comme moi de mon petit frère. Afin que ces deux témoins se tussent, j'imaginai de les marier, en quelque sorte. À [**ma lettre**]$_i$, j'en joignis donc une de la part de mon frère, qui ne savait pas écrire, pour Mlle Fauvette. J'expliquai à mon frère mon entremise, et notre chance de tomber juste sur deux sœurs de nos âges et douées de noms de baptêmes aussi exceptionnels. J'eus la tristesse de voir que je ne m'étais pas mépris sur le bon genre de Carmen, lorsque, après avoir déjeuné avec mes parents qui me gâtaient et ne me grondaient jamais, je rentrai en classe.

À peine mes camarades à leurs pupitres - moi en haut de la classe, accroupi pour prendre dans un placard, en ma qualité de premier, les volumes de la lecture à haute voix -, le directeur entra. Les élèves se levèrent. Il tenait [**une lettre**]$_j$ à la main. Mes jambes fléchirent, les volumes tombèrent, et je les ramassai, tandis que le directeur s'entretenait avec le maître. Déjà, les élèves des premiers bancs se tournaient vers moi, écarlate, au fond de la classe, car ils entendaient chuchoter mon nom. Enfin, le directeur m'appela, et pour me punir finement, tout en n'éveillant, croyait-il, aucune mauvaise idée chez les élèves, me félicita d'avoir écrit [**une lettre de douze lignes sans aucune faute**]$_j$. Il me demanda si je [**l'**]$_j$avais bien écrite seul, puis il me pria de le suivre dans son bureau. Nous n'y allâmes point. Il me morigéna dans la cour, sous l'averse. Ce qui troubla fort mes notions de morale, fut qu'il considérait comme aussi grave d'avoir compromis la jeune fille (dont les parents lui avaient communiqué [**ma déclaration**]$_k$), que d'avoir dérobé [**une feuille de papier à lettres**]$_l$. Il me menaça d'envoyer [**cette feuille**]$_l$ chez moi. Je le suppliai de n'en rien faire. Il céda, mais me dit qu'il conservait [**la lettre**]$_j$, et qu'à la première récidive il ne pourrait plus cacher ma mauvaise conduite. »

Raymond RADIGUET, *Le Diable au corps*, 1923.

L'identification du référent de la lettre pose problème. Comment traiter la référence de cet objet qui comprend à la fois le contenu, le support et le tout ainsi que la lettre

que tient le directeur ? Faut-il annoter toutes ces mentions comme faisant partie d'une seule et même chaîne ? Faut-il au contraire en faire des chaînes distinctes ? Même en tentant de séparer théoriquement ces éléments, la tâche n'est pas évidente. Une solution raisonnable est de faire une première chaîne concernant la lettre du narrateur en général, celle mentionnée au début. Puis, une seconde chaîne concernant la lettre que le directeur tient en rentrant dans la classe car le lecteur ne sait pas s'il s'agit de la même lettre au début. Cette seconde chaîne peut pourtant tout aussi être inclue dans la première. Une troisième chaîne peut s'appliquer au contenu[29] de la lettre : « ma déclaration », qui serait un singleton car cette mention n'est pas reprise. Une quatrième chaîne peut référer au support de la lettre : le papier dérobé. Cette mention est reprise par « cette feuille » bien que le référent ait quelque peu évolué entre le moment du vol et celui de la discussion avec le directeur. En tout état de cause, que faire de cette dernière mention : « la lettre » ? Le référent est-il encore le support ou la lettre dans sa globalité ? Ce type de questionnements rencontrés au début de cette thèse en a orienté le sujet afin d'obtenir des pistes de solutions.

Une deuxième tâche d'annotation a eu lieu au mois de juillet 2018, après l'annotation des quatre textes pour le corpus Democrat. Cela concernait la double annotation pour apporter de l'aide à Marine Le Mené dans la tâche de l'évaluation de l'accord inter-annotateurs pour Democrat. L'objectif de cette tâche était d'obtenir 10% du corpus annoté en double. Pour cela, il a été décidé de sélectionner 1 000 mots au début et 1 000 mots à la fin de la moitié des blocs annotés dans le projet afin que la double annotation soit plus représentative de la diversité des textes que dans le cas où ce travail aurait été fait sur des blocs entiers en moins grand nombre. J'ai donc réalisé l'annotation manuelle de deux textes de 2 000 mots chacun environ à l'aide de TXM. Ces textes ont aussi été annotés au LiLPa et sont issus des blocs « Elisabeth Seton » et « Aden Arabie ». Ils sont détaillés dans le tableau suivant :

Titre	Auteur	Source	Date	Type textuel	Genre textuel
Elisabeth Seton	L. Conan	Wikisource	1903	Narratif	Biographie
Aden Arabie	P. Nizan	Ebooks Gratuits.com	1931	Non narratif	Pamphlet

Tableau 2.3 – Textes annotés en coréférence pour la double annotation dans Democrat.

J'ai ainsi participé à diverses tâches qui n'apparaissent pas dans la version publiée du corpus Democrat, mais qui ont contribué au travail sur l'analyse des annotations dans le projet.

29. La distinction entre le contenu et le support est bien connue sous le nom de « dot-object » dans le modèle de Pustejovsky (1998).

2.4 L'exploitation des annotations : un problème ouvert

Une fois l'annotation réalisée, avec les décisions qu'elle impose tant du point de vue théorique que du point de vue technique, il reste à les exploiter. Pour travailler sur les chaînes de coréférence, il faut pouvoir effectuer des décomptes, des mesures, des comparaisons et visualiser ces résultats. Tout cela est nécessaire afin de pouvoir analyser ce phénomène complexe qui peut s'étendre sur toute la longueur d'un texte. En raison du manque d'outils performants pour cela jusqu'à présent, on pouvait observer un manque de méthodologie dans l'analyse des chaînes de coréférence. Les avancées de TXM et du projet Democrat dans ce domaine devraient permettre une évolution positive.

2.4.1 Avancées du projet : TXM

Un des objectifs du projet Democrat était de pouvoir fournir un outil d'annotation qui permette d'annoter, visualiser et analyser les chaînes de coréférence, le tout sur un support utilisé largement et maintenu [30]. Cet objectif a été rempli par le projet via la plateforme TXM (Heiden, Magué et Pincemin 2010a) qui a implémenté des fonctionnalités essentielles d'Analec (Landragin, Poibeau et Victorri 2012). Pour le volet de Democrat concernant la linguistique outillée, le livrable principal correspond donc à l'extension URS [31] de TXM [32] accompagnée d'un manuel d'utilisation. L'un des avantages de TXM est notamment de pouvoir traiter des corpus et sous-corpus ainsi que des partitions de textes.

Trois des blocs de texte présentés dans le tableau [2.2] ont été annotés à l'aide d'Analec : *Le Diable au corps*, *Douce Lumière* et l'*Est Républicain*. TXM a été utilisé pour l'annotation de *Pauline* et des textes de la double annotation : « Elizabeth Seton » et « Aden Arabie ». Le choix de ces outils a été lié aux avancées du projet Democrat, l'extension « Analec », devenue « URS », dans la plateforme TXM n'étant pas disponible au moment de l'annotation des trois premiers textes. Il est donc possible pour l'annotation dans TXM de travailler à partir de fichiers issus de GLOZZ (une combinaison de trois fichiers) ou d'Analec (au format XML-TEI). Il est aussi possible d'importer dans un corpus TXM des annotations si elles sont au format XML-TEI-URS [33].

La fonctionnalité principale de cette extension est donc l'annotation manuelle inter-

30. Sujet développé dans 2.1.1
31. Schéma développé en 2.2.1.
32. https://www.lattice.cnrs.fr/democrat/files/txm-manual-urs-extension-v1.0.pdf
33. Format abordé dans en 2.2.1.

active comme le montre la figure [2.5] à partir de l'édition de texte :

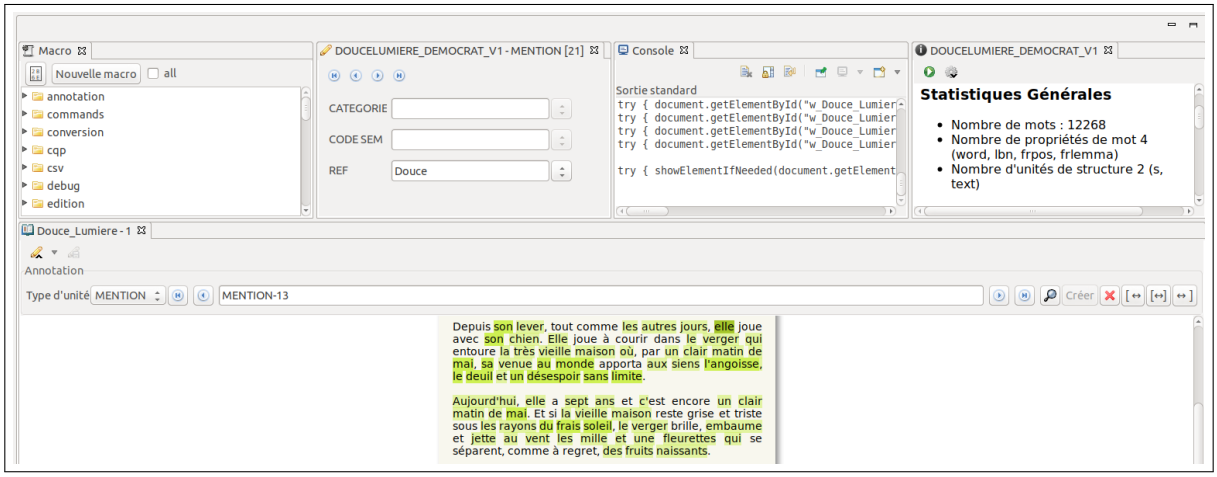


Figure 2.5 – Annotation manuelle d'un texte dans TXM.

Dans la figure [2.5], le pronom « elle » est sélectionné. Le référent qui lui a été assigné est « Douce ». Comme dans Analec, l'annotation se fait dans TXM à partir d'une « structure d'annotation », sous la forme d'un fichier à importer dans l'outil au préalable. Ce fichier est spécifique à chaque corpus car il encode le schéma d'annotation utilisé [34]. L'extension « Annotation URS » correspond à l'implémentation d'Analec et permet donc, une fois la structure d'annotation chargée, de procéder à des annotations de textes puis de les exploiter.

Il est possible de faire des requêtes URSQL [35] pour sélectionner des éléments en fonction de leur type, propriétés ou valeurs de propriétés. TXM permet aussi d'enrichir des annotations à l'aide de commandes ou de macros [36]. Une macro représente une liste de commandes automatisées pour réaliser certaines opérations sur les textes. Dans TXM, ces outils peuvent permettre de créer des unités (de nouvelles mentions), des annotations ou de supprimer des unités de manière automatique comme la macro de la figure suivante :

34. Voir la structure d'annotation de Democrat : 2.2.
35. Un langage de requête, comme GLOZZQL, à l'état de proposition par TXM au moment de la rédaction de cette thèse.
36. La macro « frpos2Categorie » permet par exemple de remplir le champ « catégorie » à partir d'une sortie de TreeTagger en récupérant la partie du discours assignée à chaque expression.

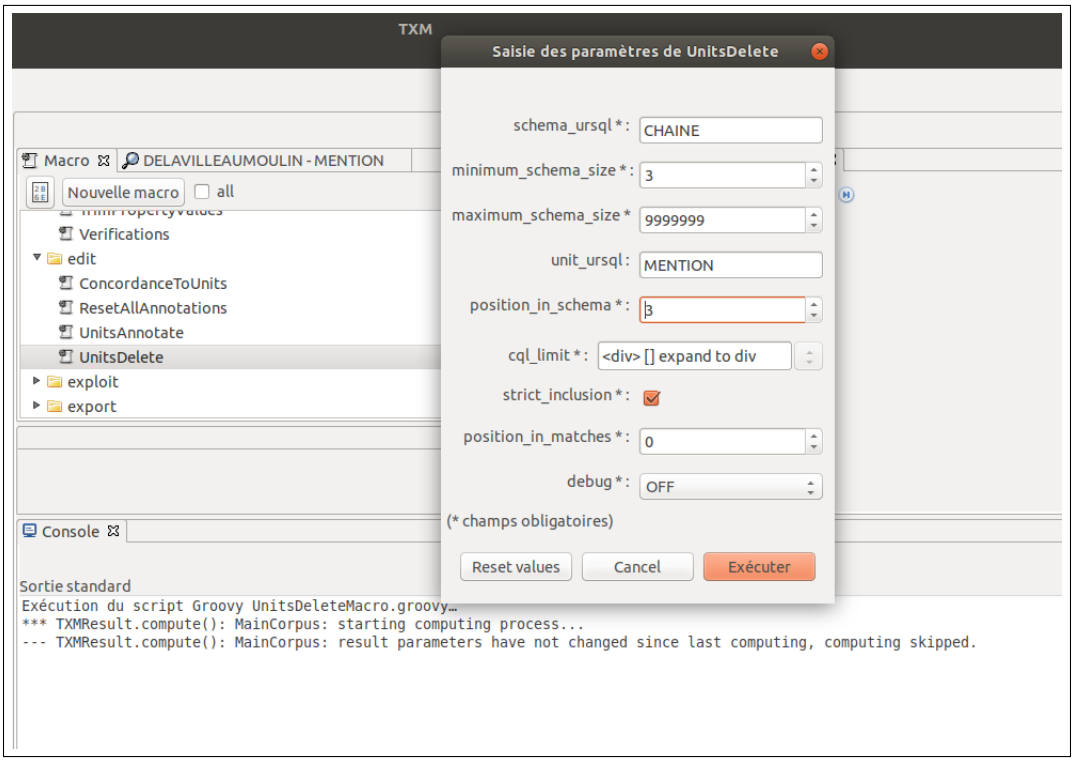


Figure 2.6 – Macro de suppression d'annotations à l'aide d'une requête URSQL.

Dans la figure [2.6], la macro « UnitsDelete » permet de supprimer les unités sélectionnées à l'aide d'une requête URSQL qui vise les unités dont la position dans le schéma est « 3 ».

TXM permet aussi de vérifier la cohérence des annotations effectuées. C'est le cas avec la macro « CheckAnnotationStructureValues » qui permet de vérifier que toute valeur d'une propriété définie dans la structure d'annotation est effectivement utilisée comme annotation. Ces valeurs correspondant aux référents, cette macro permet par exemple de vérifier que chaque référent référencé dans l'annotation a au moins deux mentions, sinon il n'est pas considéré comme faisant partie d'une chaîne. Ce type de macro représente une étape importante pour la bonne réalisation des calculs sur les chaînes de coréférence car la qualité des résultats de ces calculs est sensible aux erreurs d'annotations (manques, redondances, etc.).

En ce qui concerne l'affichage, TXM permet comme Analec de visualiser les unités en spécifiant un type d'unité (comme les mentions), ce qui est visible dans la figure [2.5]. De plus, il est possible d'exploiter ces unités afin de réaliser des extractions pour des affichages ou des décomptes. Il est également possible de réaliser les moyennes, médianes, quartiles et histogrammes des distances et des cadences entre les unités (les mentions) comme dans le tableau suivant :

Mesure	Valeur
Distance moyenne	312354 / 1901 = 164.3103629669
Distance medianne	10
Distance quartils	0 4 10 39 7564
Cadence moyenne	171172 / 2166 = 90.0431351920
Cadence medianne	5
Cadence quartils	0 2 5 13 7392
Densité référentielle	nUnites/nMots = 2166/12268 = 0.1765568960 = 17,66%
Stabilité référentielle	Voir annexe [0.2]

Tableau 2.4 – Statistiques TXM sur les unités dans « Douce Lumière ».

La distance entre les unités est calculée en nombre de mots. La cadence permet de mesurer la distance entre les unités sans les compter elles-mêmes. Les unités sont prises en compte pour le calcul de distance mais pas pour la cadence. D'autres mesures comme la densité référentielle et la stabilité référentielle sont aussi réalisables et sont présentes dans le tableau [2.4]. La densité référentielle est calculée à partir du nombre total d'unités de type « maillon » du texte divisé par le nombre de mots du texte en %. Elle indique la proportion de mentions dans le texte. La stabilité référentielle est calculée en prenant en compte le rapport entre le nombre de mentions du référent et le nombre de dénominations différentes de ce dernier. Plus cette valeur est élevé plus la stabilité référentielle est grande. Plus elle se rapproche de 1, plus le nombre de dénominations pour désigner le référent est élevé.

Concernant les schémas (pour nous, les chaînes), TXM propose également une liste de commandes dont la plupart des résultats sont répertoriés dans les tableaux suivants :

Fréquence des schémas selon leur longueur	
Longueur du schéma	**Fréquence**
1	1094
2	155
3	36
4	30
5	14
10	4
7	4
6	4
15	2
8	2
666	1
233	1
104	1
97	1
76	1
52	1
44	1
36	1
24	1
22	1
21	1
19	1
17	1
9	1

Tableau 2.5 – Statistiques TXM sur les schémas dans « Douce Lumière ».

Le tableau [2.5] répertorie les schémas du texte « Douce Lumière », classés par longueur (en unités). La figure [1.3] a été réalisée à partir de ces valeurs, tout comme les résultats du tableau suivant :

Mesures	**Valeurs m > 2**	**Valeur m > 3**
Nombre de chaînes	265	110
Longueur moyenne	8.1735849057	16.8727272727
Longueur médiane	2	5

Tableau 2.6 – Statistiques TXM sur les schémas dans « Douce Lumière ».

Le tableau [2.6] présente les résultats de moyennes et médianes des longueurs des schémas calculés à l'aide de TXM. Le calcul a été réalisé deux fois, la première fois en prenant en compte les chaînes à partir de deux maillons, la seconde en prenant en compte les chaînes à partir de 3 maillons. Comme le montrent le tableau [2.5] et la figure [1.3], les chaînes à deux maillons sont nettement plus nombreuses que les chaînes comportant plus de maillons. Les moyennes et médianes des longueurs des chaînes varient donc grandement lorsque l'on prend en compte ou non les chaînes à deux maillons. TXM permet aussi d'obtenir des diagrammes de progression (comme dans les figures [1.5] et [1.6]).

Les outils pré-existants n'étant pas véritablement adaptés aux chaînes des coréférence (Poudat et Landragin 2017), toutes ces avancées proposées par TXM représentent une base solide pour le travail d'analyse de ce phénomène discursif.

2.4.2 Constat : manque de méthodologie d'analyse des chaînes

L'analyse des chaînes de coréférence est spécifique à chaque projet. Il n'existe pas réellement de méthode scientifique universelle d'analyse de ce phénomène, d'autant plus que la mise en place d'une méthodologie d'annotation de corpus avec un accord inter-annotateurs élevé est difficile à réaliser (Fort, Nazarenko et Rosset 2012). De même que les analyses textométriques habituelles (Salem et Fleury 2009 ; Turenne 2016) ne sont pas optimisées pour le traitement des chaînes de coréférence. Nous présentons dans cette section différentes méthodes d'analyses, mais sont-elles suffisantes ?

En linguistique, Schnedecker, Glikman et Landragin (2017) isolent trois familles d'approches pour traiter de l'expression de la référence : les approches en sémantique grammaticale référentielle, les approches « à caractère discursivo-fonctionnel » et les approches « configurationnelles » / « relationnelles ». Ces approches traitent de référence et de coréférence mais elles « ne prennent pas en considération les CR[37] ». Elles ont néanmoins permis de faire émerger des paramètres essentiels aux chaînes de coréférence tels que la saillance ou la distance (Ariel 1990).

Landragin (2017) retient une routine de quatre étapes habituelles pour l'annotation des chaînes de coréférence, et par conséquent pour leur analyse, les deux allant de pair. La première correspond au repérage des expressions référentielles. La seconde est la résolution de la référence de ces expressions[38]. La troisième étape revient à construire les chaînes (conceptuellement et techniquement) et la quatrième étape est la description de leurs caractéristiques dans des annotations. Cette dernière étape permet donc de pouvoir catégoriser les chaînes et de les comparer.

37. Chaînes de Référence.
38. Avec les problèmes que cela soulève et qui seront abordés dans la partie II.

Pour catégoriser et comparer les chaînes, différentes études ont étudié des critères variés. Boudreau (2004) classe les chaînes de coréférences selon trois types. Elle introduit notamment la notion de « chaîne coréférentielle importante d'un texte » inspirée de Bergler (1997). Ce type de chaînes s'étendent sur toute la longueur du texte, elles sont donc plus longues que les autres. Leur tête se retrouve habituellement dans le titre ou les premières phrases du texte et ces chaînes contiennent les informations importantes d'un texte. Elle repère ensuite des « chaînes coréférentielles associées aux chaînes importantes », qui s'étendent généralement sur un paragraphe, puis les chaînes beaucoup plus courtes.

Le projet PEPS [39] MC4 « Modélisation Contrastive et Computationnelle des Chaînes de Coréférences » (Schnedecker et Landragin 2014) a représenté une phase d'annotation importante. En effet, chaque maillon était caractérisé par une structure de traits comportant 11 propriétés qui pouvaient avoir 78 valeurs possibles. Ce corpus a permis des analyses qualitatives en raison de la quantité élevée des traits annotés. Cependant, les 285 chaînes annotées ne suffisent pas à réaliser des analyses quantitatives représentatives. L'utilisation de l'outil Analec dans une étude sur *L'occupation des sols* de Jean Echenoz (Landragin, Tanguy et Charolles 2015), issu initialement du projet MC4, a permis d'utiliser des représentations graphiques des chaînes pour visualiser les zones de textes intéressantes en terme de densité référentielle. Cette étude a aussi utilisé des calculs de bi-grammes et de tri-grammes [40] ainsi que des décomptes de formes de référence en fonction de la structure textuelle. L'objectif de ces analyses était notamment l'étude des transitions référentielles d'un point de vue qualitatif afin de dégager des premières tendances basées sur des hypothèses linguistiques. Ces deux études ont aboutit à de nouvelles fonctionnalités dans Analec au niveau des statistiques et de la visualisation. Certaines de ces fonctionnalités sont restées à l'état expérimental dans cet outil avec l'arrivée du projet Democrat et l'implémentation d'Analec dans TXM.

Obry et al. (2017) se sont intéressées à la diachronie pour étudier l'évolution de la constitution des chaînes avec le temps. Les mesures utilisées pour cette étude ont été appliquées à l'aide de TXM pour le projet Democrat. C'est le cas de la densité référentielle par exemple, qui est obtenue en divisant le nombre d'expressions référentielles par le nombre de mots dans le texte. Elles notent que cette mesure est à mettre en relation avec la proportion d'expressions référentielles en dehors des chaînes de coréférence. Elles ont aussi calculé le nombre de chaînes de coréférence, qui indique le « caractère mono- ou pluri-référentiel du texte ». Pour schématiser, un texte mono-référentiel contiendra une seule longue chaîne alors qu'un texte pluri-référentiel en contiendra plutôt plusieurs mais plus courtes. Cette étude diachronique a aussi pris en compte la longueur des chaînes de

39. Projet Exploratoire Premier Soutien du CNRS
40. Ici, un n-gramme est une suite de n tokens (mot-forme).

coréférence. Cette mesure signifie peu de choses à elle seule, elle est à mettre en relation avec les moyennes et les tailles minimales et maximales de ces chaînes. En s'inspirant des travaux de (Perret 2000, p. 17) sur le « coefficient de stabilité », elles calculent aussi la « diversité des redénominations » dans les chaînes. Elles obtiennent cette mesure en divisant le nombre d'expressions référentielles nominales par le nombre de désignations nominales différentes pour un référent donné. Plus le coefficient de stabilité est élevé, moins les maillons de la chaîne sont variés. Cette mesure étant limitée par le rapport à la longueur de la chaîne[41], elles la mettent en parallèle avec la mesure de diversité (Pincemin et Malrieu 2014) qui dénombre le total de désignations différentes d'un référent, sans rapport à la longueur de la chaîne. Pour finir, cette étude prend aussi en compte la composition des chaînes (catégorie grammaticale des maillons, accessibilité du référent, type de référent, type du premier maillon de la chaîne).

Dans la même lignée, Schnedecker (2017) propose une approche « configurationnelle »[42] des chaînes de coréférence pour une analyse contrastive entre différents genres textuels. Cette approche prend en compte la densité référentielle, le nombre moyen de chaînes par texte, la longueur des chaînes, le coefficient de stabilité et la catégorie des maillons. Elle ajoute la prise en compte des aspects syntaxiques, thématiques, lexicaux et fonctionnels des maillons.

Avec l'impulsion du projet Democrat, ces études représentent une première proposition de méthodologie d'analyse des chaînes de coréférence d'un point de vue linguistique et statistique.

Le travail fourni au cours de cette thèse pour le projet Democrat constitue une première expérience d'annotation de corpus en coréférence. Comme tout travail d'annotation, il a été ancré dans un cadre théorique précis qui a imposé aux membres du projet de faire des choix répertoriés dans le manuel d'annotation. Les annotations réalisées pour le projet peuvent maintenant être analysées de manière statistique grâce aux avancées de TXM, ce qui permet de pouvoir avancer vers une méthodologie d'analyse des chaînes de coréférence.

41. Plus une chaîne est longue, plus le coefficient de stabilité aura tendance à augmenter car le nombre de désignations différentes n'est pas infini.

42. À la croisée entre les approches « paradigmatiques » et « syntagmatiques ».

Conclusion de la partie 1

Cette première partie constitue une introduction aux phénomènes des chaînes de coréférence et aux problèmes liés à leur annotation. Les chaînes de coréférences constituent un objet d'étude complexe car elles en impliquent d'autres, comme la référence, la coréférence et l'anaphore. Cet objet d'étude complexe est aussi utilisé dans différents domaines qui le définissent chacun différemment selon leurs besoins, c'est pourquoi un point terminologique et conceptuel a été nécessaire. Nous utiliserons donc les termes de « chaîne de coréférence » pour désigner les expressions référentielles qui désignent un même référent, sans limite de nombre de maillon. De plus, cette thèse se consacre essentiellement au problème de la coréférence et ne prendra pas en compte les phénomènes d'anaphores non coréférentes. Pour analyser les chaînes de coréférence, il est essentiel de pouvoir les annoter mais cette annotation soulève de nombreuses questions. Le projet Democrat a tenté d'y répondre avec notamment la réalisation d'un corpus d'envergure annoté en chaînes de coréférence et différentes études sur ce phénomène. Pour des raisons de faisabilité et de temps requis par la tâche d'annotation manuelle, le corpus de ce projet est annoté en coréférence stricte. Cependant, comme le soulignera la deuxième partie de cette thèse, la coréférence n'est pas toujours stricte. En effet, il est parfois difficile d'attribuer un référent à une chaîne de manière certaine car deux chaînes peuvent par exemple se recouper de manière floue, ce qui pose des problèmes aussi bien au niveau technique que conceptuel.

Deuxième partie

La coréférence non stricte

La notion de référence implique l'identification d'un référent (Charolles 2002). Cependant, il arrive parfois qu'un référent soit difficile à identifier pour différentes raisons. Cette difficulté est principalement liée à un choix à faire entre plusieurs référents potentiels. La saillance référentielle peut être un critère permettant de décider quel référent choisir.

Le choix à opérer peut concerner plusieurs référents bien identifiés mais qui ne sont pas reliés sémantiquement. Dans ce cas, le contexte est aussi un allié qui peut fournir des indices. Mais le choix peut aussi concerner des référents très proches sémantiquement, sans être coréférents pour autant car leur relation est facilement identifiable (métonymie par exemple). Qu'ils soient liés sémantiquement ou non, ces cas de figures impliquent des référents bien identifiés.

Le choix à opérer entre plusieurs référents peut aussi relever d'un flou référentiel et/ou coréférentiel. Il arrive qu'un référent évolue au fil du discours, au point que la dernière mention ne corresponde plus réellement au référent de départ. Un référent peut aussi être désigné de manière tellement imprécise qu'il pourrait raisonnablement faire partie d'une chaîne comme d'une autre, parfois des deux. Nous allons voir dans cette partie que la coréférence floue se retrouve souvent concrètement dans des cas typiques. L'étude de ces cas typiques nous amène à nous demander si les coréférences sont toujours vouées à être identifiées précisément et si elles le sont en effet.

Cette seconde partie s'attachera à distinguer et à caractériser ces différents cas de figure en s'appuyant sur des exemples en corpus pouvoir réfléchir à leur place dans les chaînes de coréférence.

Chapitre 3

Le choix du référent

Le choix du référent pour une expression dépend de critères qui le rendent saillant pour le lecteur. Il existe des cas dans lesquels l'identification précise du référent est difficile, bien que le lecteur ait un choix strict à opérer entre des référents déjà identifiés. C'est le cas des ambiguïtés référentielles, de la coréférence proche et des référents évolutifs dans une moindre mesure. Dans cette partie nous aborderons les critères linguistiques impliqués dans ces phénomènes.

3.1 La saillance : un critère de choix

La saillance est le critère qui influence la détermination de ce qui retient l'attention. « La saillance ne fait pas partie du message communiqué, mais tout le message se base sur elle, s'explique par elle, se structure en fonction d'elle » (Landragin 2005, p. 271). Cette section met en évidence l'importance de cette notion dans le choix du référent pour une expression référentielle en fonction de différents facteurs.

3.1.1 Caractérisation

Définitions

La notion de saillance (*salience* ou *saliency* en anglais) est difficile à définir de manière universelle (Hamm 2011), elle implique généralement une idée de mise en avant. En linguistique, elle a eu des acceptions et des usages distincts au fil des années. Comme le synthétise Schnedecker (2011), le terme de saillance est utilisé pour définir des phénomènes différents avec une même étiquette comme la « saillance cognitive ». Ce terme pointe aussi parfois le même phénomène avec des étiquettes différentes comme avec la « saillance visuelle » (Landragin 2004b) et la « saillance ontologique » (H.-J. Schmid 2007) qui désignent le fait que certaines entités extralinguistiques sont davantage en mesure d'attirer l'attention.

Pour souligner que la notion de saillance joue un rôle à plusieurs niveaux de la communication, Combettes (1996) définissait cette notion de la manière suivante : « la saillance est à considérer non seulement comme présence du référent dans la mémoire du récepteur, mais aussi comme “disponibilité” d'une unité, du point de vue de l'émetteur, à servir de thème constant dans les portions de texte qui vont être produites ». Nous retiendrons qu'un élément saillant est celui qui vient en premier à l'esprit. À la lecture d'un texte, cette notion est impliquée dans la résolution des anaphores : lorsqu'il s'agit d'identifier un référent pour une expression, c'est le plus saillant qui sera choisi. La saillance s'applique dans ce cas aux entités du discours.

Dans un texte, un élément est saillant lorsqu'il est mis en avant. Cette mise en avant peut avoir lieu grâce à différents mécanismes phonétiques (prosodie, intonation, etc.), syntaxiques (ordre des constituants, etc.), lexicaux (néologismes, etc.), discursifs (implicite, jeux discursifs, etc.) ou encore extralinguistiques (typographie, etc.). Plusieurs travaux portent sur les critères de saillance dans un texte, qui rendent un élément marquant. Il n'existe pas de saillance absolue pour un élément, mais plutôt des saillances relatives des entités du discours qui évoluent au fil du texte en fonction des critères et des plans d'analyse. Pour distinguer ces différents plans d'analyse, Landragin (2004b) distingue la « saillance physique (ou P-saillance) » de la « saillance cognitive (ou C-saillance) ». Pattabhiraman et Cercone (1990) différencient aussi la « saillance perceptive » de la « saillance conceptuelle ». Dans les deux cas, ces termes différencient la saillance liée aux éléments physiques du message (sémantique des mots, couleur, etc.) et les critères liés aux mécanismes cognitifs du sujet (représentations mentales, émotions, etc.).

Pour définir la saillance, il est souvent question de l'émergence d'une *Figure* sur un *Fond* (Pottier 1992 ; Talmy 2000). C'est selon ce principe que Landragin (2011a) effectue une analogie entre la « saillance visuelle » et la « saillance linguistique » en mettant en parallèle la composition des images et la composition des textes. Ces deux notions peuvent toutes deux être étudiées selon des facteurs physiques et cognitifs.

Différentes théories pour un même concept

La saillance a d'abord été abordée dans le domaine de la perception visuelle (P. Guillaume 1979) et du traitement automatique des langues. Elle est employée dans de nombreux domaines pour des tâches aussi variées que la résolution d'anaphores [1] (Winograd 1972 ; Sidner 1979), l'identification de structures informationnelles (Lambrecht 1996), la production de langage, la génération automatique de textes ou encore le résumé et la traduction automatiques. Pour utiliser cette information de saillance, il est nécessaire de pouvoir la quantifier : lui donner un score par exemple. En sémantique

1. Notamment dans les cas d'ambiguïté (sujet abordé en section 3.2.2).

computationnelle, deux approches différentes ont vu le jour (Landragin 2004a). Les « méthodes statistiques », pour lesquelles une importance est donnée à chaque facteur au préalable (Pattabhiraman 1992 ; Mitkov 1999), s'opposent aux « méthodes dynamiques » (Lappin et Leass 1994) pour lesquelles la valeur de saillance évolue en cours de traitement.

Dans la théorie du centrage (*centering theory*) (Grosz, Weinstein et Joshi 1995) comme dans de nombreux travaux à cette époque, les critères linguistiques de la saillance étaient principalement « formels » (Stevenson 2002). La proximité textuelle est souvent le critère privilégié pour déterminer quel est l'élément le plus saillant. Ce critère de récence, implique donc que plus un élément est mentionné récemment, plus il est saillant. Cette théorie accorde aussi de l'importance aux rôles thématiques. Un *agent* sera considéré comme étant plus saillant qu'un *patient* par exemple. Par la suite, d'autres travaux comme Wolters et D. Beaver (2001) ont ajouté des facteurs plus généraux en prenant en compte la combinaison de plusieurs critères comme la prosodie ou certaines constructions syntaxiques (clivées, appositions, etc.).

La théorie de la pertinence (Wilson et Sperber 1999) se rapproche également du phénomène de saillance. Elle s'inspire d'une maxime de Grice d'après laquelle l'un des objectifs du langage est la compréhension de l'intention communicative de l'interlocuteur, selon un effort conjoint. En fonction des indices donnés par le locuteur, son interlocuteur pourra faire des inférences pour identifier le message initial avec le moins d'efforts possible. Il s'agit donc d'une théorie pragmatique dans laquelle l'effort cognitif fourni par l'interlocuteur doit être justifié par la bonne identification du message.

La saillance est aussi présente à travers la notion d'accessibilité référentielle dans la théorie de l'accessibilité d'Ariel (1988). Cette théorie « a vu la lumière en réaction à l'approche localisante des expressions référentielles selon laquelle le type d'expression référentielle dépend largement du contexte dans lequel se situe son référent » (Demol 2011, p. 128). Cette approche cognitive repose sur l'idée que certaines représentations mentales de référents sont plus facilement accessibles que d'autres. Par conséquent, lorsque quelqu'un émet un message, il sélectionne des expressions référentielles particulières dans le but qu'elles puissent véhiculer les informations nécessaires à l'interlocuteur pour retrouver les représentations appropriées des référents dans sa mémoire. Ariel avance quatre facteurs linguistiques qui influencent l'accessibilité d'un référent :

- **La distance** Plus la distance entre l'expression référentielle et l'antécédent est petite, plus le référent est accessible. En effet, ce dernier a plus de chance de toujours se trouver dans la mémoire de l'interlocuteur.
- **La compétition** Plus les antécédents candidats sont nombreux, moins le référent correct sera accessible.

- **Le degré de saillance du référent** Plus il est saillant à un moment donné, plus il est accessible. Ariel avance qu'un référent qui correspond au topique [2] du discours est plus accessible.
- **L'unité** Une expression référentielle qui se trouve dans le même segment (frame, paragraphe, phrase) que l'antécédent sera plus accessible.

Ces facteurs sont à prendre conjointement car l'accessibilité est un concept complexe. De plus, ces facteurs peuvent pointer dans des directions différentes pour un même référent. Ariel propose une « échelle d'accessibilité » qui hiérarchise les différents types d'expressions référentielles selon leur degré d'accessibilité. Dans cette échelle, les formes zéro, les pronoms clitiques et les réflexifs font partie des catégories les plus accessibles car il n'y a pas de doute sur l'identité du référent dans ces cas-là en général. L'encodage de l'accessibilité basée sur la seule nature des expressions (référentielles ou non) apporte un caractère rigide à ce concept, ce qui, entre autres, a suscité quelques critiques (Reboul 1997b). Pour Kleiber (1994, p. 11), les approches cognitives de la saillance (accessibilité (Ariel 1990) et pertinence (Wilson et Sperber 1999)) sont intéressantes seulement en évitant de « minimiser le rôle sémantique propre de chaque marqueur ».

Lui aussi dans une démarche cognitive, Alshawi (1987) a développé un « modèle du contexte » (*Context Model*) qui donne pour chaque entité depuis le début du texte une valeur numérique correspondant à sa saillance. Cette valeur est appelée « activation contextuelle » (*context activation*) et est continue car elle évolue au fil du texte pour chaque entité. En effet, plus un référent est mentionné dans le discours plus il est saillant. À l'inverse, la saillance d'un référent décroît s'il n'est pas mentionné à nouveau dans les phrases qui suivent (mécanisme de « dégradation »). Ce modèle est implémenté dans le système RAP (*Resolution of Anaphora Procedure*) de Lappin et Leass (1994) et ne nécessite aucune connaissance sémantique préalable. M. Dupont (2002) s'est inspiré de ces travaux sur le calcul de la saillance pour développer son « modèle des attentes ». En plus de fonctionner pour la résolution des anaphores pronominales comme le modèle d'Alshawi, ce modèle fonctionne pour tout type de marqueur. Il s'inspire aussi de la théorie de l'accessibilité d'Ariel (1990), notamment de son échelle d'accessibilité, pour la classification des expressions référentielles. Pour lui l'accessibilité et la saillance désignent le même concept. De plus, il apporte deux nouveaux critères : « le critère de concordance » et « la plage de saillances admissibles ». La mise à jour de la saillance se fait selon trois classes de faible, moyenne ou forte saillance. Ce modèle des attentes a été implémenté dans un système de calcul de la référence : CalCoRef (M. Dupont 2003).

2. Ce à propos de quoi on dit quelque chose.

Des facteurs généraux pour la saillance linguistique

Landragin (2005) traite la saillance comme un phénomène global et propose deux principes pour distinguer les objets saillants : le « principe de primordialité » (importance) et le « principe de singularité » (originalité). L'un des deux principes sera appliqué en fonction des facteurs de saillance considérés. Pour la saillance linguistique, il relève des facteurs liés au sens et des facteurs liés à la forme, détaillés précédemment dans Landragin (2004b), présentés dans le tableau suivant :

Les facteurs liés à la forme de l'énoncé	
Facteurs	Mise en œuvre
La saillance intrinsèque au mot	Phonèmes, orthographe, nature grammaticale, etc.
La mise en avant explicite lors de l'énonciation	Prosodie, prononciation, etc.
Une construction syntaxique dédiée	Construction clivée, détachement, etc.
L'ordre d'apparition des mots	Début ou fin de phrase, répétition, symétrie, etc.
La saillance grammaticale	Fonction sujet, vocative, etc.
La saillance indirecte (grammaticale)	Transfert de saillance par lien grammatical
Les facteurs liés au sens de l'énoncé	
Facteurs	Mise en œuvre
La sémantique des mots	Référent animé, humain, agent/patient, etc.
La sémantique de l'énoncé	Thème ou rhème
La sémantique de la conversation	Topique (l'entité dont il est question)
La saillance indirecte (sémantique)	Transfert sémantique de saillance

Tableau 3.1 – Les facteurs linguistiques de saillance de Landragin (2004b).

Plus récemment, Hou et Landragin (2019) ont présenté une étude contrastive des chaînes de coréférence en comparant les phénomènes de saillance en français et en chinois. Cette étude multifactorielle reprend et affine l'inventaire de Landragin (2004b) présenté ci-dessus. Les critères de saillances sont classés en fonction de leur nature syntaxique, sémantique, textuelle ou pragmatique. La saillance d'une entité du discours est rarement le fruit d'un unique facteur [3]. La plupart du temps ce sont plusieurs facteurs de différentes natures qui jouent chacun un rôle dans la saillance d'un élément. Déterminer l'importance de chaque facteur reviendrait à attribuer une pondération à chacun. Cependant, comme le souligne Landragin (2005) à propos du poids des facteurs, « leur détermination est empirique et nécessite de coûteuses études de corpus ».

3. Hou et Landragin (2019) ont aussi montré que ces facteurs ne sont pas nécessairement universels pour toutes les langues, bien que certains soient valables dans plusieurs langues.

3.1.2 Saillance et coréférence : exemples en corpus

Dans les exemples suivants, on retrouve les critères de saillance définis dans la section précédente. Il est tout de même nécessaire de rappeler qu'il s'agit de textes écrits et que les phénomènes liés exclusivement à l'oral sont par conséquent absents.

Constructions clivées

Les constructions avec présentatif permettent de rendre un élément saillant comme dans l'exemple suivant :

Exemple [45]
« Absolument certain, répliqua **[Roger]**$_r$, puisque **[c']**$_l$est **[elle]**$_l$ **[qui]**$_l$ **[me]**$_r$ l'a dit »
Adèle BOURGEOIS, *Nemoville*, 1917. DEMOCRAT.

L'exemple [45] comporte une construction clivée. La forme prototypique de ce type de construction est :

C'est X	Qu- +Verbe
S1	S2

(Rouqier 2018)

Cette construction permet une focalisation portée par le changement de l'ordre canonique SVO (Sujet - Verbe - Objet) [4]. Ainsi, dans l'exemple [45], la forme canonique de cette construction serait « elle me l'a dit ». La construction clivée permet ainsi de mettre l'accent sur un élément plutôt qu'un autre. Dans cet exemple, le personnage féminin est mis en avant. Cependant, la même construction pourrait être utilisée pour rendre un autre élément saillant : « c'est à moi qu'elle l'a dit (et pas à toi)» versus « c'est elle qui me l'a dit (et pas une inconnue) ». Les contrastes ajoutés entre parenthèses montrent bien le changement de focalisation concernant le référent.

Dans l'exemple suivant, il y a deux constructions clivées :

Exemple [46]
« **[C']**$_e$est **[Églantine]**$_e$ **[qui]**$_e$ parle la première :
— Vous veniez de l'étang, l'avez-vous trouvé changé ?
— Non !
Et Noël ajoute aussitôt :

4. Chaque constituant occupe, par défaut, une place canonique au sein de la phrase (Milner 1989, p. 403).

> — C'est à cette époque de l'année qu'il est le plus beau ! »
> Marguerite AUDOUX, *Douce Lumière*, 1937. DEMOCRAT.

Dans l'exemple [46], certains éléments sont saillants grâce à la réorganisation de l'ordre canonique produit par la construction clivée. L'emploi du présentatif « c'est » suivi par une relative introduite par le pronom « qui » ou « que focalise l'attention sur le référent au centre de cette construction. Dans la première phrase, la clivée permet de mettre fortement en avant le personnage d'Églantine par rapport à celui de Noël. Lorsque Noël prend la parole par la suite dans l'exemple [46], ce référent devient saillant car le nom est en position de sujet et se trouve au début de la phrase. En revanche, la construction est moins lourde que la clivée et ne produit pas le même effet. Une deuxième construction clivée avec un présentatif se trouve à la dernière ligne. Elle met l'accent sur la locution adverbiale « à cette époque de l'année ». Dans une construction canonique SVO comme « Il est le plus beau à cette époque de l'année. », l'élément saillant ne serait pas l'objet (la locution adverbiale) mais plutôt le sujet.

Selon le manuel d'annotation de Democrat [5] dans les constructions clivées, seulement le syntagme focalisé est annoté comme expression référentielle. Dans les exemples [45] et [46], nous avons annoté les trois éléments (*l* et *e*) pour rendre compte du poids de cette construction au niveau de la référence. Selon le manuel, seuls les noms propres « Manine » et « Églantine » devraient être annotés afin d'éviter de créer artificiellement de longues chaînes de coréférence.

Constructions pseudo-clivées

La construction pseudo-clivée permet aussi de mettre en lumière un élément de la phrase en perturbant son ordre canonique. « Les pseudo-clivées entrent dans la catégorie générale des constructions à copule de type " A c'est B " » (Apotheloz et Roubaud 2015, p. 1). Il s'agit de constructions « spécificatives » à ne pas confondre avec les constructions « attributives ». Il y a une relation de coréférence entre A et B car A décrit le référent et B le désigne. Ces deux segments sont reliés syntaxiquement et sémantiquement par une « copule équative ». L'exemple suivant comporte une construction pseudo-clivée :

5. Page 21 : https://www.lattice.cnrs.fr/democrat/files/ANR-15-CE38-0008-DEMOCRAT_livrable_methodo.pdf

Exemple [47]
« [**Ce qui nous plaît le plus dans ces pages**]$_r$, [**ce**]$_r$ sont [**les remarques naïves et enfantines**]$_r$ [**qui**]$_r$ tombent assez souvent de la plume de notre voyageuse. »
Lucie ACHARD, *Rosalie de Constant, sa famille et ses amis*, 1901. DEMOCRAT.

Dans l'exemple [47], la chaîne de coréférence des « remarques naïves et enfantines » est annotée selon le manuel d'annotation du projet Democrat. Le segment gauche entier est annoté comme expression référentielle coréférente au syntagme nominal à droite de la copule « sont ». Le pronom « ce »[6] est aussi annoté comme expression référentielle coréférente au syntagme désignant les remarques. Ce référent est saillant grâce à la construction de la pseudo-clivée à laquelle s'ajoute la reprise par le pronom relatif « qui ».

Il y a une différence d'annotation entre les constructions clivées et pseudo-clivées dans le projet Democrat. Ces deux phénomènes syntaxiques sont pourtant proches et les deux peuvent créer artificiellement de longues chaînes de coréférence.

Constructions disloquées

Comme la clivée et la pseudo-clivée, la construction disloquée permet de rendre saillant un élément en modifiant l'ordre canonique de la phrase. Selon Bally (1965, p. 61), « la segmentation d'une phrase permet de faire de n'importe quelle partie d'une phrase ordinaire le thème ». Selon ce principe, en analyse du discours, un élément qui est disloqué est communément analysé comme étant le thème de la phrase. Un élément de la phrase peut être disloqué à droite comme dans l'exemple suivant :

Exemple [48]
« Je [**le**]$_g$ trouve bien charmant, [**ce garçon**]$_g$, poursuivit madame Boissonneault avec un soupir attendri. »
Yves BEAUCHEMIN, *Le Matou*, 1981.

Il s'agit d'une dislocation à droite dans laquelle le pronom clitique « le » est coréférent à l'élément disloqué « ce garçon ». Ils occupent tous deux le même rôle syntaxique de complément du verbe « trouver ». C'est pourquoi Berrendonner (2015) qualifie ce type de pronom de « doublet » de l'élément disloqué. Cette « redondance fonctionnelle » au niveau syntaxique est plutôt appelée « double marquage » par Blasco (1997, p. 8). Pour elle, le double marquage n'est pas une « redondance » mais plutôt une projection

6. Ce pronom entraîne parfois des divergences de traitement au niveau de la référence et de la coréférence. Ce point est discuté dans la section 4.2.3.

du clitique et donc un « étalement du paradigme du clitique ». Elle se place dans le cadre de l'Approche Pronominale[7] de Blanche-Benveniste (1987) pour préciser que différents types de constructions clivées existent dans lesquelles d'autres relations référentielles opèrent. Elle envisage différents degrés de coréférence (maximal, partiel ou inexistant) en fonction du type de dislocation. Selon elle, lorsqu'il y a double marquage, il y a effectivement coréférence entre le syntagme disloqué et le pronom clitique. Mais il s'agit du seul type de dislocation qui est définissable par ces deux critères. Dans une construction disloquée, l'élément disloqué et le pronom clitique ne possèdent donc pas toujours la même fonction syntaxique ni exactement le même référent.

L'élément disloqué peut aussi se trouver au début de la phrase comme dans l'exemple suivant (exemple [49]). On parle alors de construction disloquée à gauche :

Exemple [49]
« **[Des idées]**$_i$ j'**[en]**$_i$ avais douze à la minute... »
Fanny SEGUIN, *L'arme à gauche*, 1990.

La dislocation à gauche permet souvent de reprendre un élément introduit récemment dans le discours pour en faire l'objet d'un nouveau propos (Berrendonner 2015, p. 15). Dans l'exemple [49], le syntagme « Des idées » est antéposé puis repris par le pronom « en ». De la sorte, il est plus saillant qu'avec l'ordre canonique de la phrase « J'avais douze idées à la minute... ». Anderson, Cerisara et Gardent (2011) ont développé un système de détection de la coréférence dans les constructions disloquées à gauche et soulignent qu'il n'y a pas d'indices structurels simples[8] qui permettent l'identification des pronoms impliqués dans ce type de construction.

Dans les constructions où le syntagme disloqué est directement repris par un pronom clitique, comme dans l'exemple suivant (exemple [50]), la question de la coréférence se pose.

Exemple [50]
« **[Cette pauvre mère Chantemesse]**$_c$, **[elle]**$_c$ a au moins soixante-douze ans. »
Émile ZOLA, *Le ventre de Paris*, 1873. DEMOCRAT.

Ces constructions du type « Moi, je... » ou « Marie, elle... » ne sont pas prises en compte dans toutes les études sur les dislocations. Cette construction ressemble à une apposition et ces dernières ne sont pas annotées dans le corpus Democrat comme étant

7. « L'Approche Pronominale est un cadre théorique syntaxique qui utilise les pronoms pour donner une typologie des différentes sortes de rections verbales » Blasco (1997).
8. Comme la dépendance au verbe par exemple.

coréférentes. De même, la relation de coréférence entre l'élément disloqué et le pronom ne fait pas non plus l'unanimité. Pour Blasco-Dulbecco (2006, p. 29), « moi je » est un cas particulier plutôt présent à l'oral. Dans certains usages, les deux pronoms sont prosodiquement proches et ne fonctionnent pas réellement comme dans une construction disloquée. Elle parle plutôt d'un « sujet complexe » permettant de marquer une prise de parole.

Dans l'exemple suivant (exemple [51]), la locution prépositionnelle « quant à » participe à la saillance du référent en renforçant une construction qui est déjà focalisante.

Exemple [51]
« Quant à **[Marcelle]**$_m$, **[elle]**$_m$ paraissait de plus en plus sous la domination du docteur Desmarais »
Adèle BOURGEOIS, *Nemoville*, 1917. DEMOCRAT.

Qu'il s'agisse de constructions clivées, pseudo-clivées ou d'apposition, il nous paraît essentiel, dans un corpus annoté en coréférences, d'annoter tous les éléments coréférents. Quitte à créer des chaînes de coréférence « artificiellement » longues. Cela comprend notamment le pronom démonstratif « ce » et les pronoms relatifs « que » et « qui » des constructions clivées ainsi que les pronoms dans les appositions et les constructions disloquées. Cela permettrait de mieux pouvoir appréhender et calculer la saillance d'un référent en marquant ainsi le poids porté par ces constructions qui rendent les référents saillants.

3.2 L'ambiguïté : un choix entre des alternatives

Une expression ambiguë a plusieurs significations possibles. L'ambiguïté est donc une cause potentielle de malentendus. Plusieurs raisons peuvent générer ce phénomène, qui est en général résolu grâce au contexte et au phénomène de saillance linguistique.

3.2.1 Caractérisation

Dans les dictionnaires de langue, l'adjectif « ambigu » possède souvent ironiquement deux sens[9] : une interprétation selon plusieurs sens ou une difficulté à caractériser le sens de manière certaine. En linguistique, Fuchs (1996, p. 6) prend comme point de départ la définition suivante : « il y a ambiguïté lorsqu'à une forme unique correspondent plusieurs

9. Source : Dictionnaire de l'Académie française (https://www.dictionnaire-academie.fr/article/A9A1401)

significations ». Une expression ambiguë véhicule donc des significations différentes. Cependant, ce critère n'est pas suffisant pour définir l'ambiguïté. Fuchs (1996, p. 41) ajoute que les différentes interprétations possibles doivent être « disjointes et mutuellement exclusives ». Elles ne peuvent donc pas s'appliquer simultanément. Il y a donc un choix à effectuer entre différentes alternatives pour identifier le référent d'une expression ambiguë. Ce critère permet de distinguer l'ambiguïté des phénomènes de sur-détermination et de sous-détermination pour lesquels aucun choix n'est nécessaire. La sur-détermination concerne des expressions plurivoques[10] en raison d'une superposition de significations ajoutées les unes aux autres. La sous-détermination concerne les expressions univoques[11] mais manquant de précision : l'interprétation reste ouverte. Pour l'ambiguïté, Fuchs (1996, p. 23) parle d'« univocité dédoublée » pour décrire l'« alternative entre des sens disjoints » produite par une expression ambiguë. De plus, Fuchs (1996, p. 10) ajoute que l'ambiguïté est un phénomène qui est « prédictible en langue ». C'est-à-dire qu'il est possible de décrire l'ambiguïté liée à une expression en fonction des entrées qui lui sont associées dans un dictionnaire par exemple.

Les linguistes parlent d'ambiguïté « théorique » ou encore « virtuelle » lorsqu'elle est associée à une expression, un syntagme ou une phrase hors contexte. Cette ambiguïté est fréquemment liée à la présence d'homonymes ou à une difficulté d'attribution des rôles syntaxiques. Ce type d'ambiguïté pourrait généralement facilement être levé à l'aide du contexte. Une ambiguïté qui reste ambiguë avec le contexte est dite « effective » ou encore « référentielle ». Certaines « blagues linguistiques » (Sfar 2008) jouent d'ailleurs sur les mécanismes déclenchés par ce phénomène linguistique.

L'ambiguïté peut se trouver à différents niveaux : morphologique, lexical, syntaxique, sémantique ou encore pragmatique par exemple. Ce phénomène a généré différentes catégorisations à travers les années. Landheer (1985) avance cinq niveaux d'analyse de l'ambiguïté[12] :

- **L'ambiguïté homonymique** Elle peut être syntaxique (« Il couvre la corbeille de fleurs » (Le Goffic 1981, p. 244)) ou lexicale (« Il montrait orgueilleusement sa pêche » (Malmberg 1976)). Dans le premier cas, c'est le rôle syntaxique des éléments qui indique que la corbeille de fleurs est couverte ou bien que la corbeille est couverte de fleurs. Dans le deuxième cas, c'est l'homonymie lexicale du mot « pêche »[13] qui génère l'ambiguïté. Dans les deux cas, la phrase est une représentation unique de plusieurs interprétations différentes possibles.
- **L'ambiguïté polysémique** Dans ce cas, une seule phrase peut recevoir plusieurs

10. Qui ont plusieurs sens.
11. Qui ont un seul sens.
12. Les exemples sont ceux qui sont cités dans l'article.
13. Ce mot possède deux entrées dans le dictionnaire et peut générer différentes significations : https://www.cnrtl.fr/definition/p%C3%AAche

interprétations en fonction de la lecture « opaque » ou « transparente » (Carnap 1956) d'une phrase comme « Paul cherche un porte-monnaie en cuir ». Avec la lecture transparente, un référent spécifique est identifié : Paul cherche un porte-monnaie particulier, qui est en cuir. Avec une lecture opaque, le référent n'est pas spécifique : Paul cherche un porte-monnaie en cuir, quel qu'il soit.

- **L'ambiguïté thématique** Le thème ne ressort pas de manière explicite. L'ambiguïté thématique est caractéristique des phrases isolées et reste donc généralement virtuelle. On la retrouve en discours lorsqu'il y a des incohérences transphrastiques (volontaires ou non) comme dans la devinette citée par Le Goffic (1981, p. 228) : « – Pourquoi met-on une selle au chevaux ? – Réponse : – Parce qu'ils ne peuvent pas la mettre tout seuls. ». L'ambiguïté repose ici sur les présuppositions par rapport à la question : elles portent sur la cause et non sur l'agent. Or, la réponse concerne l'agent, ce qui génère un effet comique.
- **L'ambiguïté discursive** Elle est liée au discours et concerne des énoncés, non des phrases. Elle est « intentionnelle et connotative » et elle est souvent liée à une figure de style comme la métaphore comme dans l'exemple suivant : « Il cherche un porte-feuille bien garni. ».
- **L'ambiguïté situationnelle** Elle est parfois appelée « vagueness » et relève d'un doute sur la relation entre l'énoncé et la situation. Elle est souvent présente dans des énoncés comme « Il est déjà onze heure passées. » qui peuvent laisser supposer que le locuteur est fatigué, que son collaborateur a du retard ou encore qu'il est l'heure de manger ou de prendre un café, etc.

Par la suite, Fuchs (1996) a proposé un classement des ambiguïtés du français[14] :

- **Les ambiguïtés morphologiques et lexicales** Elles concernent les ambiguïtés liées à « la constitution des unités ». Cela concerne les problèmes d'identification des mots, de leurs frontières, de leur catégorie ou de leur flexion. C'est dans cette catégorie que se pose le cas de l'homonymie avec des exemples comme : « Il m'a fallu des quantités de scotch pour faire cette thèse », où le terme « scotch » peut renvoyer à du whisky ou à du ruban adhésif.
- **Les ambiguïtés syntaxiques** Elles concernent les ambiguïtés liées à « la constitution structures syntagmatiques ». Les difficultés d'identification des phrases et de leur structure causent plus souvent des cas d'ambiguïté à l'oral qu'à l'écrit. En revanche, les difficultés d'identification de la fonction des syntagmes et du rattachement des groupes prépositionnels, adjectivaux ou nominaux peuvent générer de l'ambiguïté aussi à l'écrit. Dans la phrase « Paul regarde le toit de la tour. », le groupe prépositionnel « de la tour » peut être rattaché soit à « le toit »[15] soit à

14. Les exemples sont ceux qui sont cités dans le livre.
15. Le toit de la tour.

« regarde » [16].

- **Les ambiguïtés prédicatives** Elles concernent les ambiguïtés liées à « la constitution des structures sous-jacentes ». Les difficultés d'identification du prédicat ou des arguments peuvent être le résultat d'ambiguïtés prédicatives comme dans l'exemple suivant : « Je l'ai vu avant toi ». Dans cet exemple, deux interprétations sont possibles : « Je l'ai vu avant que tu ne le voies. » ou « Je l'ai vu avant que je ne te voie. ».
- **Les ambiguïtés sémantiques** Elles concernent les ambiguïtés liées au « calcul des relations ». C'est le cas des difficultés de calcul de la hiérarchie des opérateurs comme dans « Tout le monde ici parle trois langues ». Est-ce qu'ils parlent tous les trois mêmes langues ou bien sont-ils simplement tous trilingues ? C'est aussi dans cette catégorie que l'on retrouve le cas des calculs du type de procès et d'actants ou encore de l'équilibre thématique de l'énoncé.
- **Les ambiguïtés pragmatiques** Elles concernent les ambiguïtés liées au « calcul des valeurs énonciatives ». Ces ambiguïtés posent le problème du calcul des valeurs référentielles du procès ou des actants comme dans « Le chien aboie. » qui peut avoir une lecture générique ou non. Cette classe d'ambiguïté regroupe aussi les problèmes de calcul des valeurs interlocutives.

Cette catégorisation des types d'ambiguïtés est différente de celle de Landheer, notamment parce que ce découpage consacre une catégorie à part entière aux ambiguïtés syntaxiques. D'autres découpages ont aussi été proposés comme celui de Galmiche (1983) qui distingue trois types d'ambiguïtés représentant des « pièges de la référence ». Ces trois types sont représentés par des distinctions entre différentes lectures pour une même phrase/expression : transparente/opaque, spécifique/non spécifique et attributive/référentielle. Le découpage s'effectue ici plutôt du point de vue de l'interlocuteur et des options qui s'offrent à lui pour lever l'ambiguïté.

L'ambiguïté est donc un phénomène qui intéresse particulièrement les linguistes, que ce soit au niveau sémantique, pragmatique, psycholinguistique, logique ou encore en traitement automatique. Pour résumer, il y a ambiguïté lorsque plusieurs sens qui s'excluent mutuellement sont associés à une seule forme linguistique. Nous nous intéresserons pour ce travail aux ambiguïtés référentielles, effectives. Celles qui ne sont pas levées grâce au contexte et qui font douter un annotateur dans son choix d'attribution d'un référent à une expression référentielle.

16. Il regarde depuis la tour.

3.2.2 Ambiguïté et coréférence

Le contexte joue un rôle important dans le déchiffrage des ambiguïtés pour la résolution des coréférences. Dans le corpus Democrat, les blocs de texte contiennent en moyenne les 10 000 premiers mots de chaque œuvre. Dans ce cadre, la levée des ambiguïtés potentielles est plus facile que dans de courts extraits oraux ou des phrases isolées. Les hypothétiques problèmes d'incertitude interprétative liés au manque de contexte pourraient donc plutôt se situer à la fin des blocs.

L'ambiguïté génère principalement des incertitudes concernant le choix de l'antécédent d'un élément anaphorique entre plusieurs candidats, comme dans l'exemple suivant :

Exemple [52]
« [**Jean**]$_{j}$ est en colère contre [**Paul**]$_{p}$. [**Il**]$_{j|p}$ ne [**l'**]$_{j|p}$accompagne pas à la fête. »

Dans l'exemple [52], la levée de l'ambiguïté détermine l'attribution (exclusive) des référents pour les pronoms « il » et « l' ». Le choix est à opérer entre les noms propres « Jean » et « Paul ». Il existe différentes stratégies pour la résolution d'une anaphore (Kail et Léveillé 1977), faisant usage des critères de saillance évoqués dans la section précédente. Une première stratégie est de choisir le syntagme nominal candidat possédant les mêmes marques lexicales que le pronom (genre et nombre). Une autre stratégie consiste à choisir le candidat qui a les mêmes relations fonctionnelles. La troisième stratégie est de choisir le candidat le plus proche du pronom sur un plan linéaire. Dans le cas de l'exemple [52], la première option n'est pas pertinente car les marques lexicales sont les mêmes pour « Pierre » et « Paul » (masculin singulier). La seconde stratégie est en contradiction avec la troisième, il s'agit donc d'une anaphore ambiguë. Cet exemple est un exemple construit et l'ambiguïté pourrait (ou non) être levée avec l'aide d'un contexte plus fourni[17]. L'exemple suivant est un exemple attesté :

Exemple [53]
« [**Bouvard**]$_{b}$ [l']$_{l}$engagea à mettre bas sa redingote. [**Lui**]$_{b|l}$, [**il**]$_{b|l}$ se moquait du qu'en-dira-t-on ! »
Gustave FLAUBERT, *Bouvard et Pécuchet*, 1881. DEMOCRAT.

Dans l'exemple [53], l'ambiguïté concerne l'attribution du référent des pronom « Lui » et « il »[18]. À la différence de l'exemple précédent où les deux pronoms ne pouvaient pas

17. Le contexte pourrait permettre par exemple de savoir lequel des deux avait prévu d'aller initialement à la fête.
18. Voir la section précédente (3.1.2) sur la saillance pour le traitement de la coréférence dans les constructions disloquées.

avoir le même référent, ici le référent est le même pour « Lui » et « il ». Il reste à définir s'il s'agit de Bouvard ou de l'autre personnage masculin. La première stratégie de résolution de l'anaphore n'est toujours pas pertinente car les deux antécédents potentiels sont aussi masculins et singuliers. Les deux autres stratégies sont aussi contradictoires car Bouvard est le sujet mais le pronom « l' » est plus proche sur le plan linéaire. L'ambiguïté est donc effective.

En production, une ambiguïté peut être volontaire ou non. Dans les deux cas, elle peut entraîner des difficultés d'interprétation du côté du récepteur. Fuchs (1996, p. 50) distingue trois types de récepteurs. Un lecteur (ou interlocuteur) humain se retrouvera face à une « équivoque » s'il ne parvient pas à trancher pour choisir une interprétation. Un linguiste pourra tenter de prédire et décrire les ambiguïtés et se retrouvera néanmoins en difficulté face aux « ambiguïtés effectives » qui ne sont pas résolubles en contexte. Enfin, un ordinateur pourra buter sur des « ambiguïtés virtuelles » qui ne posent pas de difficultés à un récepteur humain. La résolution automatique des ambiguïtés est donc un défi qui suscite un intérêt qui n'est pas nouveau pour les chercheurs (Lesk 1986 ; Hirst 1988 ; Yarowsky 1992 ; Gale, Church et Yarowsky 1992 ; Ide et Véronis 1998 ; Purandare et Pedersen 2004). En effet, les implications sont nombreuses, allant de la traduction automatique aux agents conversationnels (*chatbots*) en passant par la recherche d'information ou encore le résumé automatique. Des travaux orientés sur la résolution des anaphores ambiguës ont eu lieu et se poursuivent sur la base des schémas de Winogad (Winograd 1972). Ces schémas correspondent à des paires de phrases identiques possédant un pronom anaphorique et dont une seule expression diffère. En fonction de l'expression en question, le choix de l'antécédent de l'anaphorique varie. Ces schémas représentent la complexité de l'interprétation de la langue et font l'objet d'un challenge (Levesque, Davis et Morgenstern 2012) car leur résolution fait appel à des connaissances encyclopédiques. Ce challenge a été proposé pour complémenter voire remplacer le test de Turing[19] visant à tester l'intelligence d'un système informatique. En effet, la résolution de schémas de Winograd demande une réflexion à propos des connaissances du monde, comme le démontrent les exemples suivants :

Exemple [54]
« Joan made sure to thank Susan for all the help she had **given**.
Who had given the help ?
Answer 0 : Joan
Answer 1 : Susan »
(Levesque, Davis et Morgenstern 2012)

19. Un juge humain doit décider de la nature de son interlocuteur : humain ou machine. Un système informatique passe le test s'il parvient à se faire passer pour un humain.

Exemple [55]
« Joan made sure to thank Susan for all the help she had **received**.
Who had received the help ?
Answer 0 : Joan
Answer 1 : Susan »
(Levesque, Davis et Morgenstern 2012)

La phrase de l'exemple [54] peut être traduite par « Joan s'est assurée de remercier Susan pour l'aide qu'elle a **donnée**. ». La phrase de l'exemple [55] peut être traduite par « Joan s'est assurée de remercier Susan pour l'aide qu'elle a **reçue**. ». En fonction de l'utilisation de « donnée » ou « reçue », le référent du pronom « she »/« elle » variera.

Pour désambiguïser un terme, il est nécessaire de posséder des informations d'ordre sémantique comme des dictionnaires, des ontologies ou des réseaux sémantiques mais le sens peut aussi être automatiquement déduit à partir de corpus. Une fois les différentes significations identifiées, il faut pouvoir effectuer un choix. Pour cela, une méthode est nécessaire afin d'identifier les critères de choix. Cette méthode peut reposer sur des connaissances ou bien sur des données à partir desquelles un apprentissage (supervisé ou non) est possible pour identifier des schémas qui se répètent. En effet, la première approche consiste à exploiter des bases de connaissances pour effectuer des recouvrements entre des définitions ou des codes thématiques. La seconde approche consiste à regarder l'environnement d'un mot (son contexte) pour en identifier le sens. Il est aussi possible d'utiliser des corpus bilingues parallèles pour identifier le sens d'un mot.

Une définition parfois donnée pour l'ambiguïté est que son interprétation peut générer plusieurs paraphrases qui ne sont pas les paraphrases les unes des autres. L'ambiguïté peut donc souvent se résoudre grâce au contexte car les référents potentiels ne sont pas reliés sémantiquement. Dans le langage courant, la notion d'ambiguïté est souvent associée à des énoncés flous, pour lesquels il n'y a pas de choix à opérer. L'exemple suivant amène une réflexion sur ce sujet :

Exemple [56]
« Je lui ai dit sur un ton de plaisanterie que **son idée** était intéressante, qu'**elle** montrait les choses sous un angle auquel en effet on n'est pas habitué. »
Jules ROMAINS, *Les Hommes de bonne volonté*, 1939 - exemple discuté dans (Landragin 2007).

Dans l'exemple [56], l'anaphore peut faire penser à de l'ambiguïté pour le choix du référent du pronom « elle ». En effet, on peut avoir l'impression d'un choix à opérer entre le pronom « lui » (s'il est féminin) et le syntagme nominal « son idée ». Cependant, cet

extrait relève plutôt de la sur-détermination car si le pronom « elle » peut coréférer à « lui » ou à « son idée », il peut aussi faire référence aux deux en même temps. En effet, l'idée et la personne qui la produit sont des notions fortement reliées sémantiquement, ce qui ne rentre plus réellement dans le cadre de l'ambiguïté. Il est néanmoins impossible de faire un choix et le référent semble flou. Or, l'ambiguïté référentielle n'est pas floue[20] car elle correspond à un choix à effectuer entre deux référents clairement identifiés.

Si un référent est difficile à identifier car il y a un choix à opérer entre plusieurs référents candidats potentiels distincts (sans lien sémantique), il s'agit d'ambiguïté référentielle. L'annotation de ce phénomène est possible mais reste relativement rare en corpus, notamment dans des textes d'au moins 10 000 mots. Le phénomène de la coréférence proche implique des référents clairement identifiés (non flous) mais sémantiquement proches, ce qui peut également provoquer des questionnements au moment de l'annotation.

3.3 La coréférence proche

Il arrive que des référents soient tellement proches que l'envie est forte pour un annotateur de les intégrer dans une seule chaîne alors qu'ils ne sont pas exactement coréférents. Comment faire pour prendre en compte cette information sémantique au moment de l'annotation sans transgresser les règles d'un schéma d'annotation en coréférence stricte ?

3.3.1 Caractérisation et typologie

Ce que nous appelons « coréférence proche » est lié à un sujet qui a été mis en lumière par la thèse de Recasens (2010) puis Recasens, Hovy et Marti (2010) et Recasens, Hovy et Marti (2011) à travers une proposition de typologie de l'identité proche des relations de coréférence[21]. Cette typologie repose sur la complexité de la référence et surtout de la coréférence. Elle prend en compte le fait que des expressions peuvent avoir des référents qui entretiennent parfois des relations sémantiques fortes sans être coréférentes pour autant. Cependant, la forte proximité sémantique de ces référents est une information qu'il serait dommage de ne pas prendre en compte. Ces travaux ont permis cette prise en compte en caractérisant ces relations de coréférence proche. Ils ont établis pour cela trois degrés de coréférence permettant d'aborder la coréférence comme un continuum et de sortir du système binaire fréquemment utilisé pour la relation de coréférence entre deux expressions référentielles : coréférentes ou non. Un degré intermédiaire

20. La notion de flou sera définie dans la section 4.1.
21. Traduction littérale.

est donc ajouté entre l'identité et la non-identité entre deux référents et caractérise une « identité proche ». Cette identité proche concerne des phénomènes linguistiques comme la métonymie ou la méronymie. La typologie NIDENT proposée par Recasens, Hovy et Marti (2010) n'est pas exhaustive car elle a été construite en partant de relations de coréférence jugées problématiques par des annotateurs qui ont ensuite fait l'objet d'une catégorisation. Cette typologie considère trois degrés principaux d'identité des référents : la non-identité, l'identité et l'identité proche. De manière évidente, pour la non-identité, les référents sont différents et pour l'identité, il n'y a qu'un seul référent. L'identité proche concerne des référents proches sémantiquement ou pragmatiquement. Ce degré d'identité comporte quinze types d'identités proches qui sont regroupés en quatre catégories : la métonymie, la méronymie, la classe et la fonction spatio-temporelle.

La métonymie est une figure de style qui permet d'exprimer un concept au moyen d'un autre concept avec lequel il entretient une relation nécessaire. Ces relations classées par Recasens et al. selon six sous-catégories. Le première relation métonymique correspond au rôle ou à la fonction d'un individu pour le désigner. Les relations suivantes sont la localisation, l'organisation (entreprise ou organisation sociale à laquelle appartient le référent), la réalisation informationnelle (une histoire peut par exemple être réalisée selon différents formats : livre, film, etc.) et la représentation (une peinture, un objet, une réplique ou même une représentation mentale). La dernière sous-catégorie est une catégorie « autre » qui contient les potentielles autres relations métonymiques non répertoriées. La méronymie est une figure de style permettant de désigner un concept par le biais d'une partie de ce concept. Cette catégorie est divisée en trois sous-catégories. La première est la relation partie-tout. La seconde sous-catégorie correspond à la relation entre un objet et la matière dont il est fait. La troisième relation méronymique concerne la relation entre des ensembles dont les frontières ne sont pas bien définies. La relation de classe implique des référents de même classe mais plus ou moins génériques ou spécifiques. La catégorie des relations spatio-temporelles concerne une même entité mais réalisée selon différents critères. Elle est divisée en quatre sous-catégories : le lieu, le moment, la fonction numérique (prix, âge, etc.) et le rôle ou la fonction (par exemple une personne différente désignée par la même fonction).

L'approche de Recasens se situe dans le cadre théorique de la sémantique conceptuelle de Jackendoff (1983) et Jackendoff (2002), des espaces mentaux de Fauconnier (1984) et Fauconnier (1997), de la fusion conceptuelle de Fauconnier et Turner (2008) ainsi que de l'identité relative de Geach (1962). Elle va aussi dans le sens du point de vue de Barker (2010) à propos des degrés de similarité. La coréférence est pensée dans le cadre d'un « modèle de discours » et les entités auxquelles les expressions réfèrent sont donc des « entités de discours »[22]. La théorie des espaces mentaux de Fauconnier

22. Notion abordée en 1.1.1

cherche à retranscrire les représentations mentales qui prennent place et évoluent dans le discours. Selon cette théorie, les expressions ne réfèrent pas à des entités du monde qui nous entoure mais elles reflètent la pensée du locuteur et sa manière de l'exprimer. Chaque énoncé entraîne la création d'un espace mental en fonction de sa structure linguistique. Une expression ne possède donc pas de sens en elle-même mais un sens qui est activé en contexte. Cette approche accorde donc de l'importance aux aspects pragmatiques de la coréférence.

Le cadre établi par Recasens permet une formalisation ainsi que des visées de traitement automatique. En effet, la typologie ainsi fournie par Recasens, Hovy et Marti (2010) a été mise en application à travers un schéma d'annotation (Recasens, De Marneffe et Potts 2013). Ce schéma a été appliqué dans le but d'étudier l'accord inter-annotateurs pour ce phénomène mais aussi dans le but de fournir un corpus permettant d'appliquer des outils de détection automatique de coréférences. Ce schéma d'annotation a été appliqué sur le corpus polonais « Polish Corpus Coreference » (Ogrodniczuk, Głowińska et al. 2013). Après annotation manuelle, les liens de coréférence proche se sont montrés peu fréquents et l'accord inter-annotateurs de ce phénomène était selon eux trop bas pour être fiable.

3.3.2 La coréférence proche en corpus

Les cas concrets qui impliquent des relations de coréférence proche peuvent faire douter un annotateur dont la tâche est d'identifier les expressions référentielles (et donc les coréférences). C'est le cas de l'exemple [31], cité dans le premier chapitre comme étant un phénomène non coréférentiel concernant les deux premiers référents. En effet, il est question dans cet exemple de « la grille du potager », puis des « gros barreaux rouillés » et même ensuite des « lances pointues ». La relation entre ces trois référents ne relève effectivement pas de la coréférence stricte. Elle rentre cependant dans la classification de Recasens et al. pour la relation de méronymie entre la partie et le tout concernant les relations grille/barreaux et grille/lances.

Dans l'exemple suivant, une autre relation de méronymie est présente :

Exemple [57]
« Ils ne sortaient pas sans [**leur louchet**]$_{l}$, et coupaient en deux les vers blancs, d'une telle force que [**le fer de l'outil**]$_{f}$ s'en enfonçait de trois pouces. »
Gustave FLAUBERT, *Bouvard et Pécuchet*, 1881. DEMOCRAT.

Dans l'exemple [57], le louchet est un outil en fer ou bien comprenant du fer. Par la

suite, l'expression « le fer de l'outil » n'est pas strictement coréférente à « leur louchet » car elle désigne la matière de l'outil ou bien l'une des matières dont est composé l'outil. L'annotation de cet exemple est effectuée en coréférence stricte, les deux expressions référentielles n'ont donc pas le même indice. Cependant, il s'agit là d'un cas de méronymie qui peut être une relation objet/matière dans le premier cas ou bien partie/tout si l'outil n'est pas composé exclusivement de fer mais comporte aussi une partie en bois par exemple. Dans ce cas, il n'est pas évident de trancher, même en utilisant la classification NIDENT. De même, les relations de métonymie peuvent entraîner des réflexions à propos de la coréférence comme dans l'exemple suivant :

Exemple [58]

« Dans l'état actuel des choses, ce sont les gouvernements qui ont des initiatives et des décisions à prendre et je déplore leur manque d'action et de courage. Je constate qu'ils n'ont jusqu'ici donné aucune suite aux rapports qui leur ont été soumis par [**Bruxelles**]$_{b|c}$. Toutefois [**la Commission européenne**]$_{c}$ ne renonce pas et elle annonce la publication de nouveaux rapports. »

Pierre MENDÈS-FRANCE, *Œuvres complètes*, 1990.

Dans l'exemple [58], Bruxelles désigne la Commission européenne par métonymie. L'expression référentielle « Bruxelles » ne désigne pas directement la ville belge mais l'institution européenne qui s'y trouve. Ce rapprochement entre le lieu et l'institution est courant mais demande néanmoins des connaissances de culture générale pour pouvoir être correctement interprété. Une autre mention de la ville de Bruxelles dans ce texte ne serait pas coréférente avec l'occurrence de l'exemple [58]. L'annotation de cet exemple souligne la complexité de la coréférence proche. Dans un corpus annoté en coréférence stricte, il est possible que ces deux mentions soient ou non annotées comme étant coréférentes en fonction du choix de l'annotateur. Ce phénomène de métonymie par le lieu est pris en compte dans la typologie de Recasens et al. ce qui permet de donner un cadre plus précis pour ce type de phénomène.

La prise en considération de la relation de coréférence entre plusieurs expressions référentielles comme étant une relation « scalaire » apporte des informations plus fines linguistiquement. Cette catégorisation de l'identité proche ne représente pas pour autant la finesse des représentations cognitives humaines. De plus, l'identification du référent dans ce type de contexte reste liée au jugement humain de l'annotateur et il est plus difficile de trouver des consensus au niveau de l'annotation. L'appropriation d'un tel schéma d'annotation ainsi que la tâche d'annotation en elle-même sont donc davantage coûteuses en temps de travail. La complexité des algorithmes de détection automatique de coréférence augmente aussi dans ce cadre précis.

Si la nature d'un référent subit des mutations, comme un crapaud qui se transforme en prince ou une planche qui devient une étagère, il ne s'agit pas exactement de la coréférence proche de Recasens. Ce phénomène implique aussi une idée de continuum mais avec une idée d'évolution du référent.

3.4 Les référents évolutifs : un cas limite ?

Un référent évolutif concerne une entité dont la nature même a subi une évolution. Cette transformation est telle qu'un linguiste ou un annotateur pourra se demander s'il s'agit bien de la même entité avant et après ce changement.

3.4.1 Référence évolutive et identité

Caractérisation

Le phénomène de la référence évolutive a été discuté sans être nommé par Brown et Yule (1983) à travers des exemples repris maintes fois ensuite dans des travaux sur les référents évolutifs. Brown et Yule ont d'abord présenté la notion de cohésion textuelle de Halliday et Hasan (1976) avec leur taxonomie des types de relations cohésives. Ces relations au sein d'un texte (comme la conjonction par exemple) contribuent à lui donner un sens en maintenant la continuité sémantique de celui-ci. C'est via cette notion que Brown et Yule abordent la question des référents qui font l'objet d'une anaphore alors qu'ils ont subi une transformation. Ils citent d'abord l'exemple de Halliday et Hasan :

> **Exemple [59]**
> « Wash and core six cooking apples. Put them into a fireproof dish. [23] »
> Brown et Yule (1983, p. 191) citant Halliday et Hasan (1976)

Brown et Yule discutent cet exemple car les six pommes anaphorisées dans la seconde phrase sont celles de la première phrase qui ont été lavées et évidées. L'anaphore permet bien de relier ces deux phrases en assurant un lien sémantique entre ces dernières, maintenant ainsi la cohésion textuelle. En revanche, cet exemple permet une remise en question de la conception substitutive de la coréférence de Halliday et Hasan car les pommes ont subi une transformation et le pronom ne réfère plus exactement au syntagme nominal qu'il anaphorise. Brown et Yule proposent donc ensuite un exemple construit plus « violent » et devenu populaire pour rendre compte du changement d'état du référent :

23. Traduction : « Laver et évider six pommes à cuire. Mettez-les dans un plat ignifuge. »

Exemple [60]

« Kill an active, plump chicken. Prepare it for the oven, cut it into four pieces and roast it with thyme for 1 hour.[24] »

Brown et Yule (1983, p. 202)

Dans l'exemple [60], l'anaphore permet d'assurer la cohésion textuelle en permettant la persistance du référent malgré les changements. Cet exemple permet aussi de prendre conscience des limites de la conception substitutive, car remplacer les pronoms anaphoriques de la seconde phrase par « un poulet actif et dodu » entraînerait selon les auteurs une incompréhension du texte. En effet, cela paraîtrait irrationnel car le poulet ayant changé d'état, il ne possède plus la caractéristique d'être « actif » au moment d'être mis au four par exemple. Brown et Yule proposent donc d'utiliser un modèle de traitement permettant d'accumuler ces changements d'état du référent afin que son identité ne soit pas simplement interprétée en fonction de l'antécédent de l'anaphore mais qu'elle prenne en compte son évolution au fil du discours. Yule (1982, p. 318) reconnaît néanmoins qu'un tel modèle pourrait engendrer des représentations plus lourdes et coûteuses. C'est pourquoi Brown et Yule supposent plutôt que la référence est déterminée par une représentation mentale du discours par le lecteur/interlocuteur. Schnedecker et Charolles (1993a, p. 117) trouvent cette approche mentaliste « trop libérale du point de vue linguistique » car elle ne prend pas en compte la capacité (et surtout l'incapacité) référentielle de chaque pronom pour désigner un état ou un autre du référent au cours du discours, notamment lorsqu'il y a un changement de genre ou de nombre.

L'évolution appliquée aux référents est généralement apporté par les verbes. C'est pourquoi Charolles et François (1998) ont étudié les verbes susceptibles d'apporter du changement pour proposer une classification des prédicats transformateurs. Cette classification se focalise sur la transformation des entités concrètes et comporte trois grandes catégories de verbes. Ils peuvent indiquer soit la création, soit l'annihilation soit un changement qualitatif d'une entité. Le changement peut ne pas affecter la forme de l'entité, ou bien l'affecter par un passage du massif au comptable, ou encore modifier sa « structure partonomique[25] ».

Achard-Bayle (2001) et Achard-Bayle (2016) évoque l'idée d'un « objet-thème », regroupant les aspects ontologiques et textuels de ce type de référents particuliers. Charolles (2002, p. 50) parle de l'indétachabilité du mode de présentation des référents : « Pour Frege et les philosophes comme J. Searle qui s'inspirent de ses idées, les référents sont indétachables de leur mode de présentation. ». L'identité des entités désignées dépend donc des expressions référentielles utilisées ?

24. Traduction : « Tuez un poulet actif et dodu. Préparez-le pour le four, coupez-le en quatre morceaux et faites-le rôtir avec du thym pendant 1 heure. »

25. Une structure partonomique est une structure décomposable en parties.

Identité du référent

Tout changement appliqué à un référent affecte-t-il nécessairement son identité ? À quoi tient l'identité du référent ? Schnedecker et Charolles (1993a) abordent la question de l'« identité référentielle » liée à l'anaphore. L'identité référentielle ayant été traitée avant cela surtout en philosophie et en logique. D'un point de vue linguistique, il est effectivement intéressant de se pencher sur le rôle des pronoms dans la continuité référentielle à travers les changements opérés sur le référent par l'intermédiaire de prédicats verbaux. Les auteurs analysent que les syntagmes nominaux référentiels « individuent une entité et ils la fixent dans un monde départ ». Les noms propres pouvant accepter plus de changements que les noms communs en vertu de leur caractère de désignation plus rigide. Les pronoms quant à eux permettent de maintenir la référence, sans pour autant garantir le maintien des traits qui caractérisent l'entité de départ car ils ne possèdent pas de capacité descriptive. Schnedecker et Charolles (1993a) opèrent donc une distinction entre la coréférence et l'identité matérielle. Les pronoms ne permettent donc pas d'assurer que le référent est toujours exactement le même que lors de sa fixation dans le discours. En revanche, ils indiquent que l'entité fixée dans le discours précédemment est toujours fortement accessible[26].

Les pronoms assurent la continuité référentielle mais il existe bien un seuil à partir duquel l'identité du référent peut changer. Quels sont les critères qui maintiennent l'identité du référent ? Dans le discours, à partir de quel moment de l'évolution se crée une identité nouvelle ? Toute transformation implique-t-elle obligatoirement un changement de référent ?

Pour Schnedecker et Charolles (1993a), il faut « concevoir qu'il existe un trait sortal[27] » sous lequel l'entité désignée persiste. Dans Schnedecker et Charolles (1993b, p. 201), ils identifient deux critères de maintien du référent. Le premier est plutôt ontologique et vise la préservation des « propriétés intrinsèques à la désignation » de l'objet ainsi que de son caractère massique, qui doit rester saillant. Le second critère est phénoménologique et dépend de la perception de l'objet et du contexte focal.

Reboul (1997a, p. 8) se penche sur les aspects logiques et philosophiques de l'identité et rappelle le principe de l'identité de Leibniz : « Si A et B sont identiques, tout ce qui est vrai de A est vrai de B. ». Les qualités logiques de l'identité sont la transitivité, la réflexivité et la symétrie. Ces qualités nous rappellent à juste titre les qualités attribuées à la coréférence. Elle poursuit avec la distinction de Ferret (1993, p. 14) entre l'identité numérique, l'identité qualitative (mêmes propriétés) et l'identité sortale (même genre ou espèce).

26. La théorie de l'accessibilité d'Ariel (1990) a été abordée précédemment dans la section 3.1.1.
27. Un trait sortal est un trait d'appartenance à une classe.

Selon Apothéloz et Reichler-Béguelin (1995, p. 240), « le problème des référents évolutifs n'en est pas un : tout objet-de-discours est, par définition évolutif, car chaque prédication le concernant modifie son statut informationnel en mémoire discursive - même s'il s'agit d'une prédication non transformationnelle telle que rester assis ou ne pas bouger ». Mondada et Dubois (1995) vont aussi dans ce sens en considérant « les processus de référenciation en termes de construction d'objets de discours et de négociation de modèles publics du monde ». En revanche, pour Charolles (2001), bien qu'il y ait un niveau représentationnel dans l'interprétation de la référence, le référent existe aussi en dehors du discours. C'est pourquoi selon lui, les référents évolutifs, avec leur fonctionnement particulier, méritent un traitement spécifique. Il distingue aussi les référents évolutifs de l'évolution de la référence, qui ne porte pas atteinte à l'identité du référent.

Tout comme la référence, le phénomène des référents évolutifs est donc un sujet qui fait débat. La définition de l'identité ainsi que le point de rupture ne font pas toujours l'unanimité. D'un point de vue linguistique, la réponse n'est pas évidente, elle est le reflet de la complexité du monde qui nous entoure. Par exemple, si le référent est lié à la manière dont il est désigné, la question de l'identité d'une personne dépend-elle de son prénom ? (Elmiger 2019). Une personne transgenre effectuant une transformation n'estime pas nécessairement changer d'identité en changeant légalement de prénom ou de genre. Cela dépend bien évidemment des cas et des points de vue. Un « point de rupture » n'est finalement peut-être pas toujours nécessaire à trouver ?

3.4.2 Les référents évolutifs et les chaînes de coréférence

Au cours d'un récit, les référents des chaînes les plus longues sont fortement susceptibles de subir des évolutions. Ces référents évolutifs relèvent-ils de la coréférence ? La coréférence est un phénomène linguistique qui désigne la relation entre deux expressions référentielles qui désignent la même unité. Selon cette définition, les référents évolutifs ne sont pas coréférents passé le point de rupture. En revanche, les chaînes de coréférence sont aussi censées rendre compte du devenir discursif d'un référent tout au long d'un texte. Dans quelle mesure est-il intéressant ou non de briser ces chaînes pour qu'elles correspondent aux diverses étapes franchies par les référents ?

Lorsque l'identité se dédouble, il devrait y avoir rupture de la chaîne de coréférence. Or, les pronoms notamment permettent parfois le maintien de la continuité référentielle malgré une altération de l'identité du référent. En reprenant l'exemple [60], la chaîne pour le poulet comporterait quatre maillons, comme le montre le schéma 3.1 :

Figure 3.1 – Représentation du l'évolution du poulet dans l'exemple [60] au fil de la chaîne de coréférence.

Il n'y a aucun doute sur le fait que cette chaîne soit anaphorique car il est nécessaire d'avoir accès aux maillons précédents pour identifier chacun des maillons M2, M3 et M4. Mais s'agit-il pour autant de coréférence ?

Entre M1 et M2, le poulet n'est plus vivant, il perd donc certaines qualités, notamment celle d'être « actif ». Cela pourrait constituer un premier point de rupture. Entre M2 et M3, le poulet perd encore des propriétés, comme son identité visuelle : il n'a plus de tête ni de plumes. Cela pourrait encore être un point de rupture. Entre M3 et M4, il perd son identité numérique en passant de un à quatre morceaux. Il s'agit d'un nouveau point de rupture potentiel. Considérer tous ces points de rupture aurait pour conséquence de ne pas considérer cette chaîne comme coréférente mais plutôt comme quatre expressions référentielles isolées, à savoir quatre singletons. Cette option représente une perte du lien sémantique qui unit ces expressions et ne rend pas compte de la continuité linguistique générée par la successions des pronoms.

Les référents évolutifs peuvent concerner des entités qui se métamorphosent (Achard-Bayle 1998) car la métamorphose représente un point de rupture facilement identifiable. C'est le cas dans l'exemple suivant :

Exemple [61]

« [**l'ennemi le plus dangereux de la peau vivante**]$_{l}$ est [**le varron**]$_{l}$, [**larve d'**[**une mouche appelée hypoderme du boeuf**]$_{m}$]$_{l}$: [**cette larve minuscule**]$_{l}$ pénètre dans le corps où [**elle**]$_{l}$ subit différentes transformations, et, au printemps, [**elle**]$_{l}$ finit par se loger sous la peau de l'échine. Là [**elle**]$_{l}$ grandit jusqu'à atteindre une longueur d'environ 25 millimètres et [**perce**]$_{l}$ un trou dans la peau pour sortir à l'air libre en vue de [**sa**]$_{l}$ métamorphose en [**mouche**]$_{m}$. Une bête peut être atteinte d'[**un grand nombre de [ces varrons]**$_{l}$]$_{v}$, sa peau est gravement endommagée par les trous de sortie qui, même cicatrisés et au bout d'un an, affectent la fleur de marques indélébiles » Jacques BÉRARD, *Cuirs et peaux*, 1947.

L'exemple [61] provient de la base de données Frantext [28]. La requête pour parvenir à le trouver correspond à l'expression CQL suivante :

```
([lemma="larve"%c]) ([]1,10) ([lemma="transformation"%c])
```

Cette requête permet de rechercher les lemmes « larve » et « transformation » entre lesquels 1 à 10 mots peuvent se trouver. Ces mots n'ont pas été choisis au hasard pour cette requête. En effet, lorsqu'il est question d'une larve, il est souvent question de sa transformation et les référents évolutifs sont des référents qui subissent des transformations.

Dans l'exemple [61], la larve et la mouche ont deux noms différents. La première phrase suggère deux référents différents. Il y a donc une chaîne l pour la larve (le varron) et une chaîne m pour la mouche. Cependant, la larve subit des transformations avant sa sortie. Et lorsqu'elle perce sa sortie, il ne s'agit plus d'une larve. Pourtant la suite de pronoms continue de référer à la larve. Il est donc difficile de coller à la fois au discours et à la réalité du référent.

Les changements d'états catégoriques et successifs se retrouvent particulièrement dans les recettes de cuisine qui sont une source d'exemples par excellence pour les référents évolutifs [29] car elles impliquent une transformation des ingrédients comme dans l'exemple suivant :

28. https://www.frantext.fr/

29. Pour Goux et Rossi-Gensane (2019) le doute sur ce type de référent dans les recettes de cuisine est présent quelle que soit l'époque étudiée.

Exemple [62]

« On mange aussi [**les feuilles des plants de manioc**]$_f$; on [**les**]$_f$ fait bouillir pour [**les**]$_f$ amollir, on [**les**]$_f$ réduit en poudre très fine et on [**les**]$_f$ assaisonne avec du poisson, des vers, des grenouilles et du poivre, selon ce qu'on peut se procurer. On mange [**cette préparation**]$_p$ avec du pain de cassave et de la viande. »
John-William PAGE, *Les derniers peuples primitifs*, 1941.

L'exemple [62] provient de la base de données Frantext. La requête pour parvenir à le trouver correspond à l'expression CQL suivante :

```
([lemma="réduire"%c]) ([]1,10) ([word="poudre"%c]) ([]0,10) ([word="fine"%c])
```

Cette requête permet de rechercher le lemme « réduire » suivi d'un à dix mots suivis de la formes « poudre » et « fine » entre lesquelles il peut y avoir entre zéro et dix mots. L'intérêt de la recherche par lemme pour un verbe est de pouvoir rechercher toutes ses formes conjuguées ou non. les mots « poudre » et « fine » sont volontairement recherchés selon leur forme pour éviter de les trouver au pluriel. L'idée de cette requête était de pouvoir trouver différentes variantes de la locution « réduire en poudre ». En effet, l'action désignée par cette locution implique un changement d'état du référent qui subit une transformation. Les « jokers » ([]1,10) et ([]0,10) permettent de chercher des mots, quels qu'ils soient. Cela permet ici de pouvoir trouver diverses combinaisons avec des prépositions (réduire en poudre fine), déterminants (réduire en une poudre fine), des groupes nominaux (réduire les amandes en poudre) ou encore des adverbes (réduire en poudre très fine). Réduire un ingrédient en poudre est commun dans le registre culinaire des recettes de cuisine. Pour rester dans ce domaine, il a semblé nécessaire d'ajouter l'adjectif « fine », pour éviter tout exemple relevant du champ lexical de la guerre par exemple.

Dans l'exemple [62], on retrouve encore une succession d'occurrences du pronom « elle », qui apporte une continuité référentielle pour désigner les feuilles de manioc. Ces feuilles deviennent une « préparation » avec l'ajout d'autres ingrédients. Pourtant, avant cette étape de l'assaisonnement, les feuilles ont subi des transformations et le dernier « elle » évoque de la poudre mais il nous paraîtrait incohérent d'en faire un singleton. L'exemple suivant concerne aussi une recette de cuisine mais sans longue succession de pronoms :

Exemple [63]

« Le dimanche est jour [**des gnocchi**]$_f$. D'abord cuire [**les patates**]$_p$, dans un faitout métallique. [**Les**]$_p$ écraser à la fourchette en [**y**]$_p$ mêlant la farine. Une fois [**la boule de pâte**]$_b$ formée, et malaxée, [**la**]$_b$ découper en longs rouleaux, recouverts de farine. Puis débiter en [**innombrables petites boulettes**]$_i$. [**Les**]$_i$ rouler une à une avec l'index afin d'[**y**]$_i$ creuser une cavité pour la sauce tomate. La sauce mijote, à côté, bœuf, veau, porc. Plat de riches, plat du dimanche. Elle réduit pendant au moins deux heures, le parfum dans la cuisine et partout dans la maison. Rentrée d'une folle course dans les cités, l'odeur entre toutes, et se précipiter à la cave, à la recherche du bloc de parmesan. Une tranche à grignoter telle quelle, sans pain, avant le repas, bien sûr à l'insu de la mère. Juste avant que tout le monde se mette à table. Il y a [**des gnocchi**]$_g$ partout dans la cuisine aux vitres embuées. [**Les**]$_g$ faire passer de l'immense planche de bois sur tout ce que la maison compte d'assiettes avant de [**les**]$_g$ plonger dans l'eau bouillante. Une assiette à la fois, sans [**les**]$_g$ coller. »

Aurélie FILIPPETTI, *Les derniers jours de la classe ouvrière*, 2003.

L'exemple [63] provient de la base de données Frantext. La requête pour parvenir à le trouver correspond à l'expression CQL suivante :

```
([word="pâte"%c]) ([]1,10) ([lemma="découper"%c])
```

Cette requête correspond à la forme « pâte » suivie d'un à dix mots suivis du lemme « découper ». La recherche se fait selon la forme du mot « pâte » pour ne pas la trouver au pluriel. Le verbe « découper » proche du mot « pâte » est potentiellement lié au registre de la cuisine ou du bricolage (sans certitude pour autant).

Dans l'exemple [63], les gnocchi sont résumés en plusieurs étapes générant plusieurs coréférences strictes. D'abord les patates, puis la pâte, les boulettes et enfin les gnocchi. Peut-on considérer les boulettes comme des gnocchi ? Une réponse affirmative à cette question nécessite des connaissances extralinguistiques sur la cuisine italienne. L'annotation de cet exemple a été réalisé en coréférence stricte en se basant sur les expressions référentielles et non sur la réalité du référent. Dans la première phrase, les gnocchi ont une valeur générique. Les deux syntagmes « des gnocchi » ne sont donc pas exactement coréférents. Il y a donc quatre chaînes de coréférence (et un singleton) pour désigner cette transformation des patates en gnocchi. L'exemple suivant concerne un référent humain :

Exemple [64]

« Je connais un peu [**Debbie Harry**]$_d$. Une nuit, à [**ses**]$_d$ débuts, je [**l'**]$_d$avais emmenée voir la tour Eiffel. Et [**la blonde enfant**]$_d$ s'était étonnée : "On ne peut pas monter la nuit ?" Mais il faisait un froid glacial, c'était à Paris et deux ans avant. Depuis, [**elle**]$_d$ était devenue [**Madame Blondie**]$_b$, épousant son guitariste. Et [**elle**]$_d$ avait vendu un million d'albums. Je pouvais toujours tenter de [**la**]$_b$ joindre. »
Philippe MANŒUVRE, *L'Enfant du rock*, 1985.

L'exemple [64] provient de la base de données Frantext. La requête pour parvenir à le trouver correspond à l'expression CQL suivante :

```
([word="elle"%c]) ([word="était"%c]) ([word="devenue"%c]) ([pos!="ADJ"])
```

Cette requête permet de rechercher l'expression « elle était devenue » suivie par un mot qui n'est pas un adjectif. En effet, « elle était devenue belle », par exemple, n'implique pas réellement de changement profond chez le référent.

Dans cet exemple [64], Madame Blondie n'est plus la « blonde enfant », l'auteur en parle comme étant deux personnes différentes desquelles on ne peut plus attendre les mêmes chose. Cela suggère encore une fois de faire deux chaînes de coréférence distinctes pour désigner deux référents dont l'ADN est le même.

Il est encore une fois difficile de trancher, c'est pourquoi un relâchement des contraintes d'annotation est intéressant, comme le suggère Landragin (2018). Ce relâchement de contraintes permettrait d'inclure des référents évolutifs dans une chaîne de coréférence. Cela permettrait de faire le lien entre la larve et la mouche ou entre les différentes étapes des gnocchi par exemple. Comme nous l'avons précisé précédemment [30], dans le corpus Democrat, les référents évolutifs sont annotés dans une seule chaîne de coréférence, permettant ainsi de rendre compte de la continuité référentielle. Un schéma d'annotation permettant des sous-chaînes de coréférence permettrait d'indiquer la nature de la relation de coréférence. Ce n'est pas le cas dans le corpus Democrat et il n'est donc pas possible de retrouver les référents évolutifs ni les différentes phases de leur évolution à l'aide de requêtes dédiées [31].

Les référents évolutifs génèrent un doute à propos de l'identité du référent et cela peut rappeler le phénomène de l'ambiguïté référentielle. Cependant les choix possibles sont liés sémantiquement et correspondent plutôt à différents états sur un continuum

30. Dans la section 2.2.2.

31. Ici, la recherche des exemples servant à illustrer ce sujet a été réalisée à l'aide de requêtes basées sur des verbes de transformation ou des noms impliquant une modification de la structure partonomique du référent.

référentiel pour un référent qui subit des changements. À travers le prisme de la coréférence stricte, un référent évolutif peut générer plusieurs chaînes de coréférence. En effet, il est courant de penser que la première et la dernière mention d'un tel référent désignent des référents différents. Selon le texte, en l'absence d'un point de rupture clair et précis (comme l'éclosion du papillon qui était autrefois une chenille), les mentions intermédiaires se situent souvent dans une zone de flou coréférentiel. Cette zone de floue est maintenue dans le discours grâce aux pronoms car ils permettent de continuer de désigner une entité à travers son évolution sans prendre de décision sur un « degré de mêmeté » bien qu'ils soient aussi à même de prendre en compte les transformations subies à travers les prédicats (Charolles 1997 ; Charolles 2001).

Les référents évolutifs sont liés à une évolution temporelle, où la transformation d'un référent est représentable chronologiquement. Ce qui n'est pas le cas de la *near identity* de Recasens par exemple où le référent est différent mais proche. L'ambiguïté, la coréférence proche et les référents évolutifs sont trois phénomènes linguistiques qui attestent que le langage est chargé de subtilités. La prise en compte de chacun de ces phénomènes dans des travaux ayant des visées de traitement automatique du langage ajoutent une complexité nécessaire pour retranscrire la complexité et la richesse du langage. L'ambiguïté référentielle et la coréférence proche impliquent des référents identifiés de manière précise. Pour l'ambiguïté, il s'agit d'opérer un choix entre des référents potentiels. La levée d'une ambiguïté devrait pouvoir se faire en fonction des critères de saillance : le référent le plus saillant, celui qui ressort le plus, devrait être le candidat choisit comme référent. Or, une ambiguïté effective ne peut pas être levée et les critères de saillance peuvent mener ce choix vers des directions opposées. Pour la coréférence proche, il s'agit d'établir le degré de coréférence que peuvent entretenir deux expressions référentielles ayant des référents bien identifiés qui entretiennent des relations sémantiques. Les référents évolutifs sont à la limite de l'ambiguïté référentielle, de la coréférence proche et du flou coréférentiel car il est souvent difficile d'opérer un choix et de trancher pour décider à quel moment le référent change. Cette difficulté soulève une question d'ordre technique et conceptuel : est-il toujours nécessaire de trancher ?

Chapitre 4

La (co)référence floue : un non-déterminisme référentiel ?

Le chapitre précédent a présenté des cas dans lesquels les référents des expressions référentielles sont difficiles à identifier, sans qu'il ne s'agisse de flou. Pour l'annotation de la coréférence, des solutions existent pour les cas d'ambiguïté et de *near identity*. Ce chapitre traite du phénomène de flou (co)référentiel en partant des définitions proposées dans l'état de l'art en linguistique et dans d'autres domaines pour ensuite définir précisément le sujet au cœur de cette thèse : la coréférence floue. Ce phénomène se retrouve particulièrement dans des cas précis comme la référence à des groupes ou l'utilisation de certains pronoms.

4.1 Référence et coréférence floue

Un objet flou est un objet « dont le contour n'apparaît pas nettement» [1]. Le terme « flou » évoque couramment quelque chose de visuel mais il peut aussi qualifier un concept et s'appliquer au langage et à la (co)référence.

4.1.1 État de l'art

Sans évoquer la coréférence, de nombreux auteurs ont proposé des définitions pour les phénomènes de flou, de vague, d'ambiguïté et de généralité. Ces définitions se chevauchent parfois, apportant ainsi à ces termes la confusion véhiculée par les phénomènes qu'ils décrivent. Ce travail préexistant nous sert de point de départ et nous aide à réfléchir à propos de la (co)référence floue.

1. Source : TLFi (Pierrel, Dendien et Bernard 2004) http://atilf.atilf.fr/

Opacité référentielle

Les notions de transparence référentielle et d'opacité référentielle trouvent leur origine dans les mathématiques pour la description des propriétés des fonctions (Whitehead et Russell 1910). Ces termes ont été repris ensuite en philosophie analytique (Quine 1977). Puis en informatique (Strachey 2000) : dans un langage de programmation, une expression transparente est une expression que l'on peut remplacer par sa valeur sans que cette substitution n'ait d'effet sur le comportement du programme. De même, une expression qui ne possède pas cette propriété est opaque.

Les travaux de Quine en philosophie analytique servent de base aux travaux de linguistique sur l'opacité référentielle. Dans ce cadre, lorsque deux expressions qui désignent le même référent ne peuvent se substituer l'une à l'autre sans que cette opération ne change la valeur de vérité de la proposition, Quine (1977, p. 207) parle d'« opacité référentielle ». Plus tôt, Frege (1892) décrivait un phénomène similaire qu'il appelait « dénotation indirecte ». Un contexte opaque favorise l'apparition de ce type de phénomène[2]. Ces auteurs ont d'ailleurs relevé des conditions particulières favorisantes pour ce type de contexte comme la citation ou certaines constructions verbales (les verbes « intentionnels »)(Apothéloz 2010).

L'exemple le plus repris pour illustrer l'opacité référentielle est le mythe d'Œdipe (Galmiche 1983 ; Fuchs 1994 ; Apothéloz 2010). Il est possible de dire qu'Œdipe veut épouser Jocaste. En revanche, il n'est pas toujours vrai de dire qu'Œdipe veut épouser sa mère. « Jocaste » est pourtant coréférent à « sa mère ». Cependant, Œdipe ne le sait pas. Il ne peut donc vouloir épouser sa mère. Il s'agit donc ici d'une question de point de vue à propos de la catégorisation du référent. À ce sujet, Apothéloz (2010, p. 137) précise que la question de l'opacité référentielle dépend de « l'instance qui prend en charge la catégorisation » de l'expression référentielle. Lorsque cette instance est l'énonciateur, l'expression est transparente. Lorsque cette instance ne peut pas être l'énonciateur ou qu'il existe une ambiguïté à ce sujet, l'expression est opaque. Apothéloz appelle cette autre instance le « médiateur »[3], Kleiber (1979) parle de « sujet ».

L'opacité peut aussi se situer dans l'enchaînement de plusieurs expressions référentielles qui se succèdent. Ce phénomène d'« opacité évolutive » (Apothéloz 2010, p. 149) peut être lié à une évolution de la perception du médiateur.

2. Une expression peut aussi être opaque dans un contexte transparent (Apothéloz 2010).
3. Desclés et Guentchéva (2000) emploient aussi ce terme.

Le générique

Certaines phrases ou expressions peuvent être interprétées de manière générique. C'est souvent le cas des proverbes comme « les chiens ne font pas des chats » ou des phrases du type : « les castors construisent des barrages »[4]. En français, il n'y a pas de marqueur linguistique « dédié » qui permette d'identifier de manière certaine une expression désignant une classe d'entités (Anscombre 2017). Galmiche (1985) identifie néanmoins des structures phrastiques qui sont à même de favoriser une lecture générique comme « UN N, ça SV »[5] ou « UN N, c'est SV »[6] car elles excluent dans la majorité des cas une lecture spécifique. Ces constructions peuvent aussi avoir une lecture générique avec d'autres déterminants que l'indéfini « un », comme « les »[7] ou « le »[8] par exemple.

La généricité peut s'exprimer de différentes manières. En linguistique, des auteurs comme Kleiber (1978), Carlson (1982) et Dahl (1985) donnent cependant deux critères principaux pour qualifier une phrase générique : sa véracité et le caractère non événementiel de son propos. En effet, ce type de phrase permet d'exprimer une vérité qui se veut générale, universelle. Comme troisième critère de généricité, certains auteurs comme Jespersen (1971) et Carlson (1982) ajoutent le fait de posséder un « syntagme sujet générique ». Ce critère est discuté par Anscombre (2002), qui différencie les phrases génériques « extensives »[9] et « partitives »[10] pour lesquelles il propose aussi une classification. Le troisième critère serait plutôt selon lui de faire « intervenir une classe C d'entités et une propriété P telles que l'appartenance d'une entité x à la classe C est un argument pour que x possède la propriété P ».

Zhang (1998, p. 13) distingue les quatre phénomènes linguistique suivants : le flou (*fuzziness*), le vague / imprécision (*vagueness*), la généralité[11] (*generality*) et l'ambiguïté (*ambiguity*). Les trois premiers relèvent selon lui de la sous-détermination alors que l'ambiguïté[12] relève de la sur-détermination.

Le vague et le flou

Pour Zhang (1998), contrairement au vague, à la généralité et à l'ambiguïté, le flou est inhérent car il ne peut pas être résolu grâce au contexte. Selon lui, le flou implique une ab-

4. Cet exemple est souvent repris dans la littérature sur le générique.
5. « Un castor, ça construit des barrages »
6. « Un castor, c'est poilu » - Le SV inclut le présentatif.
7. « Les castors, ça construit des barrages »
8. « Le castor, c'est poilu »
9. « Les chiens sont des animaux »
10. « Certains chiens attaquent l'homme » (Anscombre 2002)
11. Traduction littérale du terme *generality*, que nous pourrions aussi traduire par « généricité ».
12. Phénomène linguistique abordé au chapitre précédent dans la partie 3.2.1.

sence de frontière référentielle évidente ; une expression floue possède des caractéristiques d'opacité référentielle. Selon cette définition, il s'agit du caractère flou de la dénotation qui peut avoir des conséquences sur la référence. Zhang parle donc de « frontières référentielles floues » (*fuzzy referential boundary*) pour certains concepts. Sainsbury (1991) préfère parler de concepts « sans frontière » (*boundariless*).

La notion de vague a été abordée par différents auteurs, notamment en philosophie (Peirce 1911, p. 748) puis en linguistique (Fine 1975 ; Lakoff 1970 ; Geeraerts 1993). Les acceptions des termes « vague » et « flou » peuvent varier selon les auteurs et se recoupent parfois. La définition du vague de Peirce correspond à la définition du flou de Zhang. Kempson (1977) quant à elle, distingue quatre type de « vague ». Le premier type correspond exactement à ce que Zhang appelle « flou ». Le second type ainsi que le troisième type de vague de Kempson sont pour Zhang des sous-catégories de flou. Le quatrième type correspondant plutôt à une sous-catégorie de généralité selon Zhang. Pour Zhang, le vague correspond aux expressions qui peuvent avoir plus d'une interprétation, comme dans les cas de polysémie par exemple car ces interprétations sont sémantiquement reliées.

Dans son chapitre intitulé « Les caprices de la référence », Quine (1977, p. 185) parle du caractère « vague » des termes dont la définition même implique une « zone de flou ». Selon lui, « le vague est une conséquence naturelle du mécanisme de base de notre apprentissage des mots ». En effet, certains termes possèdent des « bords flous » car leur référent se situe sur un continuum. À partir de quel pourcentage de terre mélangée à de l'eau ce mélange peut-il être dénommé « boue » et non plus « terre humide » ou « eau terreuse » par exemple ? Cette définition du vague se rapproche aussi de la définition du flou de Zhang. Comme le souligne Crystal (2008, p. 204), le terme « flou » employé par les linguistes (Lakoff 1973 ; McCawley 1993) a été inspiré par une théorie mathématique. La « théorie des ensembles flous » (Zadeh 1965) a pour objectif de modéliser la représentation humaine des connaissances en prenant en compte l'incertitude et l'imprécision. Zadeh considère l'appartenance d'un élément à un ensemble comme un continuum et non comme une valeur binaire : 0 ou 1. Il représente cette imprécision à travers un degré d'appartenance [0,1] [13] d'un élément à un ensemble.

Les anaphores à antécédent flou

La notion de flou a été appliquée au phénomène de l'anaphore avec les anaphores à antécédent flou de Landragin (2007). Pour ce type d'anaphores, le lien à l'antécédent n'est pas évident à identifier. Il peut y avoir plusieurs candidats. Il ne s'agit pas d'ambiguïté car ces candidats sont reliés sémantiquement. Le problème de l'identification de

13. Compris entre 0 et 1.

l'antécédent de ces anaphores à antécédent flou ne peut pas se résoudre à l'aide de calculs de saillance (Landragin 2007 ; Kister 1995 ; Miltsakaki 2007). Landragin (2007) propose une classification des antécédents susceptibles d'être flous : les possessifs, les groupes complexes « le N_1 de le N_2 », les coordinations, les juxtapositions et les référents évolutifs.

L'antécédent d'une anaphore qui implique un possessif est flou lorsque le possédé et le possesseur sont sémantiquement reliés. Même si ce lien relève d'une relation méronymique [14] par exemple, c'est l'identification de l'antécédent qui sera floue car les deux [15] candidats potentiels peuvent assumer ce rôle individuellement ou simultanément. C'est le cas dans l'exemple [56] abordé précédemment. Les groupes complexes du type « le N_1 de le N_2 » dont N_1 et N_2 entretiennent aussi un lien sémantique génèrent les mêmes difficultés que les possessifs dans la même situation. Les difficultés d'identification de l'antécédent sont encore les mêmes avec une coordination impliquant des pluriels. Pour la juxtaposition, la difficulté est liée à une impossibilité de savoir s'il s'agit d'une énumération ou d'une précision.

Pour résumer, en plus des référents évolutifs [16], les structures suivantes sont susceptibles de générer des anaphores à antécédent flou lorsqu'il y a un lien sémantique entre N_1 et N_2 :

Possessif « le N_1 [...] $[\text{son}]_{N12}$ [...] **il** »

Groupes complexes « le N_1 de le N_2 [...] **il** »

Coordination « les N_1 et les N_2 [...] **ils** »

Juxtaposition « le N_1, N_2, [...] **il** »

Les anaphores à antécédent flou peuvent concerner certaines anaphores abstraites [17] (Asher 2012) ou résomptives [18]. Landragin (2007) ne prend pas en compte ces cas particuliers dans sa classification. Ces anaphores ne relèvent pas toujours de la coréférence car l'antécédent n'est pas nécessairement une expression référentielle.

4.1.2 Définition du phénomène

Qu'il soit appelé « flou », « vague » ou encore « sous-spécification », ce phénomène est présent au niveau de la référence et de la coréférence. On le retrouve donc dans notre

14. La référence du possesseur comme celle du possédé est clairement identifiable.
15. Deux ou plus.
16. Leur structure est moins facilement identifiable.
17. Une anaphore abstraite renvoie à un référent abstrait. Cela regroupe les événements, les situations, les propositions ou encore les faits. Exemple : « La chute de Marie / ça s'est produit(e) alors que le directeur arrivait. » (Amsili, Denis et Roussarie 2005, p. 18).
18. Notion abordée dans la partie 1.2.3.

objet d'étude : les chaînes de coréférence. Il est alors nécessaire de définir la notion de coréférence floue et ses répercutions sur les chaînes de coréférence.

La coréférence floue

Le flou peut s'appliquer à la relation entre anaphoriques et antécédents (Landragin 2007) mais aussi à la relation de coréférence (Landragin 2014) entre plusieurs expressions référentielles (anaphoriques ou non). Lorsque deux expressions sont des maillons potentiels d'une même chaîne de coréférence et que l'interprétation du référent de l'une et/ou de l'autre se confond possiblement avec un autre référent, il s'agit de coréférence floue. Le lecteur ou l'interlocuteur peut avoir l'impression qu'il peut s'agir de l'un comme de l'autre et bien souvent des deux à la fois sans qu'il ne soit envisageable de trancher. Ce qui n'est pas le cas avec l'ambiguïté où les référents sont mutuellement exclusifs car ils ne sont pas reliés sémantiquement. La figure suivante schématise cette relation de coréférence floue :

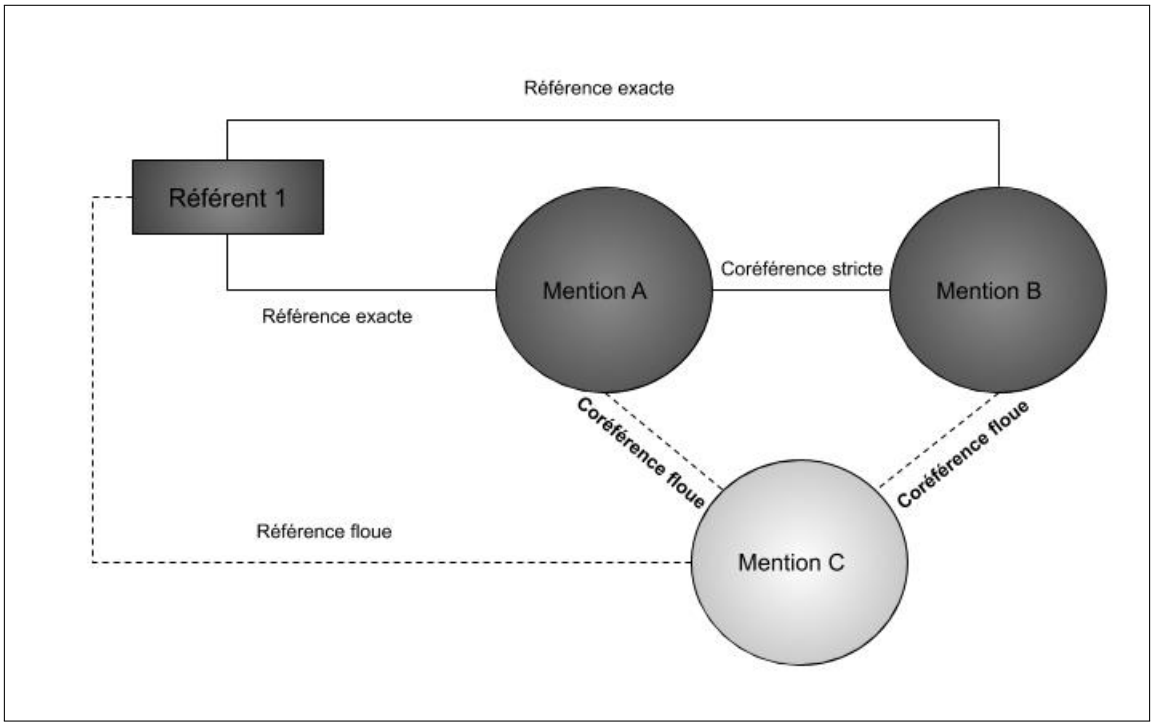


Figure 4.1 – Une schématisation des relations de coréférence stricte et floue.

Comme l'illustre la figure [4.1], dans le cas de la coréférence floue, ce n'est pas seulement le lien de la référence qui est flou. Le flou est surtout présent au niveau de la relation entre les expressions référentielles.

Les chaînes de coréférence et la coréférence floue

En ce qui concerne les chaînes de coréférence, il peut arriver qu'une ou plusieurs expressions référentielles coréfèrent de manière floue à une chaîne stricte (Delaborde et Landragin 2019). C'est le cas dans l'exemple suivant :

> **Exemple [65]**
> « — Je n'y connais personne, madame, répondit la fille ; je ne suis dans ce pays que depuis huit jours.
> — Mais allez me chercher une autre servante, **quelqu'un** ? je veux le savoir. Puisque je suis ici, je veux tout savoir. Est-elle mariée ? est-elle morte ? Allez, allez, informez-vous de cela ; courez donc ?
> La servante objecta que toutes les servantes étaient couchées, que le garçon d'écurie et les postillons ne connaissaient au monde que leurs chevaux. Une prompte libéralité de la jeune dame la décida à aller réveiller le chef, et, après un quart d'heure d'attente, qui parut mortellement long à notre voyageuse, [**on**]$_{o1}$ vint enfin lui apprendre que mademoiselle Pauline D … n'était point mariée, et qu'elle habitait toujours la ville. Aussitôt l'étrangère ordonna qu'[**on**]$_{o2}$ mît sa voiture sous la remise et qu'[**on**]$_{o3}$ lui préparât une chambre. »
> George SAND, *Pauline*, 1881.

Dans l'exemple [65], les trois pronoms « on » sont reliés par des relations de coréférence floue. En effet, il n'est pas évident qu'il s'agisse de la même personne à chaque occurrence bien que cela reste envisageable. Qu'elles soient réalisées par une servante, son chef, un garçon d'écurie ou un postillon, ce qui prime finalement c'est que les actions réalisées le soient par « quelqu'un » d'autre et non l'identification précise du référent. Les trois « on » sont reliés par un lien de coréférence floue. La relation de coréférence floue peut aussi se retrouver entre deux chaînes de coréférence stricte, c'est ce que schématise la figure suivante :

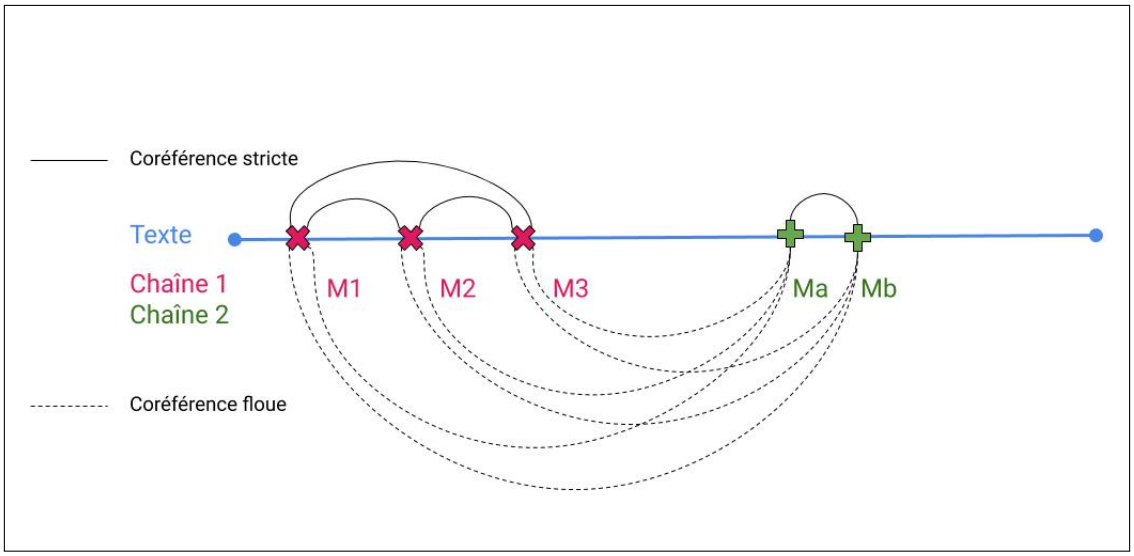


Figure 4.2 – Une schématisation de la relation de coréférence floue entre deux chaînes de coréférence stricte.

Dans la figure [4.2], les maillons Ma et Mb de la chaîne 2 sont coréférents et désignent donc strictement le même référent. Cette chaîne entretient une relation de coréférence floue avec une autre chaîne de coréférence stricte. Cette relation s'applique donc également à la relation de coréférence entre chaque paire de maillons.

Souvent confondu avec l'ambiguïté, le flou relève plutôt de la sous-détermination : lorsque le sens d'une expression linguistique est univoque mais que son interprétation reste ouverte à cause d'un « manque de précision référentielle » (Fuchs 1996). Mais on peut aussi se demander si le flou ne peut pas parfois relever de la sur-détermination : lorsqu'une expression linguistique est associée à plusieurs significations juxtaposées les unes aux autres, comme dans certains jeux de mots pour lesquels le sens du mot est à interpréter avec toute sa polysémie. Qu'il s'agisse de sous-détermination ou de sur-détermination, le flou n'est pas de l'ambiguïté car si plusieurs sens peuvent être associés à une unique forme, aucun choix n'est imposé.

Les définitions du flou et du vague appliquées à la dénotation, à la référence et même aux objets mathématiques, nous permettent de réfléchir à propos de la coréférence floue. Cette relation particulière entre des expressions référentielles a des répercussions sur le traitement des chaînes de coréférence impliquant ce phénomène, notamment lorsque cela implique des chaînes de coréférence stricte.

4.2 Cas typiques

Certaines configurations véhiculent particulièrement bien la notion de flou (co)référentiel. Ces cas de figure contribuent souvent à une identification imprécise du référent. Cela peut concerner le type de référent comme la forme de l'expression référentielle. La référence à des groupes et l'utilisation de certains pronoms sont souvent impliquées dans le phénomène de la coréférence floue.

4.2.1 Groupes pluriels

Le phénomène de la coréférence floue ne concerne pas (ou peu) les noms propres qui ont une forte valeur identifiante. Lorsque ce phénomène concerne un sujet singulier, c'est souvent parce qu'il est lié à un ou plusieurs autre(s) référent(s).

Un groupe dénote un ensemble d'entités qui possèdent quelque chose en commun. Les référents des groupes peuvent être humains ou non. Ils peuvent être désignés par des syntagmes nominaux pluriels comme « les joueurs » ou « les timbres », mais aussi par des syntagmes au singuliers contenant des noms collectifs comme « l'équipe » ou « la collection » (Argenti 2017). Un groupe peut être perçu comme un tout, mais aussi comme une collection d'individus (Mari 2013) désignés de manière plus ou moins explicite (par le biais d'une coordination ou de déterminants numéraux cardinaux par exemple).

Bien qu'il soit constitué d'entités singulières auxquelles il est possible de référer individuellement, un groupe est un référent à part entière. Il est possible d'y référer plusieurs fois et que les expressions référentielles utilisées soient reliées par des liens de coréférence stricte. Il est aussi possible, au cours d'une narration par exemple, que le groupe se divise ou bien que certains de ses membres le quittent alors que d'autres s'y ajoutent.

Les groupes flous, ou possédant des contours flous, désignés par des syntagmes comme « la foule » ou « les gens », peuvent faire l'objet de coréférence stricte bien que le référent ne soit pas toujours précisément identifié. La coréférence floue concerne le flou de la relation entre deux expressions référentielles et pas simplement le flou du lien entre une expression et son référent. Ces groupes flous sont néanmoins sujets à la coréférence floue car ils peuvent facilement générer des mouvements et des sous-groupes dont il n'est pas toujours évident de savoir s'ils sont inclus dans le groupe désigné.

Les groupes pluriels flous et/ou évoqués de manière floue sont souvent désignés ou repris par le pronom « on ». Ce pronom est vecteur de flou dans ce cas mais aussi lorsqu'il désigne des référents individuels.

4.2.2 Le pronom « On »

Référence du « on » : pronom personnel et indéfini

Le pronom « on » vient du latin « homo » signifiant « être humain ». Il est donc étymologiquement destiné à évoquer des référents humains. Il y a néanmoins quelques exceptions, souvent marquées par un certain anthropomorphisme, car il peut aussi référer à des animaux [19] ou des objets [20] par exemple.

Le pronom « on » est difficile à catégoriser (Landragin et Tanguy 2014, p. 99) : il peut être considéré comme un pronom personnel (Charaudeau 1992), un pronom indéfini (Sandfeld 1965 ; Grevisse et Goosse 2008), un pronom personnel indéfini (Pierrel, Dendien et Bernard 2004) ou encore un pronom impersonnel humain (Cabredo Hofherr 2008, p. 35). Selon les grammaires du français (Riegel, Pellat et Rioul 2004), « on » est toujours nominal, en position sujet, et désigne des référents humains animés [21] singuliers ou pluriels, masculins et/ou féminins.

Comme le souligne Gjesdal (2008, p. 15), les analyses classiques du pronom « on » lui accordent deux valeurs discursives principales qui le rendent sémantiquement « complexe » : le personnel et l'indéfini. La valeur personnelle de ce pronom est souvent réduite à une équivalence avec d'autres pronoms comme « nous » ou « je » mais aussi « tu », « il/elle » ou « ils/elles ». Il est parfois qualifié de pronom « omnipersonnel » en référence à sa capacité à être utilisé à la place de n'importe quel autre pronom. D'un point de vue syntaxique, il est intéressant de noter que le « on » peut s'accorder en fonction du référent désigné (syllepse) : singulier ou pluriel, masculin ou féminin. En revanche, comme le souligne Blanche-Benveniste (2003), pour la dislocation, seule l'association de « on » avec le pronom fort « nous » est possible [22]. Il peut aussi, mais rarement, se trouver en apposition avec des groupes nominaux, avec [23] ou sans le nous [24]. Cependant, d'un point de vue sémantique, la vision du « on » comme un simple substitut d'autres pronoms est quelque peu réductrice. En effet, l'emploi du « on » à la place de l'un de ces pronoms ajoute généralement une subtilité supplémentaire. De plus, la singularité de ce pronom est sa capacité à inclure ou exclure l'énonciateur et l'interlocuteur [25] de sa référence (Blanche-Benveniste 2003).

19. Dire « On va manger ? » avant de donner la gamelle à son chien.
20. Dire « Ba alors, on est en grève ? » à un ordinateur qui ne fonctionne plus.
21. Ils répertorient cependant certaines exceptions.
22. « Nous, on » est possible alors que « Je, on » ne l'est pas.
23. « Nous, les mômes, on compte en ronds et en balles. » CAVANNA François - Les Ritals (1978)
24. « Les filles, on voulait faire les gendarmes. » Caddéo (2000) cité par Blanche-Benveniste (2003). On retrouve cette construction plutôt à l'oral ou avec une préposition comme « En Suisse, on manque de lait. » THOMAS Édith - Pages de journal : 1939-1944 (1995).
25. Le nous peut aussi exclure l'interlocuteur.

Il est possible de faire un parallèle entre le « on » / « nous » de modestie et le « nous » de majesté pour lesquels ces pronoms sont mis pour « je » afin de marquer une distance avec le référent. Dans le premier cas, le « on » sert à masquer un aspect auto-centré du langage que peut avoir le « je ». Dans le second cas, le « nous » permet au contraire de donner plus d'importance au référent.

La seconde valeur stylistique du pronom « on », l'indéfini, correspond sémantiquement à une ou des « personne(s) sans référence spécifique », aisément substituable à d'autres pronoms indéfinis comme « quelqu'un », bien que l'équivalence sémantique ne soit pas tout-à-fait la même. Il correspond syntaxiquement à des « constructions impersonnelles et passives » Gjesdal (2008). Muller (1970, p. 52-54) parlait aussi d'une troisième « valeur stylistique » pour laquelle l'utilisation de « on » a « toujours une intention affective » comme l'ironie, l'affection ou la tendresse, entre autres. D'un point de vue discursif, « on » peut avoir différentes fonctions comme la distanciation du référent ou encore le maintien du flou par l'occultation de certaines informations.

Pour Rabatel (2001), la notion de point de vue (Rabatel 1998) permet l'interprétation de « on » comme étant personnel ou indéfini à l'aide de mécanismes sémantiques et pragmatiques. Cependant, il souligne que sa valeur de base est l'indéfini et qu'elle n'est « jamais totalement supprimée ».

Tout comme d'autres pronoms, « on » peut avoir une valeur générique et une valeur spécifique. Cette valeur générique peut être totale comme dans « Pour savoir où l'on va, il faut savoir où l'on est. ». Elle peut aussi être restreinte à certains individus comme dans « En Italie, on parle italien » où le référent de « on » a une valeur générique mais restreinte à un groupe possédant des contours flous : les italiens et/ou les personnes situées en Italie. Cette phrase demande à avoir du contexte parce que son interprétation peut être ambiguë. En effet, elle peut aussi avoir une interprétation spécifique si l'énonciateur parle de son propre usage de la langue italienne, quand il est en vacances avec sa famille par exemple.

Le pronom « on » en corpus

Nous nous sommes intéressée à la place du pronom « on » dans un sous-corpus Democrat. Ce sous-corpus a été défini pour cette thèse, les blocs de texte qui le composent sont présentés en annexe dans le tableau 0.1. Le sous-corpus étudié correspond à 30 textes narratifs et non narratifs, datant du 19ème au 21ème siècle. Cela correspond à la moitié du corpus de Democrat (environ 300 000 tokens) en sélectionnant les textes les plus récents. Dans ce sous-corpus, le pronom « on » est réparti de manière différente en fonction des genres et des types textuels comme le montrent les deux figures suivantes :

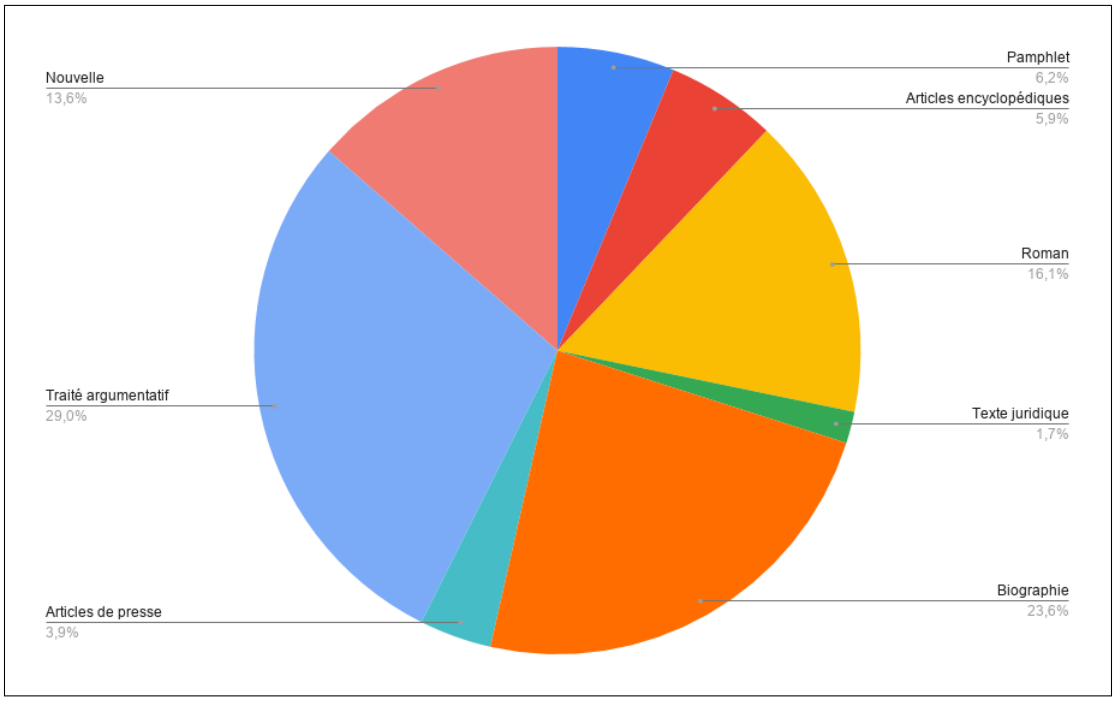


Figure 4.3 – Les occurrences du pronom « on » en fonction du genre textuel dans le sous-corpus de Democrat.

Les textes sont des blocs de texte de 10 000 mots provenant de genres différents et ces derniers ne sont pas représentés de manière homogène. En effet, il y a 10 blocs de romans, 7 blocs de textes juridiques, 5 blocs de nouvelles, 3 blocs de biographies, 2 blocs d'articles de presses, 1 bloc de pamphlet, 1 bloc de traité argumentatif et 1 bloc d'articles encyclopédiques. Nous avons donc appliqué des fréquences relatives pour les occurrences de « on » en fonction du nombre de blocs de texte de chaque genre textuel.

Dans ce corpus, la figure [4.3] montre que le pronom « on » est majoritairement présent dans le traité argumentatif (29%), les biographies (23,6%), les romans (16,1%) et les nouvelles (13,6%). Les biographies possèdent de nombreuses occurrences de « on », ce qui peut sembler étonnant pour un récit qui relate l'histoire d'une tierce personne, contrairement à une autobiographie par exemple. Il serait intéressant de regarder quels sont les emplois du « on » dans ces textes et quels sont les référents associés. Est-ce ces « on » incluent le narrateur ? Est-ce qu'ils servent plutôt de support à une description avec une tournure impersonnelle ? Par ailleurs, le traité argumentatif, qui n'est pas classé comme étant narratif, arrive en première place, loin devant les autres. La prise en compte d'autres blocs de ce genre textuel permettrait de vérifier si la tendance se maintient pour ce genre textuel. Nous pourrions également nous demander si les différents emplois de « on » dans ce bloc de texte sont plutôt des pronoms personnels ou des emplois génériques par exemple. Ces trois propositions sont des pistes qui restent encore à explorer concernant ces deux genres textuels. Les autres genres possédant le plus d'occurrences de « on »

correspondent à tous les genres narratifs présents dans ce corpus. Cette tendance est relayée dans la figure suivante :

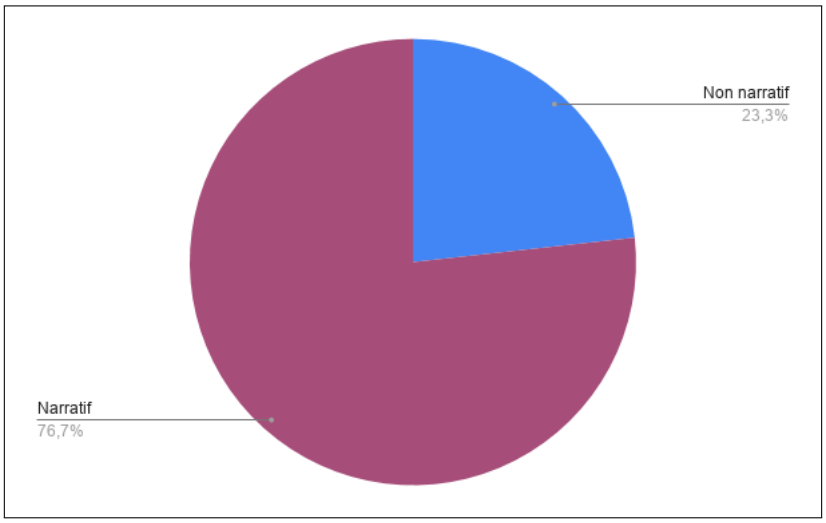


Figure 4.4 – Les occurrences du pronom « on » dans les textes narratifs et non narratifs dans le sous-corpus de Democrat.

La figure [4.4] montre que dans le sous-corpus de Democrat que nous analysons (0.1), le pronom « on » est majoritairement présent dans les textes narratifs.

Il est aussi intéressant de comparer les occurrences du pronom « on » aux autres pronoms personnels avec lesquels il peut potentiellement être substitué. Les nombres d'occurrences de ces pronoms personnels sont répertoriés dans le tableau suivant :

Pronom personnel	Nombre d'occurrences	Proportion
IL	3 929	31,3%
ELLE	2 431	19,4%
JE	2 401	19,1%
ON	**982**	**7,8%**
NOUS	907	7,2%
VOUS	736	5,9%
ILS	710	5,7%
TU	261	2,1%
ELLES	186	1,5%
TOTAL	12 543	100%

Tableau 4.1 – Pronoms personnels dans le sous-corpus de Democrat.

Le pronom « il » est de loin le pronom personnel le plus fréquent, représentant 31,3%

des occurrences. Les pronoms « elle » et « je » le suivent en représentant chacun environ 19% des occurrences. Ensuite, « on » et « nous » se suivent aussi de près en représentant chacun environ 7% des occurrences. Ce classement permet de voir que le pronom « on » est, dans ce corpus, plus fréquent que le pronom « nous ». Il se classe d'ailleurs au 4ème rang sur 9 pronoms. Cela permet d'appuyer l'importance de ce pronom dans les textes.

Nous avons vu précédemment que le pronom « on » peut aussi avoir une valeur indéfinie. Nous avons donc aussi extrait du corpus les pronoms indéfinis qui pourraient avoir une équivalence [26] avec le pronom « on » [27] dans le tableau suivant :

Pronom indéfini	Nombre d'occurrences	Proportion
Tout (tous|toute|tout|toutes)	1 455	63,1%
Quelque (quelque|quelques)	387	16,8%
Chaque	176	7,6%
Plusieurs	106	4,6%
Chacun (chacun|chacune)	88	3,9%
Certain (certaines|certain|certains|certaine)	58	2,5%
Quelqu'un (quelqu'un|quelqu'une)	22	0,9%
Quelques uns (quelques uns|quelques unes)	11	0,5%
Quiconque	3	0,1%
TOTAL	2 306	100%

Tableau 4.2 – Pronoms indéfinis dans le sous-corpus de Democrat.

Les pronoms dont le lemme est « tout » sont les pronoms indéfinis les plus fréquents pouvant potentiellement être équivalent à « on » dans ce corpus. Ils représentent 63,3% des occurrences de ces pronoms. Les autres pronoms sont moins fréquents et possèdent un nombre d'occurrences moins élevé que « on » (987).

Pour les statistiques présentées dans cette sous-section, les pronoms ont été extraits du corpus à l'aide de l'outil TreeTagger (H. Schmid 1994). Sans surprise, cet étiqueteur morpho-syntaxique ne fait pas la distinction entre les emplois personnels et indéfinis du pronom « on », ce dernier est donc classé avec les pronoms personnels. Cela ne permet donc pas une comparaison du pronom « on » en fonction de ces usages avec les autres pronoms du même usage. Cela permet simplement de donner une idée de sa présence dans notre corpus.

26. Seul ou dans un groupe nominal.

27. Pour cela, nous avons simplement retiré de la liste des pronoms indéfinis le pronom « aucun » car il ne peut pas être équivalent à « on ».

Référence du « on » : un vecteur potentiel de flou

Certains pronoms contribuent à l'identification imprécise du référent (lui-même potentiellement imprécis). C'est particulièrement le cas du pronom « on » (Landragin et Tanguy 2014). Ce pronom peut permettre d'identifier des référents de manière précise. Il est aussi et surtout un vecteur de flou notoire. Dans les mots croisés, il est parfois défini comme un « sujet vague », un « personnel un peu flou » ou encore un « pronom assez peu personnel ». Il a souvent été dénigré : « On, pronom imbécile, définit celui qui l'emploie », « On est un con qui ne dit pas son nom ». Ces dictons sont encore quelquefois utilisés pour reprendre de manière prescriptive les personnes qui utilisent ce pronom sans préciser son référent. Or, il peut parfois précisément être utile de maintenir cette information de flou référentiel et/ou coréférentiel. Notamment lorsque le référent est générique ou inconnu. Ce flou est utile dans le cas dans des textes narratifs pour représenter le vague d'une rêverie par exemple. Bien qu'il soit courant de penser que l'emploi du « on » correspond fréquemment à des emplois de français oral « familier » (Grevisse et Goosse 2008, p. 964), même l'Académie française reconnaît que :

> « Si on voulait rejeter les auteurs qui ont employé ce pronom, la France n'aurait plus d'écrivains, l'Académie française n'aurait plus d'académiciens et les anthologies littéraires ne seraient que des coquilles vides. Ostraciser ce pronom, c'est aussi oublier son origine et prendre le risque d'éliminer l'humanité »[28].

Boutet (1986) soulignait déjà aussi « l'importance de l'ambiguïté, de l'indécidabilité et de l'indétermination dans l'interprétation de "on" ».

Le référent du pronom « on » est donc souvent particulièrement difficile à identifier de manière précise. On ne sait pas toujours s'il est inclusif ou exclusif : l'énonciateur est-il compris dans le référent ? Et son interlocuteur ? Il est parfois aussi compliqué de savoir si un « on » est générique ou spécifique : le référent représente-t-il l'ensemble d'une catégorie ou un humain bien identifié ? Ce pronom peut même être dans une position intermédiaire comme dans l'exemple qui suit.

Exemple [66]
« **On** a besoin de travailler. »
Des interlocuteurs dans une queue devant Pôle Emploi - Exemple construit.

Le « on » de l'exemple [66] fait référence aux personnes qui font la queue mais il possède une valeur un peu plus générique que cela. En effet, le besoin de travailler pour vivre est énoncé ici comme une sorte de vérité générale. La valeur générique du « on »

28. Source : http://www.academie-francaise.fr/pauvre

étant véhiculée par l'emploi indéfini, elle véhicule particulièrement bien le flou référentiel. Cependant, ce peut aussi provenir de ses emplois personnels.

L'usage du pronom « on » varie quelque peu en fonction du genre textuel. À l'écrit, il est possible de retrouver des emplois typiques de l'oral dans le discours direct présent dans les textes littéraires par exemple. Ces emplois ne se retrouveront pas dans des textes encyclopédiques par exemple. Ødegaard (2006, p. 62) remarque notamment que :

> « Dans les textes littéraires, l'emploi du *on* correspond à une volonté de la part du locuteur ou du sujet de conscience de "maquiller" l'identité du ou des référent(s) afin de préserver sa face positive ou négative, voir la face négative de l'allocutaire, alors que l'emploi du pronom dans les textes non littéraires sert notamment à exprimer la modestie. »

Elle remarque aussi que le « on » est plus souvent employé de manière indéterminée que déterminée. Un usage indéterminé du « on » qui est très peu marqué par la référence est l'emploi impersonnel, à la manière de « il pleut » : « on est mardi ».

Niziołek (2019) a identifié deux constructions dans lesquelles le pronom « on », suivi d'un verbe, est vecteur de flou référentiel dans des textes fantastiques. La première est une construction « résultative » dans laquelle le référent du « on » n'est pas clairement identifié. L'accent est mis sur l'action et non sur l'acteur. La seconde construction est « évidentielle » de la forme « on » + DIRE (« dit »|« disait »). Cette construction est associée à une transmission d'information non vérifiée, sous forme de « ouï-dire ». Cela permet une mise à distance du locuteur, qui « offre à son interlocuteur la possibilité d'évaluer lui-même la fiabilité de cette information » (Kim 2004, p. 42). Selon Niziołek (2019), ces deux constructions impliquant le pronom « on » suivi d'un verbe dans des textes fantastiques font partie de stratégies d'écriture volontaires qui aboutissent à « une référence brouillée et perturbante ». Selon nous, ces constructions sont effectivement susceptibles de générer du flou (co)référentiel dans ce genre littéraire mais aussi dans d'autres.

Les successions de « on » : des chaînes de coréférence ?

Le flou référentiel que peut générer « on » a des conséquences sur les successions d'occurrences de ce pronom et leur interprétation, coréférentielle ou non. Cela contribue à faire potentiellement dépendre l'interprétation du pronom « on » de celle d'un antécédent antérieur.

Les avis divergent sur le caractère anaphorique de « on ». Certains pensent que ce dernier n'existe pas : selon Atlani (1984), « on » n'a aucune valeur anaphorique, c'est-à-dire qu'il va réactualiser sa référence à chaque occurrence. Pour Bouguerra (1999), il y a même une « vacance référentielle originelle » pour « on » : c'est une « forme

vide » que « l'allocutaire compétent » doit remplir à chaque fois. Dans les grammaires du français, il est question de référence anaphorique pour tous les pronoms. Et plusieurs auteurs comme Fløttum, Jonasson et Norén (2007) et Gjesdal (2008) parlent de complexité référentielle pour « on », avec des possibilités de sur-détermination plutôt que de sous-détermination. Dans la mesure où elle met en avant la complexité référentielle de « on », nous nous plaçons dans la lignée théorique de ces derniers auteurs. Nous suivons l'hypothèse que le pronom « on » peut partiellement s'appuyer sur la référence d'un antécédent, et qu'il peut donc avoir une valeur anaphorique et non-autonome. L'exemple suivant présente une anaphore impliquant le pronom « on » :

Exemple [67]
« [**Mon ami et moi**]$_o$ étions arrivés à 10 heures. [**On**]$_o$ est repartis à 11h. »

Dans l'exemple [67], il y a un facteur anaphorique qui intervient dans l'interprétation du pronom « on » car il renvoie à l'antécédent « mon ami et moi ». Selon Milner (1982, p. 33), un pronom ne peut pas « fonctionner comme le premier terme d'une relation d'anaphore » et un terme ne peut pas être l'anaphorique de lui-même. Selon cette définition, deux « on » consécutifs ne constituent donc pas une anaphore mais une coréférence. Le pronom « on » peut néanmoins être anaphorique, comme dans l'exemple [67]. Il peut par conséquent faire partie d'un ensemble d'expressions référentielles reliées par des anaphores coréférentielles.

En linguistique de corpus outillée et en traitement automatique des langues, il y a coréférence lorsque les référents sont strictement identiques et la construction d'une chaîne se fait uniquement lorsque toutes les relations sont strictes [29]. Cependant, avec le pronom « on », il est souvent difficile d'établir un lien de coréférence stricte et on doit faire face à un dilemme : ignorer ces occurrences ou les rendre strictes [30]. Il est possible qu'un lecteur peu attentif considère comme coréférentes deux occurrences successives de « on » sémantiquement proches, comme les deux dernières occurrences de l'exemple suivant :

Exemple [68]
« Elle parle aussi avec une sentimentalité criante. [**Ma sœur et moi**]$_s$ [**on**]$_s$ l'arrête. [**On**]$_s$ l'arrête à temps. Alors elle dit [**on**]$_f$ ne me laisse pas parler ici. Mais ce ne sont pas des paroles qu'[**on**]$_o$ a envie d'entendre, je ne sais pas pourquoi. »
Chantal AKERMAN, *Ma mère rit*, 2013.

Cependant, une lecture plus fine peut amener à introduire de l'incertitude dans la

29. C'est le cas par exemple dans le projet Democrat.
30. Ce dilemme est approfondi avec des exemples d'annotation du pronom « on » dans le projet Democrat dans la section 5.2.

référence – de l'une ou l'autre occurrence, voire toutes – et à casser ainsi la relation de coréférence. En effet, les deux premiers « on » de l'exemple [68] font référence à un groupe bien identifié qui comporte deux autres référents : le narrateur et sa sœur, annoté *s* dans l'exemple. La coréférence est stricte, notamment en raison de la dislocation[31] pour le premier « on ». La coréférence stricte du deuxième « on » est appuyée par le phénomène de répétition de la séquence « on l'arrête ». Le troisième « on » relève du discours rapporté et ne coréfère pas de manière stricte au groupe précédent comprenant le narrateur et sa sœur. En effet, il peut aussi avoir une interprétation générique bien que l'adverbe déictique « ici » renforce la référence au groupe qui inclut le narrateur en ancrant la référence dans la narration. Le dernier « on » ne coréfère pas non plus de manière stricte au premier groupe car il peut aussi avoir une interprétation générique. La coréférence entre ces deux derniers « on » n'est pas stricte non plus.

Lorsque le pronom « on » apparaît plusieurs fois de manière successive dans un texte, fait-il toujours référence à la même entité extralinguistique ? En suivant une approche descriptive sur corpus, fondée sur un ensemble d'occurrences de « on », nous tenterons de décrire le fonctionnement des exemples et d'identifier les paramètres intervenant lors de la résolution de la référence. Nous verrons que des expressions référentielles peuvent désigner le même référent de manière stricte mais aussi de manière plus vague : l'un des référents peut avoir en plus une valeur générique ou en inclure un autre. Cependant, une succession d'occurrences de « on » n'implique pas obligatoirement de relation de coréférence.

Même référent ?

Il arrive que le référent désigné par le pronom « on » ne soit pas explicité précédemment par un groupe nominal par exemple. Dans ce cas, une succession de « on » coréférents sera identifiée comme une chaîne de coréférence et non comme une chaîne anaphorique. C'est le cas dans les deux exemples suivants :

> **Exemple [69]**
> « Elle fut si aimable et si jolie dans ce badinage, que le bon maire en tomba amoureux comme un fou, voulut lui baiser la main, et ne se retira que lorsque madame D… et Pauline lui eurent promis de le faire dîner chez elles ce même jour avec la belle actrice de la capitale. Le dîner fut fort gai. Laurence essaya de se débarrasser des impressions tristes qu'elle avait reçues, et voulut récompenser l'aveugle du sacrifice qu'elle lui faisait de ses préjugés en lui donnant quelques heures d'enjouement. Elle raconta mille historiettes plaisantes sur ses voyages en province, et même, au dessert,

31. La dislocation n'est pas toujours considérée comme étant un phénomène de coréférence par tous les auteurs, voir section 3.1.2.

elle consentit à réciter à M. le maire des tirades de vers classiques qui le jetèrent dans un délire d'enthousiasme dont madame la mairesse eût été sans doute fort effrayée. Jamais l'aveugle ne s'était autant amusée ; Pauline était singulièrement agitée ; elle s'étonnait de se sentir triste au milieu de sa joie. Laurence, tout en voulant divertir les autres, avait fini par se divertir elle-même. Elle se croyait rajeunie de dix ans en se retrouvant dans ce monde de ses souvenirs, où elle croyait parfois être encore en rêve. [**On**]$_o$ était passé de la salle à manger au salon, et [**on**]$_o$ achevait de prendre le café, lorsqu'un bruit de socques dans l'escalier annonça l'approche d'une visite. »
George SAND, *Pauline*, 1881.

Dans l'exemple [69], les paragraphes précédents permettent d'identifier le maire, la mère D… (l'aveugle), Pauline et Laurence (l'actrice) comme étant les protagonistes du dîner qui sont aussi les référents des pronoms « on » de la dernière phrase. Les deux « on » de cette dernière phrase coréfèrent de manière stricte. Le deuxième « on » fait partie d'une proposition coordonnée à celle qui contient le premier « on » et ils sont tous les deux les agents de verbes qui décrivent l'enchaînement d'actions de fin de repas.

Exemple [70]
« Pourtant, mardi, il y avait bien quelques randonneurs, courageux ou inconscients, à passer devant la ferme-auberge du Ballon, les visages figés par la pluie cinglante et les jeans gorgés d'eau.
"Il manque juste le beau temps. Le soleil, c'est l'or du Ballon" résume Thierry Jeanroy, dont le bar des Démineurs est installé à cheval entre Vosges et Territoire de Belfort.
"[**On**]$_o$ bricole. La saison est vraiment médiocre. La neige, [**on**]$_o$ fait une croix dessus mais il faudrait tout de même que le temps devienne sec. [...]" »
L'Est Républicain, 02/01/2013.

Dans l'exemple [70], le contexte permet aussi d'identifier un référent aux pronoms « on » de manière non explicite. Avec ce pronom, Thierry Jeanroy s'inclut de manière claire dans le groupe auquel les « on » font référence, mais il désigne plus largement un groupe flou qui comprendrait à la fois les employés du bar mais aussi les commerçants les plus volontaires du Ballon d'Alsace sans en connaître précisément la proportion. Si ce groupe peut paraître un peu vague, les deux « on » coréfèrent de manière stricte entre eux car le référent désigné est le même : l'action représentée par le verbe de la première phrase (bricoler) est la conséquence du fait d'avoir fait une croix sur la neige.

Générique, spécifique, un peu entre les deux ?

Le pronom « on » peut permettre au locuteur de prendre de la distance avec un énoncé et de faire des descriptions ou des généralités. Le pronom « on » peut alors avoir un référent de type générique. Dans ce cas, lorsqu'il y a une succession de « on », ils peuvent très bien coréférer de manière stricte comme dans l'exemple suivant :

Exemple [71]
« Je ne sais quelle timidité, si ce n'est celle que l'[**on**]$_{o}$ éprouve en face de ce qu'[**on**]$_{o}$ n'a jamais fait, me retenait. »
Raymond RADIGUET, *Le Diable au corps*, 1923.

Dans l'exemple [71], les deux « on » sont coréférents. Ils possèdent uniquement une valeur générique. Cependant, en plus de cette valeur générique, un « on » peut aussi désigner un référent particulier en plus, comme dans l'exemple suivant :

Exemple [72]
« Lorsque quelque chose, venu de l'extérieur, m'obligeait à penser moins paresseusement à Marthe, j'y pensais sans amour, avec la mélancolie que l'[**on**]$_{g}$ éprouve pour ce qui aurait pu être. « Bah ! me disais-je, c'eût été trop beau. [**On**]$_{o}$ ne peut à la fois choisir le lit et coucher dedans. »
Raymond RADIGUET, *Le Diable au corps*, 1923.

Le premier « on » de l'exemple [72] est générique. Le second est aussi générique mais il ne coréfère pas avec le premier « on ». En revanche, il inclut certainement le narrateur : la phrase est du discours direct où le personnage se parle à lui-même. Ce flou généré par une potentielle inclusion du narrateur dans le référent générique se retrouve aussi dans l'exemple suivant :

Exemple [73]
« J'en étais victime ; mais je rougissais de les raconter. Quand [**on**]$_{o}$ est si loin de toute amitié, si seule, si triste, toute démarche difficile devient impossible. [**On**]$_{o}$ s'observe, [**on**]$_{o}$ se craint soi-même, et l'[**on**]$_{o}$ se suicide dans la peur de se laisser mourir. »
George SAND, *Pauline*, 1881.

Dans l'exemple [73], il y a une chaîne de coréférence composée de quatre « on ». Ces quatre « on » désignent un référent flou mais sont pourtant coréférents. Ils ont une valeur générique, ce qui permet une distanciation par rapport au référent, mais le fait d'utiliser le féminin pour l'adjectif « seule » inclut le personnage dans la référence en plus de la valeur générique. On retrouve ce phénomène dans l'exemple suivant :

Exemple [74]

« Allons, reprit l'aveugle, qui craignait instinctivement de déplaire à sa fille, en raison du besoin qu'elle avait de son dévouement, laissez-moi le temps de me remettre un peu ; je suis si surprise ! et comme cela, au réveil, **[on]**$_{o}$ ne sait trop ce qu'**[on]**$_{o}$ dit… Je ne voudrais pas vous faire de chagrin, mademoiselle… ou madame… Comment vous appelle-t-**[on]**$_{g}$ maintenant ! »
George SAND, *Pauline*, 1881.

Dans l'exemple [74], les deux premiers « on » coréfèrent de manière stricte et désignent « l'aveugle » tout en gardant une valeur générique, comme si cela pouvait toucher tout le monde au réveil. Le troisième « on » ne coréfère pas avec les autres bien qu'il ait une valeur générique. Il permet surtout d'éviter une tournure passive. Cependant, il inclut peut-être aussi le personnage de l'aveugle car c'est elle qui se demande comment l'appeler.

Le pronom « on » est parfois utilisé pour décrire des situations ou des paysages, il est alors souvent suivi par le verbe « voir »[32]. C'est le cas dans l'exemple suivant :

Exemple [75]

« Elle souleva le rideau et fut tout étonnée de découvrir la sapinière si proche. **[On]**$_{g}$ la voyait bien, la sapinière. **[On]**$_{g}$ la voyait sur une grande étendue, et, tout là-bas, les arbres, par places, se mettaient en rang comme les petites filles à l'école. La voix joyeuse de Noël éclata soudain : — Ah ! enfin **[on]**$_{o}$ voit clair ! »
Marguerite AUDOUX, *Douce Lumière*, 1937.

Il y a une répétition de « On la voyait » dans l'exemple [75] qui appuie la coréférence des deux premiers « on ». Ils ont une valeur générique en raison de la description mais ils sont fortement reliés aux deux personnages présents dans la pièce. Le troisième « on » est du discours direct et inclut le personnage de Noël ainsi que le personnage féminin, mais il a aussi une valeur descriptive. Le « on » de l'exemple suivant possède lui aussi une valeur descriptive :

32. Cela rappelle la structure étudiée par Niziołek (2019) abordée dans la section 4.2.2.

Exemple [76]

« [**Roxane Herbrecht et Nicolas Burcey**]$_n$: "Une explosion [**nous**]$_n$ a réveillés. Le plancher et [**notre**]$_n$ lit se sont soulevés. Une porte est tombée. Cela sentait le brûlé. [**Nous**]$_n$ avons allumé la lumière. [**On**]$_o$ voyait chez [**notre**]$_n$ voisin du dessous. C'est ensuite que [**nous**]$_n$ avons compris que l'explosion venait de chez lui. [**Nous**]$_n$ avons pris peur. [**Nous**]$_n$ avons voulu quitter notre appartement mais il n'y avait plus d'escaliers sur le palier. Alors [**nous**]$_n$ nous sommes mis au balcon et [**nous**]$_n$ avons appelé au secours." »
L'Est Républicain, 02/01/2013.

Il y a une longue chaîne de coréférence dans l'exemple [76] qui désigne Roxane Herbrecht et Nicolas Burcey et qui est principalement composée de plusieurs « nous ». Le pronom « on » permet de prendre un peu de distance en revêtant un rôle descriptif plus générique tout en référant aux deux personnages.

Équivalence avec les pronoms « Tu » et « Vous »

Les pronoms « tu » et « vous » peuvent avoir un fonctionnement similaire à « on » lorsqu'il adopte cette valeur générique/spécifique floue que nous venons d'aborder. Il s'agit principalement d'emplois à l'oral ou dans du discours rapporté comme dans les exemples suivants mais on peut imaginer les trouver à l'écrit.

Exemple [77]

(1) « C'est normal tu sais, quand on fait du sport, on transpire. »
(2) « C'est normal tu sais, quand tu fais du sport, tu transpires. »

L'utilisation du « tu » à la place du « on » dans la phrase (2) de l'exemple [77] ajoute une nuance. Cependant, le premier « tu » ne coréfère pas de manière stricte avec les suivants car ils ont une valeur générique. Cette valeur générique est accentuée par l'utilisation de « quand » dans une phrase exprimant une cause et sa conséquence sous forme de vérité générale.

Inclusion d'un référent dans un autre

Le pronom « on » désigne souvent des groupes, plus ou moins précisément identifiés. Parfois un référent peut en inclure un autre, comme dans l'exemple suivant :

Exemple [78]

« La classe finie, d'autres tourments l'attendaient sur la route qu'elle suivait en compagnie de filles et garçons regagnant leur demeure. Toutes les malices étaient bonnes à faire à cette gnangnan qui ne se défendait pas et ne se méfiait jamais. [**On**]$_{o1}$ la poussait brusquement dans un fossé vaseux, ou dans un buisson plein d'épines d'où elle sortait salie et déchirée. Quand vint la neige, elle fut toute désignée pour recevoir les boules, qu'[**on**]$_{o2}$ lui jetait de préférence au visage. Elle pensait à Noël. S'il était là, il saurait bien la défendre. Mais la ferme des Barray était peu éloignée du village, et Noël n'avait rien à faire sur la route qui conduisait au Verger, distant de plus d'un kilomètre. Il y avait bien Marguerite Dupré, une grande qui prenait parfois sa défense, mais alors c'était elle qu'[**on**]$_{o3}$ attaquait, Marguerite Dupré, dont la maison n'était pas très éloignée de celle de la petite, prenait, en même temps qu'elle, le même sentier. Mais, arrivée là, Douce ne craignait plus rien, elle courait plus vite qu'une oie et devançait facilement les méchants. »
Marguerite AUDOUX, *Douce Lumière*, 1937.

L'exemple [78] contient une chaîne de coréférence qui désigne « les méchants », mais s'agit-il des « filles et garçons regagnant leur demeure » au complet ou seulement de certains ? On a ici aussi une chaîne anaphorique floue. L'exemple suivant génère aussi ce type de flou :

Exemple [79]

« A la Gentiane, fidèle au poste dans la cabane en bois des accompagnateurs en montagne du Ballon, Jean-Louis Fretti lit et note sur son journal les températures et le temps de chaque jour. Le vocabulaire est malheureusement peu varié : pluie, brouillard, vent. Il confirme la présence des touristes : "Les gens ont réservé souvent plusieurs mois à l'avance pour Nouvel An. Ils sont là. Il y avait 150 personnes à Saint-Maurice, dimanche, pour le pot d'accueil des nouveaux arrivants. Le Langenberg est bien rempli. Le potentiel est là...". Reste qu'avec un temps pareil, [**on**]$_{g}$ peut faire un trait sur les sorties en raquette, les parcours VTT et même les sorties pédestres. "Seules les veillées trappeurs fonctionnent. [**Nous**]$_{n}$ en organisons une chaque soir. [**Nous**]$_{t}$ partons par petits groupes et [**marchons**]$_{t}$ environ 1,2 km pour [**nous**]$_{t}$ retrouver dans un chalet en bois où [**on**]$_{o}$ s'éclaire à la bougie autour d'un feu de bois. [**Les gens**]$_{l}$ font des grillades et [**on**]$_{m}$ discute du massif, des légendes locales, des animaux... [...]" »
L'Est Républicain, 02/01/2013.

Dans l'exemple [79], Jean-Louis Fretti parle des veillées trappeurs organisées pour les touristes par les accompagnateurs en montagne du Ballon d'Alsace. Le premier « on » fait partie d'un discours narratif produit par le journaliste et a une valeur générique bien qu'il désigne aussi les accompagnateurs et les touristes du Ballon. Le premier « nous » réfère aux organisateurs mais les deux « nous » qui suivent incluent en plus les touristes (tout comme le sujet 0 de « marchons ». Le « on » qui suit permet d'indiquer qu'il s'agit d'une cabane où il faut s'éclairer à la bougie mais réfère aussi aux personnes qui s'y trouvent (les accompagnateurs et les touristes). Avec « les gens », Jean-Louis Fretti s'exclut de la référence mais le « on » qui suit pourrait bien l'inclure. Cette chaîne de coréférence floue désigne un groupe large, composé des accompagnateurs et des touristes et chaque maillon de la chaîne met l'accent sur une partie de ce groupe ou y ajoute une information. Le flou de l'inclusion d'un référent dans un autre se retrouve aussi dans l'exemple suivant :

Exemple [80]

« [**Toutes ces femmes**]$_f$ se tenaient d'un côté du salon comme un régiment en déroute, et de l'autre côté, entourée de Pauline, de sa mère et de quelques hommes de bon sens qui ne craignaient pas de causer respectueusement avec elle, Laurence siégeait comme une reine affable qui sourit à son peuple et le tient à distance. Les rôles étaient bien changés, et le malaise croissait d'un côté, tandis que la véritable dignité triomphait de l'autre. [**On**]$_f$ n'osait plus chuchoter, [**on**]$_f$ n'osait même plus regarder, si ce n'est à la dérobée. Enfin, quand le départ des plus désappointées eut éclairci les rangs, [**on**]$_d$ osa s'approcher, mendier une parole, un regard, toucher, demander l'adresse de la lingère, le prix des bijoux, le nom des pièces de théâtre le plus à la mode à Paris, et des billets de spectacle pour le premier voyage qu'[**on**]$_d$ ferait à la capitale. »

George SAND, *Pauline*, 1881.

Dans l'exemple [80], les deux premiers « on » sont les anaphoriques de l'antécédent « toutes ces femmes », les deux derniers réfèrent à ce groupe de femmes moins les plus désappointées, ce qui reste flou bien que la coréférence soit stricte entre ces sous-chaînes.

Le pronom « on » peut donc référer à des entité bien différentes et permet de générer du flou référentiel et coréférentiel. Nous avons pu observer que la référence à des groupes flous n'implique pas nécessairement des relations de coréférence floue entre les différents « on » qui peuvent être impliqués.

4.2.3 Le pronom « Ce »

À l'instar des pronoms « on » et « tu » ou encore « il » et « celui-ci » (Demol 2011), le pronom « ce » peut poser des problèmes de coréférence, particulièrement au niveau de l'annotation.

Caractérisation

Selon Riegel, Pellat et Rioul (2004, p. 205-206) dans la la grammaire méthodique du français, il s'agit d'un pronom démonstratif neutre de forme simple, les formes composées étant « ceci » et « cela ». Ils répertorient plusieurs emplois de « ce » :

> « La forme neutre atone *ce* s'emploie d'une part comme **sujet clitique** (elle s'inverse comme les pronoms personnels sujets) **du verbe *être*** éventuellement modalisé par *pouvoir* ou *devoir* (*C'est gentil - Ce devrait être facile*), mais a été progressivement remplacée par *cela*, puis par *ça* (sauf devant le présent de l'indicatif du verbe *être* : **Ça est gentil* (belgicisme), mais *Ça serait gentil*. Elle joue également le rôle d'**antécédent « support non-animé » d'une relative** (*Ce qui se conçoit bien s'énonce clairement*) **ou d'une subordonnée interrogative portant sur le c.o.d.** (*Dis-moi ce qu'il a encore fait*).
>
> La forme *ce* fonctionne régulièrement comme **relais formel et sémantique pour estomper la disconvenance de nombre entre deux groupes nominaux respectivement sujet et attribut**(XI : 6.1) : ?? *Sa passion est les livres / Sa passion, c'est/ce sont les livres*. Elle figure enfin dans **les expressions figées** *sur ce, ce faisant*. »

Ils précisent aussi que la forme « ça » est morphologiquement simple et cependant issue de « cela », c'est pourquoi elle se comporte comme une forme composée. Le premier usage de « ça » qu'ils relèvent est la référence déictique à un référent non catégorisé ou la décatégorisation péjorative d'un référent. Un autre usage de cette forme qu'ils relèvent est l'anaphorisation de propositions, dont l'antécédent ne possède pas de genre ou de nombre. Enfin, « ça » peut aussi permettre de neutraliser le genre ou le nombre d'un antécédent qui est souvent générique, comme dans « Les enfants, ça fait du bruit ».

Référence et coréférence

Certains éléments n'ont pas un accès direct à la référence mais ils peuvent avoir une importance dans les chaînes de coréférence. C'est un point qui a été soulevé dans l'ar-

ticle de Landragin (2011b) avec la distinction entre les maillons forts et les maillons faibles [33] d'une chaîne de coréférence. Les adjectifs « fort » et « faible » s'appliquent à la fois à la nature de l'expression référentielle mais aussi à son degré de contribution à la coréférence. Selon cette théorie, abordée précédemment dans la section 1.2.3, un maillon fort est une expression référentielle qui a un accès direct au référent comme les groupes nominaux, les noms propres ou encore les pronoms. À l'inverse, un maillon faible ne réfère pas directement, il s'agit plutôt d'un indice qui contribue au rappel cognitif du référent, comme par exemple un sujet non exprimé d'un verbe (*sujet zéro*) ou bien l'accord en genre ou en nombre. Étant donnée cette théorie, on peut se demander à quel type de maillon correspond le pronom « ce ». Est-ce qu'il s'agit d'une expression référentielle qui a un accès direct au référent ou est-ce que la référence est moins directe ? [34]

L'annotation de « ce » dans le projet Democrat

Le manuel du projet Democrat (2.3.1) décrit en particulier quelles sont les expressions référentielles à annoter (ou non), cela implique donc de faire certains choix. Ces choix sont le fruit de réflexions prenant en compte des contraintes aussi bien théoriques que pratiques. Il est nécessaire de rappeler que le pronom « ce » n'est pas l'objet d'étude du projet Democrat, le manuel d'annotation prévoit cependant quelques cas particuliers liés au pronom démonstratif neutre (de forme simple ou composée). C'est le cas du présentatif par exemple. Dans « c'est gentil », le « c' » reprend un adjectif. Or, les adjectifs ne sont pas considérés comme étant référentiels, c'est pourquoi ni le « c' » ni « gentil » ne seront annotés comme des mentions dans Democrat. Dans le cas de la construction C' + ÊTRE + QUE comme pour « C'est que je suis le roi », le « c' » n'est pas référentiel et donc pas annoté dans le projet Democrat.

Dans les constructions avec un groupe nominal comme « Un homme est venu, c'était Paul », il n'y a pas de problème car le pronom et le nom sont coréférents et donc annotés comme des mentions. En revanche, la manuel distingue les constructions clivées des pseudo-clivées :

Exemple [81]

(1) « **C'**est [la linguistique]$_x$ qui m'intéresse. »
(2) « [**Ce** qui m'intéresse]$_x$, [**c'**]$_x$est [la linguistique]$_x$. »

Les deux phrases de l'exemple [81] sont issues du manuel d'annotation de Democrat.

33. Définis dans la section 1.2.3.
34. La question de savoir si la prise en compte des maillons faibles doit être faite dans les chaînes de coréférence dépendra des besoins d'un projet par exemple.

Dans le cas de la construction clivée, seul le groupe nominal est annoté alors que dans les pseudo-clivées comme dans la phrase (2), le pronom et le groupe prépositionnel le sont aussi. Dans le cas des phrases clivées, le pronom « ce » élidé n'est pas pleinement référentiel. Il nous paraît cependant important d'annoter ce type de marqueur de la référence, c'est pourquoi il serait donc plutôt à catégoriser dans ce type d'usage comme un « maillon faible » (Landragin 2011b, p. 13).

Le manuel d'annotation de Democrat prévoit aussi le cas des anaphores résomptives dans lesquelles, un pronom démonstratif neutre peut être un anaphorique. Il reprend alors une ou plusieurs propositions [35]. Dans ce cas, il est prévu d'annoter le pronom comme une expression référentielle qui peut potentiellement aussi être reprise, mais son antécédent ne sera pas annoté. C'est un biais du manuel d'annotation qui est le fait de contraintes techniques d'annotation : en effet, il n'est pas évident d'annoter tout un paragraphe avec de potentielles césures comme étant le premier maillon de la chaîne car les outils utilisés permettent une annotation linéaire. Dans l'exemple suivant : « Il a dit qu'il avait cassé le vase en trébuchant. C'est faux ». Il peut sembler évident que les deux « il » coréférèrent en fonction du contexte, sinon il peut y avoir une ambiguïté. Le « C' » réfère à la proposition « il avait cassé le vase en trébuchant » mais seul le pronom est annoté.

Lorsqu'il y a une référence générique, il peut y avoir coréférence (sauf s'il y a un retour au spécifique). Le manuel prévoit d'annoter le groupe nominal et le pronom lorsqu'il y a une référence générique de ce type comme dans « Un chat, ça miaule » ou « Un café, ça réchauffe toujours ».

Le pronom « ce » ne peut pas être impliqué dans des emplois génériques de la forme « ce N, ils » [36] (Guérin 2014). Guérin conclut que son mode de donation [37] qui ne le présente pas uniquement comme faisant partie d'une classe (contrairement à « un » indéfini) ne lui permet pas d'être impliqué dans une anaphore générique de ce type. Ce type d'anaphore n'est de toutes façons pas exactement coréférente.

Le pronom « ce » se trouve aussi dans des expressions figées. De manière générale, il n'y a pas de référence dans une expression figée : dans l'expression « ça me fait une belle jambe », « une belle jambe » ne réfère pas. Ce phénomène est comparable à l'expression « est-ce que » qui est devenue figée avec le temps. Dans ce cas, on n'annotera pas « ce ». Dans l'expression « ce me semble », le « ce » est impersonnel, il ne réfère pas et ne sera donc pas annoté. En revanche, le cas de « ça me semble » est différent car « ça »

35. Le projet CO2 exclue le pronom « ça » et ses dérivés (« cela », « c' » et « ce ») pour cette raison.

36. Comme pourrait l'être « un » : « J'ai adopté un chat, parce qu'ils sont affectueux. » (Guérin 2014) versus *J'ai adopté ce chat, parce qu'ils sont affectueux.

37. Le mode de donation correspond au sens d'une expression (Frege 1892).

reprend souvent toute une situation, dans ce cas le pronom serait plutôt à traiter comme une anaphore résomptive. Dans le cas d'autres expressions comme « Et ce|ça|cela », « ce disant|faisant », « pour ce faire », « sur ce », il s'agit d'expressions figées bien que l'on puisse néanmoins avoir une idée de ce à quoi réfère « ce ».

Hypothèses

Pour dépasser le cadre du manuel d'annotation et continuer la réflexion à propos du pronom « ce » dans les chaînes de coréférence, nous avons formulé, avec Frédéric Landragin, trois hypothèses principales à propos du pronom « ce » dans les chaînes de coréférence que nous avons ensuite tenté de vérifier dans un petit corpus.

La première hypothèse suppose que le pronom « ce » puisse apparaître dans tout type de chaîne et ne concerne pas un type de référent en particulier. En ce sens, il peut être comparable aux autres maillons. Les principaux types de maillons étant les noms propres, des groupes nominaux et des pronoms (dont les concurrents de « ce » : « ça » et « cela » par exemple). La seconde hypothèse suppose que « ce » apparaît essentiellement en première position, qui est une position cruciale, mais jamais au milieu de la chaîne du fait des constructions dans lesquelles il est susceptible d'apparaître (par exemple les présentatifs). La dernière hypothèse est que le pronom « ce » peut très bien être un « maillon fort » parce qu'il s'agit d'un pronom et qu'il peut par exemple être repris par la suite. Cependant, il peut aussi bien être un « maillon faible » qui aura simplement une portée au niveau local.

Exemple [82]

1. « Marthe était à moi ; **ce** n'est pas moi qui l'avais dit, **c'**était elle. » (Le Diable au corps)
2. « Marthe ignorait **ce** que c'est que d'être mutine. » (Le Diable au corps)
3. « La cloche se cassant, le chat en profite, même si **ce** sont ses maîtres qui la cassent et Ø s'y coupent les mains. » (Le Diable au corps)
4. « Car il me semblait que **ce** qui jusqu'ici avait entravé mes désirs, **c'**était la peur du ridicule, de me sentir habillé, lorsqu'elle ne l'était pas. » (Le Diable au corps)
5. « Les chiens de berger ne sont plus **ce** qu'ils étaient... » (L'Est Républicain)
6. « À partir de demain, **ce** sera Marguerite Dupré qui te fera lire. » (Douce Lumière)

7. « Ah ! Dieu ! où suis-je ? est-**ce** un rêve que je fais ? » (Pauline)
8. « Pauline savait toute la vie de Laurence, même **ce** qui ne lui avait pas été raconté, et **cela** plus que tout le reste peut-être. » (Pauline)

Pour vérifier ces hypothèses, nous avons extrait de notre corpus les occurrences du pronom « ce », dont l'exemple [82] constitue une partie. Dans ces phrases, « ce » et ses dérivés peuvent avoir des référents abstraits et des référents concrets. Dans les phrases 1, 3 et 6 le référent est humain. Dans la phrase 5, le référent est animal et dans les autres cas, le référent est abstrait : la peur, la vie ou encore un rêve par exemple. Ainsi, le pronom « ce » peut autant faire partie d'une chaîne de coréférence dont le référent est abstrait que concret. Le pronom « ce » peut par conséquent être comparable aux autres types de maillons d'une chaîne dans le sens où il n'est effectivement pas caractéristique d'un type de référent en particulier. En revanche, nous avons observé que le pronom « ce » était moins fréquent que les autres pronoms (en particulier pour les référents humains) et qu'il apparaît en majorité dans des chaînes courtes voire très courtes. Ce pronom est donc comparable aux autres maillons mais seulement dans la mesure où il ne nécessite pas de méthodologie particulière.

En ce qui concerne la seconde hypothèse, nous supposions[38] que le pronom « ce » apparaissait essentiellement en première position. Dans ces quelques exemples représentatifs de notre corpus, le pronom « ce » figure en première position six fois sur huit. Bien que « ce » soit souvent le premier maillon, il ne semble pas avoir un rôle d'ancrage référentiel, qui est plutôt le rôle du premier maillon car il sert d'introduction marquée d'un nouveau référent. Il a plutôt un rôle de maillon complémentaire, lié à la continuité référentielle et à la cohérence discursive, en particulier en raison du fait qu'il apparaisse à une position proche d'une expression référentielle plus marquée comme un groupe nominal par exemple. À ce propos, la répartition des maillons d'une chaîne peut varier à l'intérieur d'un texte (figures 1.3 et 1.4). La distance intermaillonaire peut être plus ou moins régulière selon les chaînes. Les expressions coréférentes d'une chaîne peuvent apparaître de manière plus ou moins éparse dans le texte ou alors être plus rapprochées par endroits en formant de petits groupes au fil du texte. Le pronom « ce » fait plutôt partie du deuxième cas de figure, car il apparaît souvent proche d'un autre maillon de la même chaîne, et même souvent dans de petites chaînes qui seraient un seul de ces deux petits groupes.

Notre dernière hypothèse supposait que « ce » pouvait être un maillon fort car il s'agit d'un pronom et qu'il peut être repris par la suite. Nous supposions aussi que « ce » pouvait être un maillon faible en raison de sa portée qui est bien souvent seulement au niveau local. On observe que le pronom « ce » apparaît souvent à l'intérieur de construc-

38. Toujours avec Frédéric Landragin.

tions syntaxiques qui comprennent un ou plusieurs maillons de la chaîne. C'est le cas dans la phrase 3 de l'exemple [82], avec « ce sont ses maîtres qui » où la référence est plutôt marquée par le groupe nominal et le pronom relatif. Il y a aussi dans cette chaîne deux autres exemples de maillons faibles : le sujet zéro et le pronom réfléchi clitique (qui n'est d'ailleurs pas annoté dans Democrat). On retrouve le même phénomène dans les autres exemples, comme dans la phrase 7 : « est-ce un rêve que je fais » où la référence est plutôt marquée par le groupe nominal bien qu'introduit par « ce ». Cet autre maillon qui apparaît dans les mêmes constructions que « ce » est souvent un maillon plus « fort » dans le sens où sa catégorie grammaticale est plus marquée d'un point de vue référentiel. C'est le cas des noms propres, des groupes nominaux et des pronoms personnels et relatifs. Le rôle de « ce » apparaît donc plutôt comme étant celui d'un complément syntaxique, parfois nécessaire, et pas réellement celui d'une expression référentielle. D'après ces observations, le pronom « ce » serait donc plutôt un maillon faible. Que l'on peut ignorer ou prendre en compte dans les chaînes de coréférence mais en précisant qu'il s'agit d'un maillon faible.

Nous avons donc observé que le pronom « ce » est comparable aux autres expressions référentielles car il n'est pas caractéristique d'un type de référent. Il apparaît néanmoins plutôt dans des chaînes courtes. Il est le pronom démonstratif le plus fréquent et reste aussi fréquent que la plupart des pronoms personnels [39].

Le pronom « ce » a plutôt un rôle de maillon complémentaire même s'il est souvent en première position de la chaîne et il s'agit donc plutôt d'un maillon faible car il sert souvent de complément syntaxique dans une construction avec une expression référentielle plus marquée au niveau de la référence. La typologie des chaînes de coréférence est un travail en cours dans le projet Democrat avec par exemple le calcul de la distance intermaillonnaire. Cela pourrait permettre de clarifier les typologies de chaînes de coréférence en fonction des types de maillons et donc de clarifier le rôle de « ce », à plus grande échelle dans les chaînes.

Ce travail sur le pronom « ce » dans les chaînes de coréférence mériterait une étude quantitative, avec un corpus plus grand. Cela nous permettrait de mettre en perspective l'importance de ce pronom dans les chaînes de coréférence. Le pronom « ce » gagnerait aussi à être analysé face à des marqueurs linguistiques comparables (comme le pronom « cela » par exemple), afin d'appréhender de manière plus large la distinction entre maillon fort et maillon faible.

39. Dans le sous-corpus de Democrat par exemple (cf. le tableau en annexe 0.1), la simple forme « ce » apparaît 1 160 fois, et « c' » 494 fois, contre 982 pour le pronom « on » (cf. le tableau 4.2).

4.3 Résout-on vraiment les (co)références ?

Il est maintenant établi que certaines expressions référentielles génèrent du flou (co)-référentiel. Ce flou peut jouer un rôle et il n'est pas toujours nécessaire de le dissiper pour pouvoir comprendre un texte par exemple.

4.3.1 L'intérêt du flou

Un parallèle est possible entre le flou (co)référentiel et visuel. Dans les deux cas, la connotation est fréquemment négative. Une photographie floue est bien souvent une photo ratée, quand bien même le moment capturé est celui recherché. Le flou peut aussi être « artistique » lorsqu'il est maîtrisé et volontaire, la connotation est alors positive. En peinture, le fameux « sfumato » de Léonard de Vinci est une technique de peinture qui produit un effet vaporeux apporté notamment par les contours imprécis des formes des éléments. Cette technique est obtenue par l'application d'un glacis[40] et permet de suggérer la perspective et l'atmosphère de la peinture. En photographie ou au cinéma, le flou peut permettre de mettre en avant un sujet net par contraste[41] mais il peut aussi servir à adoucir un élément de la composition visuelle lorsqu'il est utilisé sur un visage lors d'un portrait par exemple. Il peut aussi représenter le mouvement lorsqu'il est dû au déplacement d'un élément ou d'un personnage[42] pour représenter la vitesse d'une voiture ou la la poésie d'une cascade. Si le flou peut être recherché sur le plan visuel, il l'est aussi parfois dans le langage. Comme nous l'avons abordé dans la sous-partie 4.1.1, le flou peut être présent distinctement au niveau de la dénotation, de la référence ou encore de la coréférence.

Au niveau de la communication humaine, Boutet (1986) remarque qu'une de ses caractéristique « réside certainement dans le fait que les échanges se produisent dans une certaine approximation, un certain flou et non dans la clarté ou l'univocité. ». La communication entre des locuteurs ou la compréhension d'un texte par un lecteur n'est donc pas toujours exactement limpide, qui demandent parfois des ajustements, mais qui se poursuivent.

Selon Quine (1977), il est parfois préférable de ne pas « chercher de remèdes à la présence du flou ». Comme sur le plan visuel, cette présence dans le langage est même parfois utile et volontaire. C'est le cas avec les énallages par exemple. L'énallage est une figure de style qui correspond à une transposition de temps, mode, nom ou personne par un autre temps, mode, nom ou personne. En ce qui concerne les expressions référentielles, cela se

40. L'application de couches successives de pigments.
41. Cet effet est appelé « bokeh ».
42. On parle alors de « flou cinétique ».

traduit souvent concrètement par une substitution de pronom ou de nom. Là où certains peuvent voir une faute grammaticale, d'autres soulignent plutôt son utilité pragmatique dans le discours (Bonhomme 2014). En effet, une énallage permet de capter l'attention de l'interlocuteur ou du lecteur en utilisant une formulation que l'on ne s'attend pas à retrouver à cet endroit. De nombreux exemples existent dans les publicités (dont le but est de capter l'attention) comme avec le « Think different » d'Apple pour lequel on attendrait plutôt l'adverbe « differently ». Il y en a aussi régulièrement en politique : « Votez utile ». Cette figure de style est aussi dans les textes narratifs avec des substitutions de personnes comme avec l'usage du présent dans un récit à l'imparfait pour ajouter du relief à la narration et marquer l'aspect soudain d'une action. On retrouve aussi cette figure de style dans le rap, style musical impliquant à la fois l'engagement du « je » et la parole collective du « nous » (Pecqueux 2005) ou du « on ».

Pour Benveniste (1966, p. 174), les pronoms de la troisième personne diffèrent de ceux de la première et de la deuxième personne par leur nature et leur fonction. La troisième personne représente « l'absent » et correspond plutôt à une « non-personne ». Les énallages qui passent de la « personne » (première et deuxième) à la « non-personne » (troisième) ajoutent une distance avec la personne désignée. Ce phénomène est fréquent avec le passage de la deuxième à la troisième personne (ou au « on »). Il peut alors avoir divers objectifs distincts. Ils peuvent servir à rabaisser ou humilier comme lorsqu'un supérieur hiérarchique adresse un « Alors, on prend son aprèm ? » à un employé qui quitte le travail à 17h30. Au contraire, ils peuvent aussi avoir une fonction bienveillante dans les énoncés hypocoristiques du type « On avait un gros malheur ? » adressé à un enfant qui pleure. Cette fonction est véhiculée par l'usage de l'imparfait, qui met la notion de malheur à distance temporellement, comme si le chagrin était déjà passé. L'usage du pronom « on » à la place de la deuxième personne atténue aussi la notion de malheur en la généralisant.

Qu'il soit présent sous la forme d'une énallage ou non, le flou référentiel peut s'avérer utile dans le discours et est employé plus souvent qu'on ne le pense. C'est le cas avec les « on » de l'exemple [65] où ce qui prime est que le travail soit fait par « quelqu'un » d'autre et non l'identification précise du référent.

Il n'est pas toujours utile de tenter de résoudre le flou. Comme le montrent les expériences de Ferreira, Bailey et Ferraro (2002), les lecteurs peuvent tout à fait se contenter d'énoncés imprécis et poursuivre leur lecture sans que cela ne leur pose de problème.

4.3.2 La théorie Good-enough

Lorsqu'un groupe est désigné de manière floue ou plus généralement lorsque le référent d'une expression est difficile à identifier de manière exacte, le cerveau humain est-il perturbé ? A-t-il réellement besoin de pouvoir identifier le référent de chaque expression référentielle avec précision ?

Les résultats des expériences en psycholinguistique sur les représentations « good-enough » de la compréhension du langage ont prouvé que les sujets sont capables de traiter certaines expressions de manière superficielle. La première expérience en question (Ferreira, Bailey et Ferraro 2002) a montré grâce à des enregistrements de mouvements oculaires que les lecteurs pouvaient avoir des difficultés à interpréter correctement les phrases à « garden-path ». Ce sont des phrases dont la première partie envoie le système de compréhension du langage dans une mauvaise direction par rapport au sens de la phrase. La seconde expérience (Ferreira et Stacey 2000) a montré que la structure des phrases passives est fragile car elle nécessite d'assigner les rôles sémantiques dans un ordre atypique : le patient avant l'agent. Étant donné que ces structures sont fragiles, elles ont besoin d'être renforcées par la suite. Si elle ne le sont pas, la compréhension de la phrase a de forte chances de ne pas être correcte.

4.3.3 Application à la (co)référence

Ces résultats nous font repenser la manière de modéliser les chaînes de coréférence dans les textes. Il serait intéressant de se pencher sur la résolution de chaînes de coréférence en prenant en compte cette théorie selon laquelle la compréhension du langage – et donc la résolution d'une (co)référence – serait parfois simplement « bien assez bonne ». C'est un sujet qui a été abordé par Charolles (2014, p. 7) avec la question de la profondeur de traitement des annotations des expressions référentielles.

Landragin (2007) dans un article sur les anaphores à antécédent flou cite Prandi (1987, p. 135) au sujet des ambiguïtés structurales liées à la portée des prépositions « ce genre d'ambiguïté structurale, très fréquent dans l'expression nominale, reste d'ailleurs sans conséquences sur l'interprétation, du fait, probablement, de son caractère systématique et non annulable ».

La théorie good-enough permet d'avancer que la compréhension partielle d'un texte n'empêche pas le bon déroulé de la lecture. Cela implique que le lecteur n'a pas toujours besoin de résoudre le flou pour avancer dans un texte. Au niveau de l'annotation, cela nous permet de donc de repenser l'annotation des chaînes de coréférence en prenant en compte cette théorie.

Recasens, Tolchinsky et Marti (2014) ont testé le concept de *near-identity*[43] d'un point de vue psycholinguistique. Leur hypothèse de départ est que le traitement de la coréférence de manière binaire est une « simplification excessive » (*oversimplification*) et que la prise en compte d'un degré de coréférence intermédiaire (la coréférence proche) est nécessaire. Les résultats de leurs expériences ont montré qu'il existe une classe en dehors de la coréférence stricte et de la non coréférence. Les participants ont parfois annoté les expressions de cette classe de coréférence proche comme étant coréférentes et parfois non et selon des degrés de coréférence allant de 1 à 4. Les cas causant le plus de désaccords ont généré des temps de réaction élevés. Les auteurs en ont donc conclu que cette catégorie de coréférence proche entraîne une complexité de traitement plus élevée. Ils ont aussi observé que la *near-identity* est plus proche de la coréférence que de la non coréférence. Enfin, ils reconnaissent aussi que la notion de coréférence proche est connexe aux notions d'ambiguïté et de vague.

Ødegaard (2006, p. 99) avait tenté une classification des valeurs de « on » et en a conclu que : « Il est souvent difficile, voir impossible, de placer *on* dans telle ou telle catégorie, ce qui est dû notamment au fait que les frontières entre les valeurs sont floues, mais également à la référence vague du pronom. ». La solution serait-elle par conséquent de tolérer que les expressions référentielles et leurs relations ne puissent pas toujours rentrer dans des cases ? Est-ce qu'il pourrait être envisageable d'avoir une « case » floue, ou un trait spécifique pour les relations de coréférence floue ?

Le flou référentiel et coréférentiel peut apparaître sous différentes formes. Il existe des cas typiques de ce phénomène qui sont : les groupes pluriels et certains pronoms, comme « on » et « ce » en particulier. Dans chacun de ces cas de figure, le flou se manifeste de différentes manières. Il s'agit souvent d'une superposition de sens potentiels, relevant plutôt de la sur-détermination. Cette superposition implique plusieurs référents, correspondant souvent à des groupes, déjà parfois flous eux-même. Elle peut aussi régulièrement impliquer une référence générique qui se mêle à une référence spécifique. Des expériences psycholinguistiques nous ont montré qu'il n'est pas toujours nécessaire de résoudre ce flou pour comprendre un texte. Une autre manière d'appréhender l'annotation de ce phénomène en corpus est sûrement possible, en prenant en compte le fait que le flou coréférentiel ne doit pas toujours être levé mais plutôt caractérisé comme tel.

43. Voir section 3.3.1

Conclusion de la partie 2

La coréférence correspond à un cadre strictement établi. Cependant, ces deux chapitres abordent des cas de figure plus subtils, correspondant à la complexité et à la richesse de la langue, qu'il n'est pas toujours aisé de catégoriser. Dans le cas des ambiguïtés référentielles, un choix entre plusieurs référents est possible. Des critères de saillance peuvent parfois aider à les résoudre. Pour la coréférence proche, le choix n'est pas aisé car les référents sont proches, bien que précisément identifiés. Dans le cas des référents évolutifs, une zone de flou s'installe, entre l'ambiguïté et la coréférence proche. Dans les cas de coréférence floue, le choix n'est pas possible, en raison du flou et de la proximité sémantique entre les référents. Ce phénomène apparaît en particulier avec les groupes pluriels et certains pronoms comme « on » et « ce » ainsi que leurs dérivés. Un lecteur n'a pas toujours besoin de choisir un référent pour avancer dans sa lecture, pourquoi imposer un choix de référent au moment de l'annotation s'il est impossible à réaliser en pratique ? De plus, prendre en compte ce type de phénomène dans l'annotation pourrait aider des systèmes de traitement automatique de la langue à mieux intégrer les subtilités du langage.

Troisième partie

De la linguistique de corpus vers le TAL

De nombreux corpus existent pour la coréférence stricte mais de plus en plus de projets prennent en compte des cas de coréférence non stricte. Comment les phénomènes impliqués sont-ils traités dans l'annotation de corpus ?

Une analyse de ces phénomènes en corpus, notamment dans un cadre strict comme le projet Democrat, sera utile pour juger de l'intérêt de prendre en compte les phénomènes liés à la coréférence non stricte au niveau de l'annotation.

Des propositions pour préciser le cadre du projet Democrat sont nécessaires, d'une part pour s'assurer du respect de la coréférence stricte dans les chaînes. D'autre part pour permettre d'indiquer lorsqu'une relation de coréférence est stricte ou floue. Sans ces précisions, des confusions sont possibles.

Les confusions générées par un manque de précision du cadre peuvent perturber les résultats de systèmes de détection automatique de la coréférence car les annotations du corpus manquent de justesse. Cependant la prise en compte de la coréférence floue dans un système automatique entraîne aussi des conséquences. C'est ce que nous proposons d'explorer dans les chapitres suivants : le chapitre 5 traitera de l'annotation de la coréférence non stricte et floue, le chapitre 6 proposera des recommandations pour le traitement de ce phénomène en corpus.

Chapitre 5

La coréférence non stricte et floue à l'épreuve de l'annotation

Dans les approches classiques impliquant des corpus annotés en coréférence, il y a construction d'une chaîne seulement lorsque la coréférence est stricte : le référent est exactement le même pour tous les maillons sans qu'aucun doute ne soit possible. Les deux chapitres précédents ont abordé des phénomènes de référence générique, ambiguë, vague ou floue et de coréférence non stricte comme la coréférence proche et la coréférence floue. Est-il possible d'intégrer ces phénomènes dans l'annotation des chaînes de coréférence ?

5.1 L'annotation de la coréférence non stricte

Plusieurs corpus ont tenté de prendre en compte des relations entre référents qui ne sont pas exactement coréférents. Trouver un cadre pour la coréférence non stricte n'est pas évident, malgré un cadre précis et des relations bien identifiées.

5.1.1 Des tentatives dans différents projets

Certains projets offrent la possibilité d'indiquer lorsque la référence n'est pas *exactement* mais *quasiment* identique. Il ne s'agit pas toujours des mêmes phénomènes dans tous ces projets. Il n'est pas inutile de le rappeler : l'annotation d'un corpus nécessite toujours de faire des choix concernant les phénomènes à annoter mais aussi la manière de les annoter. Ces choix sont motivés par un cadre théorique et des contraintes techniques. Les projets et corpus suivants permettent une annotation des relations de coréférence en prenant en compte des cas de coréférence non stricte.

Le programme ACE (Automatic Content Extraction) (Doddington et al. 2004) s'intéresse à un certain type d'entités bien classifiées. Elles ne correspondent pas exactement à toutes les expressions référentielles. Les relations entre ces entités et les événements auxquels elles prennent part sont aussi prises en compte. Dans le projet, ils distinguent

cinq types de relation (qui sont aussi sous-catégorisées en 24 types/sous-types au total) : le rôle, la relation partie-tout, la localisation, la proximité et les relations sociales. En plus de l'annotation des relations, il y a une tâche d'annotation de la coréférence avec la possibilité d'indiquer lorsqu'il y a des cas de métonymie — lorsqu'un concept est désigné par un autre concept avec lequel il entretient une relation, le lien de coréférence n'étant pas strict dans ce cas. Le corpus est en anglais et comporte aussi de l'arabe et du chinois. L'annotation est au format XML pour indiquer les relations : leurs arguments sont les entités et les attributs sont les types de relation.

Le projet OntoNotes (Pradhan et al. 2011) s'intéresse entre autres à la coréférence pour les entités et les événements. Il y a une distinction entre deux types de coréférence : la coréférence « identique » (sans prise en compte des génériques, des sous-spécifications ou des résumés) et la coréférence « appositive » (qui est annotée comme un attribut comme dans « John, un linguiste, est venu à la conférence »). Dans le corpus Democrat par exemple, l'apposition n'est pas une mention à part entière et ne figure pas dans la mention.

Le schéma d'annotation du corpus WikiCoref (Ghaddar et Langlais 2016), un corpus constitué de documents du Wikipédia anglais, se base sur celui d'OntoNotes mais avec quelques différences. Il y a une distinction entre la coréférence identique, la coréférence attributive (pour les appositions) et la coréférence attributive dans des constructions copulatives (comme dans « Macron est le président de la république »). L'annotation de ce corpus est réalisée avec l'outil MMAX2 (Müller et Strube 2006).

La typologie NIDENT proposée par Recasens, Hovy et Marti (2010) et évoquée précédemment dans la section 3.3.1 prend en compte des relations de coréférence stricte et proche. Les trois types de relation de cette typologie sont : la non-identité, l'identité et l'identité proche. L'identité correspond à la coréférence stricte car le référent est exactement le même. La coréférence proche comporte quinze types de relations différentes réparties en quatre catégories : la métonymie, la méronymie, la classe et la fonction spatio-temporelle.

Pour le polonais, le Polish Coreference Corpus (Ogrodniczuk, Glowinska et al. 2014) a distingué les relations de coréférence identique et quasi-identique en s'inspirant des travaux de Recasens sur la *near-identity*.

La première version du corpus ARRAU [1] (Poesio et Artstein 2008) en 2008 prend en compte l'ambiguïté référentielle. C'est aussi le cas du corpus Phrase Detectives (Chamberlain, Poesio et Kruschwitz 2016) pour le projet AnaWiki, qui a été annoté grâce à la collaboration sur le web pour en faire un corpus de grande taille. Dans ces deux corpus les ambiguïtés référentielles sont annotées en se basant sur l'avis des annotateurs

1. « Anaphora Resolution and Underspecification ».

pour donner un score aux alternatives possibles.

Pour le français oral, le corpus ANCOR (Muzerelle, Lefeuvre, Antoine et al. 2013) a distingué les anaphores directes (ou fidèles), indirectes (ou infidèles), pronominales, associatives et associatives pronominales.

Dans le corpus Democrat, les sujets zéro sont annotés comme des mentions à part entière bien qu'ils ne soient pas présents physiquement. Les référents évolutifs sont aussi annotés dans une seule chaîne de coréférence malgré un changement d'état ou de forme (voire d'identité) du référent.

Tous ces corpus opèrent une distinction entre une relation de coréférence identique et une relation de coréférence qui n'est pas complètement identique mais clairement identifiée. Ces distinctions reposent sur des critères linguistiques précis pour des relations facilement identifiables en théorie. Cependant, aucun ne prend en compte la coréférence floue ni le doute que peut générer pour l'annotateur la relation entre des référents proches sémantiquement.

5.1.2 Au-delà de la coréférence proche

La coréférence proche concerne des référents proches sémantiquement mais qui sont bien identifiés. L'ambiguïté concerne des référents pour lesquels un doute est possible mais qui ne sont pas proches sémantiquement. La coréférence floue est à la croisée de ces deux concepts : les référents potentiels sont proches sémantiquement mais un doute existe dans l'identification du référent.

Nous avons observé dans le chapitre précédent que certains pronoms sont particulièrement porteurs de flou. Tout comme une succession de « on », les « nous » peuvent aussi désigner des groupes flous et donc potentiellement entretenir des relations de coréférence floue. Dans l'exemple suivant, il y a une chaîne de coréférence composée de cinq « nous » dont le référent est un groupe flou incluant le narrateur :

Exemple [83]

« [**Nous**]$_n$ allions chaque jour, après dîner, à la gare de J…, à deux kilomètres de chez [**nous**]$_n$, voir passer les trains militaires. [**Nous**]$_n$ emportions des campanules et [**nous**]$_n$ les lancions aux soldats. Des dames en blouse versaient du vin rouge dans les bidons et en répandaient des litres sur le quai jonché de fleurs. Tout cet ensemble me laisse un souvenir de feu d'artifice. Et jamais autant de vin gaspillé, de fleurs mortes. Il fallut pavoiser les fenêtres de [**notre**]$_n$ maison.

Bientôt, [**nous**]$_n$ n'allâmes plus à J… Mes frères et mes sœurs commençaient

d'en vouloir à la guerre, ils la trouvaient longue. »
Raymond RADIGUET, *Le Diable au corps*, 1923.

Le référent du groupe flou peut probablement être la famille du narrateur, mais est-ce que cela inclut aussi les parents ou simplement les enfants ? Est-ce qu'il s'agit du même groupe à chaque fois ? Le fait de ne pas avoir de réponse claire à ces questions ne perturbe pas réellement la lecture. Ces questions se posent lors de l'annotation. L'exemple [83] est issu du corpus Democrat et la chaîne de coréférence floue a donc été traitée comme une nouvelle chaîne indépendante des autres bien qu'elle soit reliée sémantiquement à la chaîne de coréférence comprenant le narrateur et sa famille. Quelques chaînes qui pourraient correspondre à la famille du narrateur sont visibles dans la figure suivante :

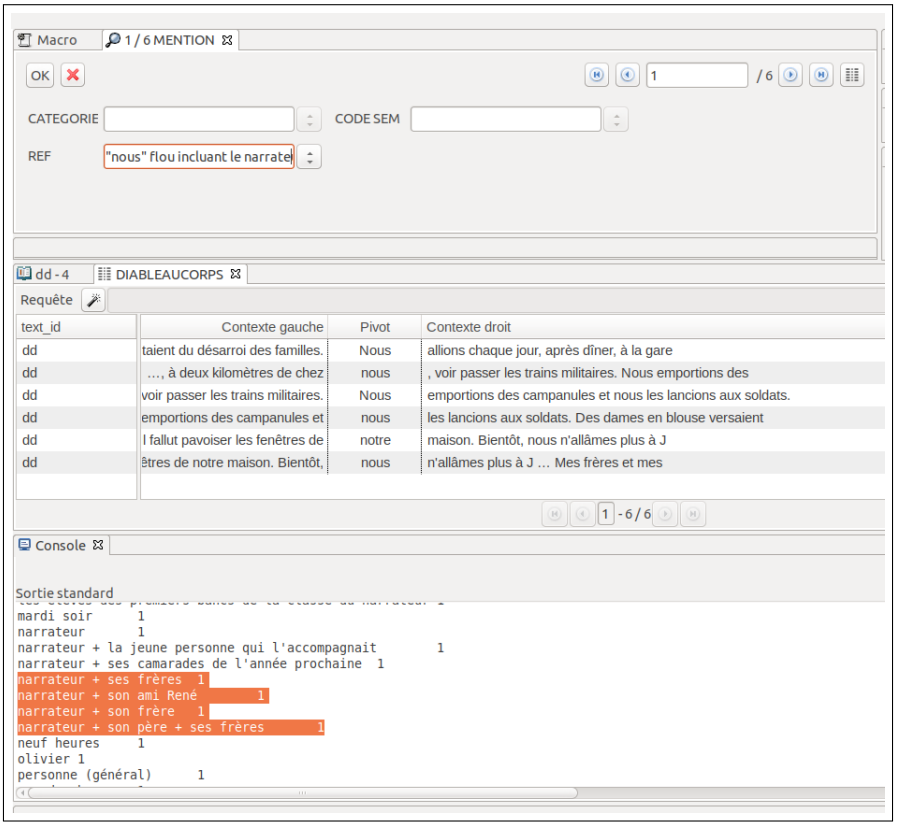


Figure 5.1 – Des chaînes proches mais séparées dans l'annotation d'un bloc de Democrat.

L'annotation présente dans la figure [5.1] a été réalisée au début de ce travail de thèse, sans que les phénomènes de coréférence non stricte ou floue n'aient été encore évoqués et avec une volonté de rester dans le cadre de la coréférence stricte : une chaîne = un référent. Si un doute est présent, une nouvelle chaîne est créée. Dans le cas de cette annotation, le doute réside dans le manque de précision avec lequel le narrateur désigne sa famille, d'autant plus qu'il ne précise pas quels sont les membres qui la composent. Cet exemple est représentatif des chaînes qui pourraient être reliées sémantiquement, sans relever de la coréférence stricte ou proche pour autant. En effet, l'identification du référent

et des relations entre les expressions suscite le doute en raison du flou de la description du référent. Le cadre strict du projet Democrat implique de créer une nouvelle chaîne. Cependant, relier ces chaînes tout en indiquant que la coréférence n'est pas stricte pourrait constituer une solution pour ne pas perdre cette information. L'absence de solution dans le projet Democrat a mené les annotateurs à faire des choix d'annotation qui ne correspondent pas à de la coréférence stricte.

5.2 Différentes conduites d'annotation de la coréférence non stricte dans le corpus Democrat

Le flou référentiel ne pose pas de problème au lecteur qui se satisfait aisément de cette imprécision parfois volontaire dans la narration. En revanche, ce phénomène est problématique lorsqu'il s'agit de mettre des étiquettes sur les expressions référentielles comme c'est le cas dans l'annotation d'un corpus en coréférence tel que Democrat. Malgré un cadre strict du projet, il réside une variabilité dans l'annotation de ce phénomène qui nécessite un cadre spécifique le prenant en compte. Les 58 textes du corpus Democrat ont été annotés par 31 collaborateurs du projet aux statuts variés, mais toujours avec des connaissances en linguistique. C'est notamment cette variété d'annotateurs qui a nécessité un manuel d'annotation fourni et précis. Malgré tout, le flou référentiel a généré des variations d'annotation inter mais aussi intra-annotateur [2]. Nous avons relevé diverses approches d'annotation de coréférence floue dans ce corpus que nous avons ensuite catégorisées.

Pour analyser l'annotation de la coréférence floue dans notre sous-corpus [3], le point de départ a été le constat effectué dans le chapitre précédent. En effet, comme nous l'avons vu au chapitre 4, il existe des cas particuliers propices au flou, comme les groupes pluriels ou l'emploi de certains pronoms (en particulier « on » et « ce » ainsi que leurs dérivés). Pour trouver les chaînes de coréférence impliquant un phénomène de flou (co)référentiel, la première étape a donc été de rechercher, à l'aide de TXM, les chaînes dont la dénomination (le label donné par l'annotateur) comporte ces pronoms ou commence par un déterminant au pluriel. En l'absence d'indices dans le label de la chaîne, il a fallu effectuer une recherche des pronoms en question dans le texte pour ensuite relever la manière dont ils ont été annotés. Cette méthodologie a fait remonter une observation principale : il peut y avoir une confusion entre les référents génériques, indéfinis ou flous.

2. On n'est pas toujours d'accord avec ses propres annotations quelques temps plus tard. C'est une remarque vérifiée personnellement au cours de ce travail de thèse.

3. Ce sous-corpus est présenté en annexe 0.1.

5.2.1 Regrouper les référents génériques

Une première approche relevée dans la première version du sous-corpus Democrat consiste à regrouper les référents qui possèdent une valeur générique, en fonction de leur forme linguistique. Le bloc de texte « Aden Arabie » possède trois chaînes correspondant à ce critère : « Ongenerique »[4] (102 maillons), « VousGenerique » (32 maillons) et « hommegenerique » (7 maillons). La chaîne des « on » génériques comporte essentiellement des maillons correspondant au pronom « on », comme dans les exemples suivants :

Exemple [84]
« [**On**]$_{\text{Ongenerique}}$ parle autour de moi du départ, [**on**]$_{\text{Ongenerique}}$ me fait des recommandations, je respire dans un vertige que je devais trouver agréable. [**On**]$_{\text{Ongenerique}}$ me dit adieu, je file comme un mort. »
Paul NIZAN, *Aden Arabie*, 1931.

Dans l'exemple [84], les trois « on » sont annotés avec le même référent « Ongenerique ». Cependant, leur valeur n'est pas réellement générique mais plutôt indéfinie. Ils sont tous inclus dans une chaîne de coréférence stricte. On peut cependant se demander s'il ne s'agit pas d'une chaîne de coréférence floue car le référent peut être le même pour ces trois mentions mais sans qu'il n'y ait de certitude qu'il s'agisse exactement des mêmes personnes à chaque fois. Dans le même texte, une autre succession de « on » pose problème quelques paragraphes plus loin :

Exemple [85]
« Mais quels cadeaux fait l'océan quand les jours ont passé, quand [**on**]$_{\text{Ongenerique}}$ a coupé tant de fuseaux horaires qu'[**on**]$_{\text{Ongenerique}}$ s'embrouille dans ses calculs si l'[**on**]$_{\text{Ongenerique}}$ veut savoir ce que font vos amis à Paris, s'ils dorment ou s'ils mangent ? [**On**]$_{\text{Ongenerique}}$ peut dire qu'[**on**]$_{\text{Ongenerique}}$ est hors d'atteinte, matériellement invulnérable. Il ne faut pas chercher midi à quatorze heures : cela signifie quelque chose de tout à fait simple et important, que les armatures de l'ancien esprit sont perdues : il faudra lui en trouver d'autres et la découverte ne va pas de soi. »
Paul NIZAN, *Aden Arabie*, 1931.

Les cinq « on » de l'exemple [85] sont aussi annotés avec le même référent « Ongenerique » et sont donc aussi inclus dans la chaîne de coréférence mentionnée dans l'exemple [84]. Cette relation est problématique car il n'y a aucun lien de coréférence entre les « on » de cet exemple et ceux de l'exemple précédent. Les trois premiers « on » de l'exemple [85] seraient plutôt à regrouper dans une unique chaîne de coréférence stricte

4. Ce nom est le label donné par l'annotateur au référent des expressions référentielles correspondantes.

car ils ont un lien de coréférence et assurent ainsi une continuité référentielle précise. Les deux « on » suivants sont aussi génériques et coréférents, mais sont-ils coréférents de manière stricte avec les trois précédents ? Dans l'exemple [84], il ne s'agit pas réellement de référence générique mais plutôt de l'identification d'un groupe flou, lui-même désigné de manière floue. Toujours dans le même texte, quelques paragraphes plus loin, deux nouveaux « on » sont annotés avec le référent « Ongenerique » :

Exemple [86]

« **Les voyageurs** sont condamnés à ne voir des maisons où vieillissent les hommes sédentaires que des murs de toutes les couleurs, avec des curiosités simplement architecturales. Je fus ce voyageur : circuler sur de petits vapeurs écaillés, sur des dhows indigènes de l'un à l'autre bord de ce profond canal des enfers, rebondir sur les remparts de l'Afrique et de l'Arabie, ces mouvements du désordre n'imitent pas longtemps les allures de la liberté. [**On**]$_{\text{Ongenerique}}$ sent une espèce de boule de métal qui tourne à l'intérieur de la vie : elle heurte les organes, plus [**on**]$_{\text{Ongenerique}}$ remue, plus elle les blesse. »
Paul NIZAN, *Aden Arabie*, 1931.

Dans l'exemple [86], il s'agit bien de deux « on » possédant une valeur générique. En revanche, avec cette annotation ils se retrouvent eux aussi dans la même chaîne de coréférence que les « on » des deux exemples précédents. Pourtant, ils ne sont pas coréférents avec les « on » de l'exemple [84], ni avec ceux de l'exemple [85], bien que certains soient génériques aussi. Les deux « on » de l'exemple [86] peuvent aussi coréférer de manière floue avec les voyageurs mais aussi avec le narrateur en raison de la succession du syntagme nominal « Les voyageurs » et du syntagme verbal « Je fus ce voyageur ». Le premier évoque un type de personnes de façon générique et le second syntagme revient au narrateur plus précisément. L'utilisation du « on » permet ensuite de reprendre ces deux entités de manière floue, la coréférence entre les deux « on » étant stricte en revanche.

Cette approche a le mérite de vouloir identifier les pronoms « on » dont la valeur est générique. Cependant, les maillons identifiés dans cette chaîne ne sont pas tous génériques. Certains sont indéfinis ou entretiennent des liens de coréférence floue. La valeur générique liée aux pronoms semble difficilement pouvoir être présente dans de longues chaînes de coréférence. En revanche, avec un groupe nominal générique comme « l'homme », cela semble plus évident. La chaîne « hommegenerique » issue du même bloc de texte s'étend de la page 4 à la page 17 et réfère à l'homme en général. Placer tous les « on » génériques dans une même chaîne en fait des mentions coréférentes. Cependant, ils ne le sont pas tous pour autant. Cette approche présente donc deux problèmes principaux.

Cette démarche a aussi été adoptée au début de ce travail de thèse pour le premier bloc

annoté (« Le Diable au corps ») qui contient deux chaînes de coréférence dans lesquelles les maillons ne sont pas tous coréférents. La première chaîne concerne deux « nous » génériques et la seconde concerne 16 « on » génériques qui ne sont pas tous coréférents.

Dans bloc de texte « Le Capitaine Fracasse », il y a aussi une longue chaîne de coréférence de 44 maillons contenant principalement des « on », un « son » et un « vous »[5]. Étant donnée la longueur de la chaîne, il semble peu probable que tous les maillons soient effectivement coréférents. Certains « on » permettent la description lorsqu'ils sont suivis des verbes suivants à l'imparfait : « voir », « croire », « se trouver », « déboucher ». Ils peuvent alors être coréférents. Le lien de coréférence entre deux « on » est encore plus évident dans le cas d'énoncés impliquant la conjonction « si » comme dans l'exemple suivant :

Exemple [87]
« Pourtant, si l'**[on]**$_{\text{référent générique}}$ eût persisté, sans redouter les égratignures des broussailles et les soufflets des branches, à suivre jusqu'au bout l'antique allée devenue plus obstruée et plus touffue qu'une sente dans les bois, **[on]**$_{\text{référent générique}}$ serait arrivé à une espèce de niche de rocaille figurant un antre rustique. »
Théophile GAUTIER, *Le Capitaine Fracasse*, 1863.

Dans l'exemple [87], les deux « on » sont sans aucun doute coréférents en raison de la structure qui implique une subordonnée conditionnelle (*protase*) introduite par la conjonction « si », suivie d'une conséquence (*apodose*). Les deux étant liées. Dans cette chaîne, d'autres maillons ne sont pas coréférents entre eux ni aux premiers, comme ceux des deux exemples suivants :

Exemple [88]
« — **[On]**$_{\text{référent générique}}$ ne saurait mieux élucider mes paroles, répondit l'acteur, et vous parlez de cire. »
Théophile GAUTIER, *Le Capitaine Fracasse*, 1863.

Dans l'exemple [88], le « on » se trouve dans un dialogue et inclut le locuteur (l'acteur). Il a certainement une valeur générique mais il ne peut pas être coréférent aux autres maillons de la chaîne « référent générique ». Il devrait plutôt être un singleton car il ne coréfère en fait à aucune autre expression référentielle dans le texte. C'est aussi le cas du « on » de l'exemple suivant :

5. Il semble que ce maillon doive plutôt être un singleton.

Exemple [89]

« Les principaux emplois de la comédie s'y trouvaient représentés, et, s'il manquait un personnage, [**on**]$_{\text{référent générique}}$ racolait en route quelque comédien errant »

Théophile GAUTIER, *Le Capitaine Fracasse*, 1863.

Dans l'exemple [89], le pronom « on » est indéfini. Il semble désigner des personnes à la tête de la comédie, probablement les gens qui s'occupent de recruter le personnel. Ce « on » permet de ne pas nommer explicitement les personnes désignées et fonctionne à la manière d'un passif : c'est l'action effectuée qui compte et non qui la réalise. En tout état de cause, ce « on » devrait lui aussi être un singleton et car il ne coréfère avec aucune autre mention et encore moins avec celles de la chaîne dont le référent est « référent générique ».

L'annotation du bloc de texte de « La Morte amoureuse » regroupe aussi des référents génériques qui ne sont pas nécessairement coréférents. Il existe une chaîne « on générique » (30 maillons) et une chaîne « vous générique » (4 maillons). Le problème de ces deux chaînes reste le même : tous leurs maillons ne sont pas coréférents. Dans l'exemple suivant, les trois « vous » génériques (page 24) sont effectivement coréférents :

Exemple [90]

« Elle [**vous**]$_{\text{vous générique}}$ faisait commettre avec elle l'infidélité que [**vous**]$_{\text{vous générique}}$ eussiez commise avec d'autres, en prenant complétement le caractère, l'allure et le genre de beauté de la femme qui paraissait [**vous**]$_{\text{vous générique}}$ plaire. »

Gustave FLAUBERT, *Bouvard et Pécuchet*, 1881.

Cependant, ces trois « vous » de l'exemple [90] ne coréfèrent pas avec le premier maillon de la chaîne « vous générique » qui apparaît 19 pages plus tôt, à la page 5, dans l'exemple suivant :

Exemple [91]

« J'étais, tout éveillé, dans un état pareil à celui du cauchemar, où l'on veut crier un mot dont [**votre**]$_{\text{vous générique}}$ vie dépend, sans en pouvoir venir à bout. »

Gustave FLAUBERT, *Bouvard et Pécuchet*, 1881.

Les 30 maillons de la chaîne « on générique » ne sont eux non plus pas tous coréférents. Deux des maillons de cette chaîne correspondent au pronom « votre » possédant une valeur générique comme le montre l'exemple suivant :

Exemple [92]

« Être prêtre ! c'est-à-dire chaste, ne pas aimer, ne distinguer ni le sexe ni l'âge, se détourner de toute beauté, se crever les yeux, ramper sous l'ombre glaciale d'un cloître ou d'une église, ne voir que des mourants, veiller auprès de cadavres inconnus et porter [**soi-même**]$_{\text{on générique}}$ [**son**]$_{\text{on générique}}$ deuil sur [**sa**]$_{\text{on générique}}$ soutane noire, de sorte que l'[**on**]$_{\text{on générique}}$ peut faire de [**votre**]$_{\text{on générique}}$ habit un drap pour [**votre**]$_{\text{on générique}}$ cercueil ! »
Gustave FLAUBERT, *Bouvard et Pécuchet*, 1881.

Dans l'exemple [92], « soi-même », « son » et « sa » ont une interprétation générique (un prêtre en général). Ces trois maillons sont bien coréférents. Cependant le « on » a plutôt vocation à désigner un référent extérieur, plutôt indéfini, mais pas coréférent aux trois premiers maillons. Les deux « votre » semblent quant à eux coréférents entre eux et avec les trois premiers maillons, mais éventuellement de manière floue car l'énallage de changement de personne ajoute une subtilité dans l'interprétation. De plus, les autres maillons de la chaîne « on générique » de ce bloc de texte ne sont pas coréférents aux maillons de l'exemple [92].

Le bloc de texte « Sarrasine » possède aussi la même approche d'annotation : il y a une chaîne « nous générique » (18 maillons), une chaîne « vous générique » (16 maillons) et une chaîne « on générique » (11 maillons). Dans chacune de ces chaînes, tous les maillons ne sont pas nécessairement coréférents.

Cette première approche consiste donc à effectuer des regroupement d'expressions dont le référent est générique[6] alors qu'elles ne sont pas toutes coréférentes. De plus, il existe parfois une confusion entre un référent générique, qui réfère à une catégorie générale de référent (qui peut référer à tout le monde avec de potentielles restrictions), et un référent indéfini, qui désigne un référent non identifié.

5.2.2 Regrouper les référents indéfinis

Dans le bloc de texte « Bouvard et Pécuchet », une chaîne de 42 maillons possède un « référent indéfini », elle ne comporte que des « on ». Dans l'exemple suivant, le pronom « on » se situe dans un dialogue, au discours direct :

6. En tentant la plupart du temps de séparer les « on », les « nous » et les « vous » génériques alors que leur forme ne les empêche pas d'être coréférents.

Exemple [93]

« — Tiens, dit-il, nous avons eu la même idée, celle d'inscrire notre nom dans nos couvre-chefs.
— Mon Dieu, oui, [**on**]$_{\text{référent indéfini}}$ pourrait prendre le mien à mon bureau !
— C'est comme moi, je suis employé. »
Gustave FLAUBERT, *Bouvard et Pécuchet*, 1881.

Dans l'exemple [93], le pronom « on » désigne en effet un référent qui n'est pas défini. Il peut s'agir de n'importe quelle personne au bureau du personnage Pécuchet. Dans l'exemple suivant, le « on » est aussi annoté dans la chaîne « référent indéfini » :

Exemple [94]

« [**On**]$_{\text{référent indéfini}}$ aurait dit qu'il portait une perruque, tant les mèches garnissant son crâne élevé étaient plates et noires. »
Gustave FLAUBERT, *Bouvard et Pécuchet*, 1881.

Dans l'exemple [94], le pronom « on » fait donc partie de la même chaîne de coréférence que le « on » de l'exemple [93] alors que leurs référents, bien qu'indéfinis, ne sont pas les mêmes. En effet, le « on » de l'exemple [94] n'implique pas les personnages, il sert de tournure impersonnelle pour permettre une description. Ces deux « on » apparaissent sur la même page mais ils ne sont pas coréférents. Ils ont pourtant été annotés avec le même référent. Les « on » de cette chaîne n'ont donc pas vocation à être coréférents.

Dans le texte « De la ville au moulin », différents pronoms (plusieurs « on » ainsi qu'un « qui » et un « où ») ont aussi été annotés avec un « référent indéfini ». Ce label concerne principalement des cas de référence générique ou des cas dans lesquels le passif pourrait être utilisé comme dans l'exemple suivant :

Exemple [95]

« C'est grand-mère qui nous avait élevés tandis que nos parents travaillaient. À sa mort, trois ans plus tôt, j'étais déjà grande et forte et ma mère avait décidé que je resterais à la garde des jumeaux, et qu'[**on**]$_{\text{référent indefini}}$ m'adjoindrait une femme de ménage pour m'éviter les gros travaux. »
Marguerite AUDOUX, *De la ville au moulin*, 1926.

Dans l'exemple [95], « on m'adjoindrait » pourrait être remplacé par « il me serait adjoint ». L'usage du « on » rend la phrase légèrement moins impersonnelle : bien que le référent ne soit pas identifié, il doit s'agir d'un référent humain. Cependant, d'autres maillons sont identifiés avec ce label « référent indéfini », comme les deux « on » génériques coréférents dans l'exemple suivant :

Exemple [96]

« Il y a les chemins qu'[**on**]$_{\text{referent indefini}}$ ne voit pas mais dont [**on**]$_{\text{référent indefini}}$ devine le tracé capricieux au passage des charrettes. »
Marguerite AUDOUX, *De la ville au moulin*, 1926.

Le premier problème dans cet exemple [96] est que ces deux « on » coréfèrent entre eux mais ne coréfèrent pas aux autres maillons avec un « référent indéfini ». Il y a donc une nouvelle fois une grande chaîne pronominale qui n'est pas à proprement parler une chaîne de coréférence mais qui est pourtant annotée comme telle. Il serait donc intéressant de faire le tri dans ce type de « sur-chaînes » afin de s'assurer que les maillons des chaînes de coréférence soient bien tous coréférents. Le second problème avec ce label « référent indéfini » est qu'il regroupe des phénomènes différents : une forme proche du passif et une valeur générique, avec de la coréférence stricte comme dans l'exemple [96] ou floue comme dans l'exemple suivant :

Exemple [97]

« **Je** ne voulus pas attendre le réveil complet de la mignonne pour approcher le biberon de sa bouche. Elle le prit sans méfiance, mais à peine l'eût-elle pressé qu'elle le repoussa et renvoya en pluie toute la gorgée de lait. Il y eut dans ses yeux subitement ouverts un étonnement indigné, et aussitôt elle se mit à crier comme jamais elle ne l'avait fait encore.

Tout le jour elle cria et repoussa de ses petites mains l'horrible chose qu'[**on**]$_{\text{référent indefini}}$ voulait l'obliger à mettre dans sa bouche. Lasse et ennuyée, j'essayai de divers moyens pour la faire boire, mais tous furent inutiles. »
Marguerite AUDOUX, *De la ville au moulin*, 1926.

Dans l'exemple [97], le « on » peut référer à la narratrice mais l'utilisation de ce pronom et non du « je » ajoute une distance. On peut supposer que c'est pour cette raison qu'il a été annoté avec un « référent indéfini ».

Cette approche consiste donc à réaliser une seule chaîne pour les référents indéfinis. Elle a aussi été observée dans les blocs de texte « Nemoville » et « Ventre de Paris » qui possèdent chacun une longue chaîne « référent indéfini ». Malgré les justifications que l'on peut trouver pour l'annotation de ces mentions dont le référent est indéfini[7], le problème de cette annotation reste le même que pour l'approche précédente[8] : une seule « chaîne » regroupe des maillons coréférents et non coréférents.

7. Lorsque les notions de généricité et d'indéfinitude ne sont pas confondues.
8. Approche vue dans la section 5.2.1 qui regroupait les référents génériques

5.2.3 Regrouper tous les pronoms « on »

Les deux approches précédentes avaient pour point commun de regrouper les référents possédant un même aspect sémantique (comme la généricité ou l'indéfinitude) dans une chaîne alors que tous les maillons ne sont pas coréférents. L'approche suivante va encore plus loin en regroupant tous les pronoms « on » dans une même chaîne. Le premier bloc de texte concernant « Jean-Christophe » a été annoté avec une chaîne de 51 maillons nommée « ON »[9]. Les maillons de cette chaîne peuvent avoir un référent générique comme dans l'exemple suivant :

Exemple [98]
« – [**On**]$_{\text{ON}}$ ne doit pas céder aux enfants, quand ils pleurent. Il faut les laisser crier. »
Romain ROLLAND, *Jean-Christophe*, 1905.

Dans l'exemple [98], le pronom « on » est le premier de la chaîne. Il a une valeur générique car il prend place dans un énoncé qui a une tournure de proverbe, avec une valeur de vérité générale. D'autres maillons de cette même chaîne ont aussi une valeur générique ou indéfinie mais ne sont pas coréférents avec tous les autres. Le problème posé par cette approche est alors le même que dans les sections précédentes. De plus, certains maillons de la même chaîne peuvent avoir un référent bien identifié, qui n'est ni générique, ni indéfini, ni flou. C'est le cas dans l'exemple suivant :

Exemple [99]
« Plus le chemin était mauvais, plus Christophe le trouvait beau. [...] Parfois, [**on**]$_{\text{ON}}$ rencontrait sur la grande route un paysan dans sa carriole. Il connaissait grand-père. [**On**]$_{\text{ON}}$ montait auprès de lui. C'était le paradis sur terre. Le cheval filait vite, et Christophe riait de joie, à moins qu'[**on**]$_{\text{ON}}$ ne vînt à croiser d'autres promeneurs »
Romain ROLLAND, *Jean-Christophe*, 1905.

Dans l'exemple [99], les trois pronoms « on » réfèrent à un groupe qui semble être constitué du narrateur, de Christophe et du grand-père. Il s'agit en tout cas de compagnons de voyage et ces pronoms n'ont pas de valeur générique. Ces trois « on » méritent donc une chaîne de coréférence à part, plutôt que de faire partie de la même chaîne de coréférence que tous les autres « on » du texte.

Cette approche d'annotation regroupe les « on » et se base donc plutôt sur la forme

9. La même approche a été observée pour l'annotation du deuxième bloc de texte concernant « Jean-Christophe » de Romain Rolland ainsi que dans la biographie « Madame de Hautefort ». Dans ce dernier texte, il semble qu'il n'y ait pas de pronom « on » dont le référent puisse avoir une interprétation précise.

de l'expression plutôt que sur son sens. Or, tous les « on » ne sont pas nécessairement coréférents. La chaîne « ON » comporte 51 maillons, le texte comporte 49 « on » [10]. Un « nous » et deux « son » qui peuvent avoir une interprétation générique sont aussi annotés dans cette chaîne. Cette approche possède donc aussi une part de réflexion à propos de la généricité. En effet, la généricité peut être portée par différentes formes linguistiques.

Le pronom « on » peut être générique, indéfini, flou ou encore désigner un référent précis. Nous avons vu dans les sections précédentes que tous les « on » génériques ne sont pas nécessairement coréférents. La même remarque peut être observée pour les « on » indéfinis. Mettre tous les « on » dans une même chaîne de coréférence revient à prendre en compte uniquement la forme de l'expression linguistique. L'annotation manuelle de ce type d'information n'est donc pas nécessaire étant donné qu'il est possible de les extraire automatiquement. Il arrive que des expressions référentielles soient coréférentes lorsqu'elles ont la même forme. C'est généralement le cas des noms propres et cela peut arriver avec des groupes nominaux ou des pronoms. Cependant, dans un texte, plusieurs « il » possèdent souvent des référents différents par exemple. L'approche d'annotation observée dans cette section correspond donc à créer une « sur-chaîne » de « sur-chaînes ».

5.2.4 Regrouper les référents génériques selon la structure textuelle

Le bloc de texte de Democrat contenant les articles de Wikipédia regroupe des articles encyclopédiques sur les girafes, les singes et les zèbres. Pour l'annotation du pronom « on », des chaînes différentes ont été créées pour chaque article du bloc de texte : « ONGir », « ONSin » et « ONZeb ». Cependant, certains « on » non coréférents restent annotés dans une même chaîne comme dans les deux exemples suivants :

Exemple [100]
« La girafe vit en Afrique, dans la savane. [**On**]$_{\text{ONGir}}$ la trouve du Tchad jusqu'en Afrique du Sud. »
Article Wikipédia (« Girafe » - Section « Répartition géographique »)

Exemple [101]
« D'après l'union internationale pour la conservation de la nature (UICN), [**on**]$_{\text{ONGir}}$ comptait 155 000 girafes vivant dans la nature en 1985 contre 97 500 en 2015 »
Article Wikipédia (« Girafe » - Section « Conservation »)

10. L'un des « on » du bloc de texte a été oublié lors de l'annotation.

Les deux « on » des exemples [100] et [101] sont donc annotés comme étant des maillons d'une même chaîne alors qu'il n'y a pas de relation de coréférence entre eux, hormis leur valeur générique. Il est intéressant de noter que les maillons des trois chaînes (« ONGir », « ONSin » et « ONZeb ») ne sont pas tous des « on ». Dans la chaîne « ONGir », il y a un « son » :

Exemple [102]
« L'expression "peigner la girafe" signifie ne rien faire ou perdre [**son**]$_{\text{ONGir}}$ temps. »
Article Wikipédia (« Girafe »)

Le pronom possessif « son » a ici une valeur générique car il désigne n'importe quelle personne, comme les autres « on » de la chaîne dans laquelle il est annoté. Dans la chaîne « ONSin », il y a aussi un « nous » :

Exemple [103]
« La première description "scientifique" des singes qui [**nous**]$_{\text{ONSin}}$ soit parvenue date du IVe siècle av. J.-C. et revient au philosophe grec Aristote. »
Article Wikipédia (« Singe »)

Ce « nous » a lui aussi une valeur générique et ne désigne pas seulement l'auteur de l'article et ses collègues mais une communauté plus large d'initiés, dont les frontières sont floues. Dans la chaîne « ONZeb », il y a deux maillons, un « nos » et un « nous », qui possèdent également une valeur générique :

Exemple [104]
« De [**nos**]$_{\text{ONZeb}}$ jours, il est presque impossible de distinguer le crâne d'un zèbre de celui d'un cheval, mais [**nous**]$_{\text{ONZeb}}$ pouvons penser que les équidés qui colonisèrent les savanes tropicales devinrent des zèbres ».
Article Wikipédia (« Zèbre »)

C'est donc cette valeur générique qui semble lier tous ces pronoms à l'intérieur de ces trois chaînes. Mais cela suffit-il pour qu'ils soient coréférents ? À notre sens, cette vision est quelque peu réductrice. Un effort de partition des expressions possédant une valeur générique en fonction des articles a été opéré lors de l'annotation de ce bloc. Cela va dans le sens du respect de la structure textuelle pour l'analyse des chaînes de coréférence. En effet, il est normal qu'un « on » générique issu de la partie à propos des singes ne soit pas coréférent avec un « on » générique issu de la partie sur les girafes, surtout si ces trois articles ont été rédigés indépendemment. Cependant, ces « chaînes thématiques » ne sont pas des chaînes de coréférence car tous les maillons qu'elles contiennent ne sont

pas tous coréférents. Ce principe d'annotation pourrait paraître semblable aux « chaînes topicales » de la ressource ANNODIS (ANNOtation DIScursive) (Péry-Woodley et al. 2011). Ces chaînes comportent des segments [11] qui possèdent un même topique. Une chaîne topicale regroupe pour « la majorité » (Colléter et al. 2012) — et donc pas uniquement — des propositions à propos d'un même référent. L'annotation de ce type de chaîne dans ANNODIS consiste aussi à repérer les « indices » qui permettent leur identification. Ces indices pourraient correspondre aux maillons d'une « chaîne de référence », selon Federzoni, Ho-Dac et Rebeyrolle (2020). L'annotation des « on » du bloc de texte de Wikipedia dans Democrat ne correspond donc pas aux chaînes topicales d'ANNODIS.

Une approche liée à la structure textuelle (mais aussi au sens) est également observable dans le premier bloc de texte « Mademoiselle Fifi ». Un effort de séparation des référents du pronom « on » a été réalisé avec l'annotation de quatre chaînes de coréférence impliquant des « on » : « ON » (10 maillons), « OnTOUS » (4 maillons), « On » (2 maillons) et « OnGenRouille » (3 maillons).

La chaîne « ON » possède 10 maillons qui se répartissent de la page 1 à la page 10 du bloc de texte. En se fiant aux analyses réalisées précédemment sur ce corpus, autant de pronoms « on » répartis sur autant de pages devraient avoir une faible probabilité d'être effectivement coréférents. En effet, la première mention concerne une description : « La pluie tombait à flots, une pluie normande qu'**on** aurait dit jetée par une main furieuse ». Elle ne coréfère déjà pas avec la seconde mention qui se rapporte plutôt à des « on dit » : « et **on** le disait brave homme autant que brave officier ». Selon nous, tous les maillons de cette chaîne devraient être des singletons.

La chaîne « OnTOUS » possède des maillons sur les pages 6, 7 et 8. Ces pages sont inclues dans la plage de la chaîne « ON » (pages 1 à 10) mais cette chaîne de coréférence semble avoir été identifiée selon un critère sémantique. La chaîne de coréférence semble concerner les gens prenant part à une réception : « **on** entra dans la salle à manger » page 6 et « **On** s'assit. » page 7. Cependant, dans la phrase suivante : « **On** arrivait au dessert ; **on** versait du champagne. », les deux maillons ne coréfèrent pas entre eux ni avec les autres maillons de la chaîne. Le premier « on » peut impliquer le premier groupe de personne de manière floue mais il a plutôt vocation à décrire la situation temporelle de manière générale. Le second « on » désigne de manière indéfinie les gens qui versent le champagne, il ne s'agit probablement pas des invités qui prennent part à la réception.

La chaîne « On » possède deux maillons qui se trouvent sur la page 12 du bloc de texte. Ces deux maillons ne sont pourtant pas coréférents. Le premier « on » permet de réaliser une description, bien qu'il puisse impliquer des personnages et le narrateur de

11. Ces segments peuvent correspondre à des propositions.

manière floue : « tout au bout **on** apercevait des arbres ». Le second « on » a plutôt une valeur indéfinie et n'a pas de lien sémantique avec le premier : « La marche rapide du convoi disait bien pourtant qu'**on** enterrait ce défunt-là sans cérémonie, et, par conséquent, sans religion. ».

La chaîne « OnGenRouille » apparaît sur la page 17 du bloc, au début d'une section (ou d'un chapitre) intitulée « LA ROUILLE », comme le montre l'exemple suivant :

Exemple [105]
« LA ROUILLE
Il n'avait eu, toute sa vie, qu'une inapaisable passion : la chasse. Il chassait tous les jours, du matin au soir, avec un emportement furieux. Il chassait hiver comme été, au printemps comme à l'automne, au marais, quand les règlements interdisaient la plaine et les bois ; il chassait au tiré, à courre, au chien d'arrêt, au chien courant, à l'affût, au miroir, au furet. Il ne parlait que de chasse, rêvait chasse, répétait sans cesse : "Doit-**[on]**$_{\text{OnGenRouille}}$ être malheureux quand **[on]**$_{\text{OnGenRouille}}$ n'aime pas la chasse !"
Il avait maintenant cinquante ans sonnés, se portait bien, restait vert, bien que chauve, un peu gros, mais vigoureux ; et il portait tout le dessous de la moustache rasé pour bien découvrir les lèvres et garder libre le tour de la bouche, afin de pouvoir sonner du cor plus facilement.
[On]$_{\text{OnGenRouille}}$ ne le désignait dans la contrée que par son petit nom : M. Hector. ».
Guy de MAUPASSANT, *Mademoiselle Fifi*, 1882. Bloc 1

Le deux premiers « on » de l'exemple [105] sont effectivement coréférents (avec une valeur générique) mais ils ne le sont pas avec le troisième « on ». Ce dernier possède aussi une valeur générique. Cependant le référent n'est pas le même. Le nom choisi pour désigner le référent de cette chaîne comporte le nom donné à la section : « Rouille ».

L'annotation du bloc de texte de « Mademoiselle Fifi » [12] pour les pronoms « on » semble liée à l'ordre linéaire du texte et parfois même à sa structure textuelle (paragraphe et section). Cette approche semble logique car les « on » coréférents sont souvent proches linéairement. Deux « on » issus de chapitres, sections ou paragraphes différents auront moins de chances d'être coréférents [13]. Un effort de rapprochement sémantique a aussi été opéré avec la création de la chaîne concernant les invités. Cependant, il reste plusieurs maillons dans chaque chaîne qui ne sont pas coréférents.

Dans ces différents textes, ces approches ont le mérite de vouloir séparer les maillons

12. Les deux autres blocs de texte de « Mademoiselle Fifi » ont été annotés avec la même approche.
13. Toujours en se basant sur les analyses de ce corpus.

qui proviennent d'endroits différents dans le texte, supposant qu'ils ne puissent pas être coréférents. Cependant, les chaînes ainsi produites ne sont toujours pas des chaînes de coréférence stricte.

5.2.5 Regrouper les expressions génériques ou floues coréférentes

Contrairemet aux approches précédentes, certains annotateurs ont fait l'effort de séparer référents génériques qui n'étaient pas coréférents, tout en gardant le lien de coréférence entre les expressions effectivement coréférentes, comme dans l'exemple suivant :

Exemple [106]
« La trinité présente une immense carrière d'études philosophiques, soit qu'**[on]**$_{\text{on générique 4}}$ la considère dans les attributs de Dieu, soit qu'**[on]**$_{\text{on générique 4}}$ recherche les vestiges de ce dogme répandu dans le vieil orient »
François-René de CHATEAUBRIAND, *Génie du Christianisme*, 1802.

Dans l'exemple [106], le label donné au référent des deux pronoms « on » est « on générique 4 ». Ces deux « on » sont effectivement coréférents. L'annotateur ou l'annotatrice a identifié 8 chaînes de « on » génériques distinctes qui ne coréfèrent effectivement pas entre elles. Cette approche nous semble tout à fait correcte tant du point de vue technique que du point de vue théorique. Cette même démarche a été appliquée pour identifier 19 chaînes de « nous » génériques, 5 chaînes de « vous » génériques ainsi que 3 chaînes distinctes désignant l'homme de manière générique [14].

Pour le second bloc de texte annoté pour ce travail de thèse, la réflexion à propos du flou et de l'annotation commençait à émerger. La dérive d'annotation à propos des « on » génériques ou flous dans une même chaîne n'a pas été réitérée. Il y a une chaîne concernant deux « on » génériques qui sont bien coréférents :

Exemple [107]
« quand **[on]**$_{\text{“on” générique}}$ connaît bien le passé, disait-elle, **[on]**$_{\text{“on” générique}}$ dirige mieux le présent. »
Marguerite AUDOUX, *Douce Lumière*, 1937.

Il y a aussi plusieurs chaînes dont le référent est flou et qui impliquent des « on » ou des « nous » comme dans l'exemple suivant :

14. « un homme générique » (3 maillons), « homme générique 1 » (4 maillons) et « l'homme générique » (4 maillons).

Exemple [108]

« Il parla de la sapinière voisine où l'espace était grand, où il y avait un ruisseau dans lequel **[on]**$_{\text{"on" flou}}$ pouvait se baigner et un étang dans lequel **[on]**$_{\text{"on" flou}}$ pouvait pêcher. »
Marguerite AUDOUX, *Douce Lumière*, 1937.

Dans l'exemple [108], le référent est annoté comme étant « flou » car il s'agissait selon nous d'un flou situé entre le générique et les enfants. Le grand-père parle d'un lac dans lequel il est possible de se baigner et d'un ruisseau dans lequel il est possible de pêcher (1ère interprétation). Cependant, il en parle aux enfant dans le but qu'ils y aillent. Le rôle du grand-père faisant autorité sur les enfants, le verbe « pouvoir » peut être interprété comme une autorisation et pas seulement comme une capacité. Il peut donc aussi s'agir d'un lac et d'un ruisseau dans lequel les enfants peuvent respectivement se baigner et pêcher (2ème interprétation). Les deux interprétations n'étant pas nécessairement exclusives, le terme « flou » a été utilisé pour désigner le référent. Cependant, nous avons estimé que ces deux mentions désignaient le même référent de manière stricte, bien que ce dernier soit flou.

Dans le bloc de texte « Pauline » et le deuxième bloc de l'Est Républicain, annotés ensuite pour ce travail de thèse, les « on » (génériques ou non) ont aussi été correctement annotés dans des chaînes en fonction de leurs relation de coréférence.

Cette approche consiste donc à séparer les chaînes qui ne sont pas coréférentes. Cela revient donc, lorsque cela est nécessaire, à créer plusieurs chaînes de coréférence dont le référent est générique par exemple.

5.2.6 Ne pas regrouper (ou annoter) les expressions dont le référent est générique ou indéfini

Le premier bloc de texte concernant le code civil ne possède aucune chaîne de coréférence identifiée comme générique ou dont le référent est indéfini. Pourtant, le texte comprend des occurrences d'expressions dont le référent est générique, comme dans l'exemple suivant :

Exemple [109]

« La preuve de l'intention résultera d'une déclaration expresse, faite tant à la municipalité du lieu que l'**on**]$_{M826}$ quittera, qu'à celle du lieu où [**on**]$_{M829}$ aura transféré son domicile. »
Code civil - Bloc 1.

Dans l'exemple [109], les deux pronoms « on » sont traités comme des singletons. Nous considérons pourtant qu'ils entretiennent un lien de coréférence. Cette relation de coréférence repose sur la construction syntaxique de la phrase (verbe + tant à ... + qu'à ...) et sur la relation sémantique entre ces deux segments : la personne (générique) qui quitte une municipalité est la même qui transfère son domicile dans une autre municipalité.

Le second bloc de texte concernant le code civil ne possède pas non plus de chaîne identifiée comme correspondant à un référent générique, indéfini ou flou. Le texte ne contient que 3 « on » qui possèdent tous une valeur générique ou indéfinie mais qui ne sont pas coréférents.

Dans le bloc concernant la « Convention pour la protection du milieu marin de l'Atlantique du nord-est » il n'y a pas non plus de chaîne identifiée comme correspondant à un référent générique, indéfini ou flou. De plus, aucun des « on » n'est même annoté comme étant un singleton. Ils ne correspondent donc pas à des mentions et ne sont donc pas considérés comme des expressions référentielles. C'est aussi le cas des blocs de texte de la « Convention relative au renforcement de la Commission interaméricaine du thon tropical établie par la convention de 1949 entre les États-Unis d'Amérique et la République du Costa Rica ("convention d'Antigua") » et de la « Convention pour l'unification de certaines règles relatives au transport aérien international (convention de Montréal) ».

Il pourrait sembler que la démarche de considérer les expressions dont le référent est générique ou indéfini comme des singletons (ou même de ne pas les annoter comme des mentions) soit réservée aux textes non narratifs. Cependant, l'annotation du bloc de texte correspondant à la biographie « Elisabeth Seton » présente une approche similaire. Il ne semble pas y avoir de chaîne dont le référent soit générique, indéfini ou même flou. Les pronoms « on » ne sont parfois pas annotés comme des mentions. Lorsqu'ils le sont, il sont essentiellement des singletons. Pourtant certaines mentions sont selon nous coréférentes comme dans l'exemple suivant :

Exemple [110]

« Si j'ai quelque chose à demander, c'est par là qu'il faut que j'appelle ; alors paraît la sentinelle, armée de pied en cap, qui se promène avec un long fusil ; et tout cela, parce qu'**[on]**$_{M53}$ veut se préserver de la terrible contagion qu'**[on]**$_{M56}$ suppose que nous avons apportée de New-York… »
Laure CONAN, *Elisabeth Seton*, 1903.

Dans l'exemple [110], les deux « on » désignent un groupe de personnes qui n'est pas identifié précisément : il peut s'agir les habitants du lieu, des personnes qui le gouvernent, etc. Bien que le référent de ce groupe soit difficile à identifier, la structure de la phrase suggère que le référent des deux « on » est le même. Ces gens veulent se préserver de la contagion qu'ils pensent que les autres ont ramenée de New-York.

Cette approche ignore donc complètement la question des référents génériques, indéfinis ou flous et de la (co)référence du pronom « on ».

5.2.7 Identifier systématiquement un référent

Il semble que l'un des deux blocs de texte de Democrat comprenant la collection d'articles issus de l'Est Républicain ne possède pas de chaîne dédiée au flou, au générique ou au pronom « on ». Un effort d'annotation en coréférence stricte avec une identification systématique du référent a été opéré pour les occurrences de ce pronom. Une chaîne a par exemple été créée avec « la rédaction » comme référent. Ses maillons sont des pronoms comme « on », « nous » et « vous » (discours direct). Cependant, le référent de ces mentions n'est pas toujours uniquement la rédaction du journal, comme dans l'exemple suivant :

Exemple [111]

« Face à une formation gravelinoise dont **[on]**$_{\text{la rédaction}}$ connaît la force de frappe, les Nancéiens ont fait preuve d'un cœur gros comme ça. »
L'Est Républicain - Bloc 1

Dans l'exemple [111], le pronom « on » peut désigner la rédaction du journal mais il peut aussi avoir une valeur générique (la force de frappe gravelinoise, bien connue de tous). Par ailleurs, l'annotation de tous les « on » censés désigner la rédaction dans une seule et même chaîne peut amener des interrogations par rapport à la notion de chaîne de coréférence en raison des différents articles et auteurs.

L'annotation du bloc de texte « Rosalie de Constant » possède une chaîne de 56 maillons désignant le « narrateur ». Dans la plupart des cas, les maillons sont coréférents et désignent effectivement uniquement le narrateur comme dans la première phrase :

« [**Nous**]$_{\text{narrateur}}$ aurions pu commencer, selon l'usage, cette biographie trois ou quatre cents ans avant la naissance de notre héroïne ». Cependant, quelques maillons ont aussi une valeur générique, comme dans l'exemple suivant :

Exemple [112]
« [**On**]$_{\text{narrateur}}$ pourrait [**s**]$_{\text{narrateur}}$'y tromper à la douceur voulue de ce style, mais Rosalie ne trouva jamais dans sa seconde mère les soins, l'indulgence que son cœur exclusif et sensitif eût désirés. S'il n'y eut pas guerre ouverte, ce fut plutôt une négation de rapports ; [**on**]$_{\text{narrateur}}$ [**se**]$_{\text{narrateur}}$ demandera souvent en lisant dans la suite la correspondance du père et des filles quel rôle jouait cette nouvelle mère, et [**on**]$_{\text{narrateur}}$ en viendra à conclure qu'elle jouait pour le moins un rôle fort effacé. »
Lucie ACHARD, *Rosalie de Constant, sa famille et ses amis*, 1901.

Dans l'exemple [112], les deux premières mentions sont coréférentes entre elles et les trois autres aussi [15]. Selon nous, il devrait donc y avoir deux chaînes de coréférence pour tous ces maillons car les deux premiers ne coréfèrent pas aux trois autres. Les deux premiers maillons ont une valeur générique alors que les trois derniers pourraient bien désigner le narrateur, bien que l'usage du pronom « on » par énallage permette de prendre une certaine distance avec le référent et implique donc potentiellement un lien de coréférence floue avec la chaîne désignant le narrateur.

Vouloir identifier à tout prix un référent est risqué. Cela peut entraîner un regroupement de maillons qui ne sont pas tout à fait coréférents, ou au contraire une séparation de chaînes qui pourraient être coréférentes.

Toutes ces démarches d'annotation observées dans les 30 textes de notre sous-corpus de Democrat peuvent posséder des points communs mais aussi diverger complètement. Dans les deux cas, il s'agit d'un manque d'homogénéité et de régularité qui n'est pas recherché pour un corpus annoté qui servira de base d'entraînement à des outils automatiques ou à des analyses linguistiques.

15. Bien que le manuel ne recommande pas d'annoter les pronoms clitiques réflexifs. Selon nous, cette information est pourtant intéressante à annoter, notamment pour aider les systèmes de détection automatique dans leur apprentissage. En effet, le « se » dans cet exemple a une valeur réflexive : se demander à soi-même. Cependant, avec le pronom « on », le clitique « se » peut aussi être interprété de manière non réflexive : se demander à chacun (« On se demande nos numéros. »).

5.2.8 Remarques sur la notion de chaîne de coréférence et la structure textuelle

Peut-il y avoir coréférence entre des textes concaténés ?

Dans un bloc de texte constitué d'une concaténation de petits articles de presse, est-il juste de considérer que des mentions peuvent être coréférentes alors qu'elles ne sont pas issues du même article et qu'elles ne sont par conséquent pas nécessairement écrites par le même auteur ?

Il est parfois question des mêmes entités dans des articles différents : le référent est le même, les expressions référentielles peuvent donc être considérées donc coréférentes. Est-ce qu'il peut y avoir des relation de coréférence entre des expressions issues de textes différents ?

Est-ce que le seul critère pour que des expressions référentielles fassent partie d'une même chaîne de coréférence est leur relation de coréférence ? Comme le soulignent les figures [1.5] et [1.7] par exemple, les longues chaînes de coréférence sont rares. Une chaîne désignant le narrateur ou un personnage principal peut s'étaler sur la longueur d'un roman, mais n'y a-t-il pas parfois des points de rupture qui nécessiteraient de découper ces chaînes en sous-chaînes ? Guillot-Barbance et Quignard (2019) ont étudié les chaînes dans leur rapport avec la structure textuelle dans les *Essais sur la peinture* de Diderot. Ils ont remarqué que les chaînes étaient dans ce cadre précis essentiellement « cantonnées à une dimension très locale » comme le paragraphe. Ils concluent aussi que leur étude « donne une illustration de la manière dont les chaînes de référence [...] créent un maillage local qui n'est ni totalement dépendant ni totalement indépendant de la structure textuelle ».

Dans l'autre bloc de Democrat contenant des articles de l'Est Républicain, celui annoté au cours de ce travail de thèse, plusieurs articles évoquent les mêmes incidents, comme dans l'exemple suivant :

Exemple [113]
« [article] Polochon a disparu
Polochon a pris la fuite après l'explosion.
Effrayé par l'explosion, Polochon, un petit chat appartenant à [**Roxane Herbrecht**]$_r$ a disparu. [**Sa propriétaire**]$_r$ le recherche activement. Si vous apercevez dans le secteur de la rue du Tramway, le chat au pelage de couleur noire et pourvu d'un collier avec une petite plaque, prière de contacter [**la jeune femme**]$_r$ »
L'Est Républicain - Bloc 2

L'exemple [113] contient un article issu du même bloc que l'extrait [76] qui relatait l'explosion dans un article précédent : « **Roxane Herbrecht** et Nicolas Burcey : "Une **explosion** nous a réveillés. Le plancher et notre lit se sont soulevés." ». Les deux articles évoquant une explosion et la présence du nom propre à deux reprises nous ont laissé supposer qu'il s'agissait de la même personne. Au moment de l'annotation, les mentions désignant Roxane Herbrecht ont été annotées avec le même référent dans les deux articles pour éviter de perdre cette information, alors qu'elles correspondraient probablement plutôt à des sous-chaînes pour un même référent. Il existe d'autres exemples de ce type dans ce bloc de texte de l'Est Républicain. Le fait de garder un lien de coréférence entre les différents articles permet aussi de pouvoir suivre le devenir discursif du référent en question.

La question se pose aussi pour le bloc de texte contenant les articles de Wikipédia qui sont eux aussi une concaténation de plusieurs articles. Dans la section 5.2.4, la démarche consistant à séparer les chaînes de « on » génériques en fonction des articles était louable[16] car la coréférence de ce type de pronom est peu probable à travers deux textes différents. Pourtant, comme nous l'avons vu pour les articles de presse, cela reste envisageable pour les noms propres. Pour les groupes nominaux génériques la question se pose aussi : si le groupe nominal « les singes » est présent dans un articles sur les girafes, peut-il coréférer avec les expressions désignant les singes en général dans un autre article ?

Quelle(s) limite(s) pour la coréférence des expressions à valeur générique ?

Dans le texte « Capitaine Fracasse », plusieurs « vous » et un « votre », dont la valeur est générique, sont tous[17] annotés avec le référent « vous ». Certains « vous » sont coréférents de manière certaine comme dans l'exemple suivant :

16. Bien que les maillons ne soient pas tous coréférents au sein de chaque chaîne.
17. Huit au total.

Exemple [114]

« Les ronces, aux ergots épineux, se croisaient d'un bord à l'autre des sentiers et [**vous**]$_{vous}$ accrochaient au passage pour [**vous**]$_{vous}$ empêcher d'aller plus loin et [**vous**]$_{vous}$ dérober ce mystère de tristesse et de désolation. »
Théophile GAUTIER, *Le Capitaine Fracasse*, 1863.

Dans l'exemple [114], les trois « vous » ont une valeur générique et sont coréférents. Ils forment une chaîne de coréférence stricte, notamment en raison de leur proximité. Mais lorsqu'il s'agit de généricité, quelle est la distance maximale acceptable pour le maintien de la coréférence ? Est-ce d'ailleurs le seul critère ? Deux pages plus loin, la description continue dans l'habitation après être passée par un cabinet, une grotte ou encore une écurie :

Exemple [115]

« Au-dessus de la cheminée de forme antique, un massacre de cerf dix cors épanouissait son bois, et le long des murailles grimaçaient sur les toiles rembrunies des portraits enfumés représentant des capitaines cuirassés ayant leur casque à côté d'eux ou tenu par un page, et fixant sur [**vous**]$_{vous}$ des yeux profondément noirs seuls vivants dans leurs figures mortes »
Théophile GAUTIER, *Le Capitaine Fracasse*, 1863.

Étant donnée la prolongation de la description sur plusieurs pages, il est compréhensible d'annoter ce « vous » dans la même chaîne que les précédents. Les « vous » et le « votre » suivants sont dans le même cas : ils prennent place dans la description du lieu et il n'y a pas de rupture dans le discours comme dans la structure textuelle, comme un changement de chapitre par exemple. Avec ce type de rupture, la coréférence ne serait probablement pas maintenue avec un référent aussi vague.

5.2.9 Classification

Lors de l'annotation des chaînes de coréférence, la coréférence floue peut donc engendrer différentes conduites d'annotation que nous avons abordées dans les sections précédentes. Le tableau suivant regroupe les textes de notre sous-corpus de Democrat et récapitule l'approche suivie par chaque annotateur ou annotatrice. Les différentes approches correspondent aux numéros des titres des sections dans lesquelles elles sont présentées, que nous rappelons ici :

1. Regrouper les référents génériques ;
2. Regrouper les référents indéfinis ;
3. Regrouper tous les pronoms « on » ;

4. Regrouper les référents génériques selon la structure textuelle ;
5. Regrouper les expressions génériques ou floues coréférentes ;
6. Ne pas regrouper (ou annoter) les expressions dont le référent est générique ou indéfini ;
7. Identifier systématiquement un référent ;

Titre	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Aden Arabie	✓	-	-	-	-	-	-
Articles Wiki	-	-	-	✓	-	-	-
Bouvard et Pécuchet	✓	✓	-	-	-	-	-
Code Civil 1	-	-	-	-	-	✓	-
Code Civil 2	-	-	-	-	-	✓	-
Code de procédure pénale	-	-	-	-	-	-	-
Convention univ	-	-	-	-	-	-	-
Convention marin	-	-	-	-	-	✓	-
Convention aéro	-	-	-	-	-	-	-
Convention thon	-	-	-	-	-	✓	-
Douce Lumière	-	-	-	-	✓	-	-
De la ville au moulin	-	✓	-	-	-	-	-
Élisabeth Seton	-	-	-	-	-	✓	-
Est Républicain 1	-	-	-	-	-	-	✓
Est Républicain 2	-	-	-	-	✓	-	-
Génie du christianisme	-	-	-	-	✓	-	-
Jean-Christophe 1	-	-	✓	-	-	-	-
Jean-Christophe 2	-	-	✓	-	-	-	-
Madame de Hautefort	-	-	✓	-	-	-	-
Mademoiselle Fifi 1	-	-	-	✓	-	-	-
Mademoiselle Fifi 2	-	-	-	✓	-	-	-
Mademoiselle Fifi 3	-	-	-	✓	-	-	-
La morte amoureuse	✓	-	-	-	-	-	-
Le capitaine Fracasse	✓	-	-	-	-	-	-
Le Diable au corps	✓	-	-	-	-	-	-
Le ventre de Paris	-	✓	-	-	-	-	-
Nemoville	-	✓	-	-	-	-	-
Pauline	-	-	-	-	✓	-	-
Rosalie de Constant	-	-	-	-	-	-	✓
Sarrasine	✓	-	-	-	-	-	-
TOTAL	**6**	**4**	**3**	**4**	**4**	**5**	**2**

Tableau 5.1 – Sous-corpus de Democrat : les différentes conduites d'annotation.

Les blocs de texte correspondant au « Code de procédure pénale » et à la « Convention univ » ne contiennent pas de pronoms « on » et ne semblent pas contenir d'expression dont le référent est générique, indéfini ou flou. Le bloc « Convention aéro » semble être dans le même cas, il ne possède qu'un pronom « on » qui est un singleton, ce qui semble logique. Le bloc « Bouvard et Pécuchet » regroupe quant à lui deux approches d'annotation qui ne sont pas exclusives.

Certaines des approches regroupées dans le tableau [5.1] pourraient être qualifiées de dérives d'annotation. Selon nous, l'approche adaptée est l'approche 5 : celle qui consiste à bien séparer les mentions qui ne sont pas coréférentes sans pour autant chercher à identifier un référent à tout prix, même lorsque la relation est floue. Comme cela a été le cas dans l'annotation du texte de Chateaubriand notamment (section 5.2.5). Nous avons tiré une classification de ces « dérives », elles peuvent être regroupées en deux catégories :

1. **Un sur-regroupement de maillons** : l'annotation de tous les référents génériques et/ou flous dans une unique chaîne de coréférence bien que tous ces maillons ne soient pas tous coréférents (conduites d'annotation 1, 2, 3, 4 et parfois 7). Cela revient souvent aussi à annoter ensemble des mentions dont le lien de coréférence est flou.
2. **Un sous-regroupement de maillons** : l'annotation, dans différentes chaînes, d'expressions référentielles sémantiquement reliées car elles ont un lien de coréférence floue (conduites d'annotation 6 et parfois 7).

Ces variations dans l'annotation de la coréférence non stricte et floue concernent principalement des pronoms, le pronom « on » revenant fréquemment car il est un marqueur linguistique qui véhicule particulièrement bien le flou référentiel.

Avec le sur-regroupement de maillons, il est impossible de détecter automatiquement les sous-chaînes effectivement coréférentes. Le sous-regroupement de maillons entraîne le problème inverse : l'annotation de ces expressions dans des chaînes différentes est une perte d'information sémantique. Cette perte est problématique lorsque l'on tente de représenter la structure discursive et informationnelle d'un texte.

Même dans un cadre strict comme le projet Democrat, il peut y avoir des variations dans l'annotation de la coréférence floue entre plusieurs annotateurs différents mais aussi pour un même annotateur. En suivant le manuel d'annotation de Democrat, les annotateurs devraient surtout effectuer des sous-regroupements de maillons pour respecter le critère de coréférence stricte. Or, nous avons relevé une majorité de cas de sur-regroupements (tableau [5.1]), ce que ne recommande pas le manuel. C'est pourquoi il est nécessaire de trouver un cadre précis et dans lequel il est permis d'indiquer lorsque le lien de coréférence est flou.

Il est possible que certains annotateurs envisagent la généricité comme étant un référent unique et que d'autres envisagent que la coréférence est impossible entre deux expressions à valeur générique. Nous pensons néanmoins que la raison de certaines de ces démarches d'annotation est plutôt liée à la méthode d'annotation du corpus. L'annotation par expressions référentielles a peut-être induit les annotateurs en erreur. Apposer une étiquette sur une expression est une opération mentale différente que celle qui correspond à relier des expressions entre elles pour en faire une chaîne. Il est possible que les annotateurs qui ont apposé le label « référent générique » sur des expressions, n'avaient plus conscience au moment de l'annotation que cela allait créer une unique chaîne de coréférence. De plus, les erreurs d'annotation manuelles sont inévitables. Au cours d'une étude sur l'annotation des expressions temporelles, Setzer et Gaizauskas (2001) avaient relevé différentes causes d'erreur d'annotation : un manque de compréhension du manuel d'annotation ou encore la fatigue qui peut être liée à la monotonie de la tâche d'annotation ou à l'annotateur lui-même.

La confusion qui a généré des conduites d'annotation impliquant des sur-regroupements de maillons provient aussi probablement de la consigne recommandant de faire la distinction entre les différents référents du pronom « on » :

> « 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". »

Le manuel du projet Democrat précise néanmoins que seuls les liens de coréférence stricte sont pris en compte, la distinction des référents de « on » ne dispense pas du critère de coréférence pour la création des chaînes. La consigne n'était présente que pour le pronom « on », or les exemples issus des différentes approches nous ont montré que la question touche les référents génériques, indéfinis ou flous et peuvent être portés par différents pronoms (« on », « vous », « nous », etc.) ainsi que des groupes nominaux. Il est plus aisé de regrouper des chaînes contenant le même mot-clé que de diviser des chaînes non coréférentes car cela demanderait un nouvel effort d'annotation manuelle. C'est pourquoi une consigne claire à ce sujet est nécessaire.

Même des annotateurs experts comme ceux du projet Democrat n'ont pas trouvé de consensus tacite pour l'annotation du flou. Pour uniformiser l'annotation de ce phénomène, il est nécessaire de fournir un cadre plus précis au niveau du manuel d'annotation. Si le projet ne prend pas en compte la coréférence floue, le cadre doit aussi préciser plus clairement qu'il faut séparer les chaînes de coréférence qui semblent proches mais qui ne possèdent pas strictement et assurément le même référent. Si la coréférence floue est prise en compte, le manuel doit indiquer à quel niveau renseigner la nature de cette relation : via le label du référent ou via les relations entre les mentions par exemple (ou les deux).

Que l'on prenne en compte la coréférence floue ou non, il est important de se poser la question des référents génériques. La généricité est souvent portée par le pronom « on » mais elle peut l'être aussi plus rarement par le biais d'autres pronoms comme « nous », « tu » ou encore « vous » [18] par exemple. Annoter tous les pronoms génériques dans une même chaîne de coréférence pourrait avoir un intérêt seulement dans le cas d'une hiérarchisation de cette chaîne en sous-chaînes coréférentes. Cependant, comme le démontrent les exemples [85] et [86], tous les « on » génériques d'un texte ne coréfèrent pas nécessairement (alors que c'est le cas pour certains). C'est pourquoi la solution à privilégier serait de créer des sous-chaînes génériques, qui sont en réalité des chaînes de coréférence stricte, à la manière de l'annotation du texte de Chateaubriand abordé dans la section 5.2.5. Pour les « on » de l'exemple [85], le référent pourrait être « On-generique-fuseaux » afin que le nom du référent soit explicite et singularisant. Pour l'exemple [86], le référent pourrait être « On-generique-voyageurs ».

Les articles non narratifs comme les articles encyclopédiques et les articles de presse ont pour objectif de rapporter des faits. Par leur nature, ils laissent moins de place au flou que les romans dans lesquels un groupe peut être désigné de manière floue plus facilement. Nous avons néanmoins relevé des cas dans lesquels l'annotation en coréférence stricte pose des difficultés dans chaque genre textuel. Cette étude reste qualitative et pourrait aussi mériter d'être menée à une plus grande échelle pour pouvoir en tirer des conclusions générales. Face à ces difficultés pour l'obtention d'annotations uniformes du phénomène de la coréférence floue dans un cadre strict, nous proposons de la prendre en compte directement dans le schéma d'annotation.

Malgré les tentatives d'autres projets pour annoter la coréférence non stricte, le flou des relations de coréférence en français n'a jamais été pris en compte en corpus alors qu'il génère des conduites d'annotation hétérogènes. Nous avons pu identifier des situations propices à la coréférence floue. Nous les avons ensuite caractérisées pour retenir que la coréférence floue s'exprime fréquemment à travers la référence à des groupes et l'utilisation de certains pronoms (le pronom « on » en particulier). L'étude de ces cas particuliers

18. Nous avons aussi vu l'exemple du pronom possessif « son » dans cette section.

dans des textes en français du corpus Democrat et de leur annotation au sein du projet nous a permis de relever différentes approches d'annotation. Certaines annotations correspondent à une volonté de regrouper les référents indéfinis, d'autres à regrouper les référents génériques, en les confondant parfois. L'annotation d'un bloc de texte regroupe tous les pronoms « on ». D'autres annotations effectuent un regroupement des référents génériques en fonction de la structure textuelle. Certaines annotations ne regroupent des expressions génériques ou floues que lorsqu'elles sont effectivement coréférentes. Des annotations ont aussi pris le parti de ne pas regrouper les expressions génériques ou indéfinies coréférentes, les considérant parfois comme étant non référentielles. Pour finir, certaines annotations ont tenté d'identifier un référent précis systématiquement. Certaines de ces approches pourraient être qualifiées de « dérives d'annotation ». En effet, le manuel d'annotation recommande l'annotation des relations de coréférence stricte et ces dérives ne respectent pas ce critère. Le choix de l'annotation en coréférence stricte est pourtant compréhensible car il paraît simple d'annoter uniquement les relations sur lesquelles il n'y a aucun doute. Cependant, l'annotation du pronom « on » en particulier a généré des comportements de sur-regroupements (fréquents) et de sous-regroupements (moins fréquents) de maillons. L'annotation en coréférence stricte n'est finalement pas si évidente et un cadre plus précis et permettant d'indiquer les relations de coréférence floue serait selon nous nécessaire. De plus, les chaînes de coréférence représentent l'évolution référentielle d'une entité au fil d'un texte et il est regrettable de perdre des informations sur cette entité en mettant la coréférence floue de côté.

La prise en compte de la coréférence floue aurait pu être envisageable de manière simple dans le corpus Democrat. Ce corpus est annoté en coréférences au niveau des unités (selon le schéma URS). Si les chaînes étaient réellement coréférentes de manière stricte (avec plutôt des sous-regroupements de maillons que des sur-regroupements de maillons), il aurait suffi de relier entre elles les chaînes ou les singletons dont la coréférence est floue. Ce lien pourrait s'effectuer au niveau des relations ou des schémas (URS) en indiquant que la relation de coréférence est stricte ou floue.

Chapitre 6

Des recommandations pour le traitement de la coréférence non stricte en corpus

Lors du chapitre précédent, nous avons vu que le cadre strict du projet Democrat n'a pas suffi pour obtenir des annotations homogènes en coréférences, notamment lorsque des phénomènes de généricité et d'indéfinitude étaient présents. Les annotateurs confondent parfois ces concepts entre eux mais aussi avec des référents plus précis. De plus, ces phénomènes génèrent souvent un flou coréférentiel. Certains référents génériques sont quelques fois simplement génériques, mais ils peuvent aussi impliquer d'autres référents de manière floue. C'est souvent le cas du narrateur par exemple. L'hétérogénéité de l'annotation de ces relations de référence et de coréférence nous pousse à envisager des recommandations plus précises pour l'annotation ainsi que la prise en compte de la coréférence floue.

6.1 Des recommandations pour l'annotation

La chaîne de traitement habituelle en linguistique de corpus commence par la description du phénomène, puis la création d'un corpus et la rédaction d'un manuel d'annotation qui sera appliqué à travers un schéma d'annotation. Ensuite, vient l'utilisation d'outils pour l'annotation puis l'exploration des données annotées. Dans leur très grande majorité, ces outils incitent à trancher. Or, trancher est justement incompatible avec la notion de flou. La dernière étape de cette chaîne de traitement est l'utilisation d'outils de traitement automatique des langues pour identifier le phénomène de manière automatique. La catégorisation qui en résulte est encore une fois stricte. Pourquoi ne pas construire une chaîne de coréférence même lorsque le référent est vague ou indéterminé ?

6.1.1 Objectifs

Envisager l'appartenance à une catégorie de manière non binaire n'est pas nouveau. Blasco-Dulbecco (1999) distinguait des degrés de coréférence, notamment en lien avec la valeur de généricité que peut porter un pronom. En effet, selon elle, le pronom clitique n'est pas toujours complètement anaphorique, comme lorsqu'il a une valeur générique[1] ou une divergence d'accord pour un élément disloqué[2].

Dans le cas des « références méréologiques »[3] (*mereological references*), Poesio, Sturt et al. (2006) s'inspirent notamment des travaux de Ferreira, Bailey et Ferraro (2002)[4] pour proposer une hypothèse (*Justified Slopiness Hypothesis*) selon laquelle une expression anaphorique ambiguë n'est pas toujours perçue comme étant problématique. Cette hypothèse est aussi détaillée dans Poesio, Reyle et Stevenson (2008). À la suite de ces travaux, Versley (2008) en propose une variante (*Generalized Slopiness Hypothesis*) en s'inspirant aussi de Smith et Brogaard (2000) et de leur considération pour la référence vague.

La théorie des ensembles flous[5] (Zadeh 1965) représente l'idée d'appartenance partielle à une classe. Cette théorie considère que les catégories peuvent avoir des limites qui ne sont pas bien définies et qu'il existe une gradualité dans le passage d'une situation à une autre. Le point intéressant de cette théorie est la fonction d'appartenance. Il n'y a plus seulement deux états : celui d'appartenir ou non à une classe mais un degré d'appartenance à cette classe. Cette théorie est donc une généralisation de la théorie mathématique des ensembles mais en admettant des situations intermédiaires. Cette théorie des ensembles flous est un outil mathématique qui permet de gérer l'imprécision et l'incertitude dont on pourrait se servir pour modéliser la coréférence floue. Cependant, cela nécessiterait une modélisation permettant d'estimer le degré de coréférence de chaque maillon à une chaîne, ce qui est rarement possible. Une solution plus simple serait préférable pour réduire la complexité de la tâche d'annotation.

Landragin (2007) propose plusieurs solutions pour la gestion de l'annotation des anaphores à antécédent flou, potentiellement applicables à la coréférence floue en général. Il propose tout d'abord de nommer un antécédent flou, qui puisse servir de référent. Il propose ensuite trois principes : « le principe des alternatives », « le principe des possibles » et « le principe de double balisage ». Le principe des alternatives consiste à

1. Voir section 4.1.1 sur le générique.
2. Voir section 3.1.2 sur les disloquées.
3. Il s'agit d'une classe d'expressions anaphoriques qui peut regrouper deux référents pour en créer un troisième. Poesio, Sturt et al. (2006) donnent l'exemple de l'entité « train » (désignée par un pronom) qui englobe une locomotive et un waggon.
4. Voir section 4.3.2 sur la théorie *good-enough*.
5. Théorie évoquée rapidement dans la section 4.1.1.

regrouper les deux candidats potentiels pour former un antécédent. Dans ce groupe formé par cet antécédent figure aussi l'antécédent correspondant à l'idée de l'auteur sans que l'annotateur puisse l'identifier strictement avec certitude car ils se recoupent. Le principe des possibles correspond à regrouper l'ensemble des candidats possibles comme antécédent, alors qu'ils ne sont pas tous plausibles contrairement au principe des alternatives. Le principe de double balisage concerne plutôt les antécédents qui pourraient correspondre à des empans de texte beaucoup plus longs, comme un paragraphe, correspondants à des entités de discours.

Ces diverses propositions ainsi que la notion de *near-identity* de Recasens (2010) représentent une volonté de prise en compte de la complexité de la représentation de la coréférence pour un humain au sein d'une opération de catégorisation (comme la tâche d'annotation). Étant donnée cette complexité de représentation (universelle) de la coréférence, il semble indispensable de donner un cadre très précis aux annotateurs devant annoter un corpus en coréférences. Ce cadre doit notamment prendre en compte les différents phénomènes linguistiques abordés dans le chapitre précédent. Il doit aussi établir des définitions claires de ces phénomènes et en donner des exemples. En effet, l'annotateur doit pouvoir se référer à des consignes précises pour les cas complexes de ce type. Le manuel d'annotation doit donc distinguer les notions de généricité, d'indéfinitude et de flou pour ensuite en donner des exemples assortis de consignes d'annotation. Ces consignes suivent un schéma d'annotation qui doit aussi proposer des solutions d'annotation de ces phénomènes linguistiques complexes.

La réalisation d'un schéma d'annotation implique de définir les phénomènes qu'il va pouvoir reconnaître. Le schéma que nous proposons permet l'identification de la coréférence floue. Concrètement, cela demande à l'annotateur de prêter attention aux cas particuliers vecteurs de flou que nous avons identifiés dans la section 4.2, comme la référence à des groupes ou encore certains pronoms (« on » notamment). Une attention particulière doit aussi être apportée vis-à-vis de la référence générique qui peut impliquer des entités de manière floue.

6.1.2 Proposition 1 : des précisions dans le manuel d'annotation

Pour prévenir les dérives d'annotations relevées dans le chapitre précédent, le manuel d'annotation devrait contenir les précisions suivantes :

1. Une distinction est nécessaire entre les notions de généricité, d'indéfinitude et de flou :
   — **Généricité** : un référent générique désigne une catégorie générale de référents qui peut référer à « tout le monde » avec des restrictions éventuelles liées à

des conditions spécifiques. On retrouve ce type de référent en particulier dans les proverbes : « Quand on veut on peut ».
- **Indéfinitude** : un référent indéfini est un référent qui n'est pas identifié. Par exemple : « On m'a volé mon vélo. ».
- **Flou** : un référent flou est difficilement identifiable mais il peut être impliqué dans des relations de coréférence stricte. En revanche, la coréférence floue peut impliquer des référents flous et stricts mais c'est la relation entre les référents qui est floue. La généricité et le pronom « on » peuvent être vecteurs de flou coréférentiel, notamment lorsque le narrateur s'inclut de manière floue dans une phrase générique : « Quand on est vieille comme moi, on se fatigue plus vite ».

2. Ces trois concepts n'impliquent pas nécessairement des relations de coréférence entre tous les référents qui possèdent cette valeur. Par exemple, certains « on » génériques peuvent être coréférents, d'autres non. Il est donc nécessaire de distinguer chaque référent au moment de l'annotation en donnant des labels permettant de les identifier (ne serait-ce qu'au moyen de numéros : « on-générique-1 », « on-générique-2 », etc.).
3. Mettre un label sur une expression revient à identifier une chaîne de coréférence potentielle. Toutes les expressions possédant le même label se retrouvent dans une même chaîne de coréférence. Il est donc important de ne pas donner le même label à des référents différents, en particulier pour les référents génériques.
4. Le pronom « on » est souvent impliqué dans des relations de coréférence floue. Il est aussi important de noter qu'il peut faire l'objet de relations de coréférence stricte et qu'il est nécessaire de les distinguer.
5. Dans des blocs de texte qui sont une concaténation de différents textes [6], mieux vaut distinguer des sous-chaînes génériques [7] qui apparaissent dans des textes différents par exemple. En effet, il est toujours moins coûteux de les regrouper par la suite automatiquement plutôt que de devoir faire le tri manuellement.
6. Les relations de coréférence sont strictes par définition. Il apparaît cependant parfois que des relations soient floues. Dans ce cas, il est nécessaire de les distinguer au niveau de l'annotation. Si aucun doute n'est possible sur l'identification du référent, la relation de coréférence sera stricte. Si un doute est possible et qu'il ne s'agit pas d'ambiguïté, la coréférence sera floue.

Ces remarques et précisions peuvent venir compléter un manuel d'annotation pour la coréférence tel que Democrat afin de préciser la conduite à adopter lors de l'annotation.

6. Idéalement, un bloc de texte annoté en coréférences ne devrait pas être une concaténation de plusieurs textes différents. Voir discussion en 5.2.8.
7. Comme pour « la rédaction » dans l'exemple [111].

Ces précisions vont de pair avec notre deuxième proposition qui est une adaptation du schéma d'annotation du projet Democrat afin de prendre en compte la coréférence floue.

6.1.3 Proposition 2 : de nouvelles propriétés dans le schéma

Le schéma d'annotation que nous proposons envisage d'intégrer la notion de coréférence floue au moment de l'annotation. Cela consiste à ajouter un trait caractérisant la relation de coréférence afin de préciser si elle est stricte ou floue.

Notre proposition de schéma d'annotation s'inspire du schéma proposé par le projet Democrat [8]. Il correspond au modèle d'annotation de type Unités-Relations-Schémas (URS), développé à l'origine dans le logiciel GLOZZ (Widlöcher et Mathet 2009) et implémenté dans le logiciel Analec (Landragin, Poibeau et Victorri 2012) puis par extension dans le logiciel TXM (Heiden, Magué et Pincemin 2010a).

L'annotation de la coréférence dans Democrat se fait dans un premier temps via l'identification manuelle du référent de chaque expression référentielle (*unité*), comme le montre le schéma suivant :

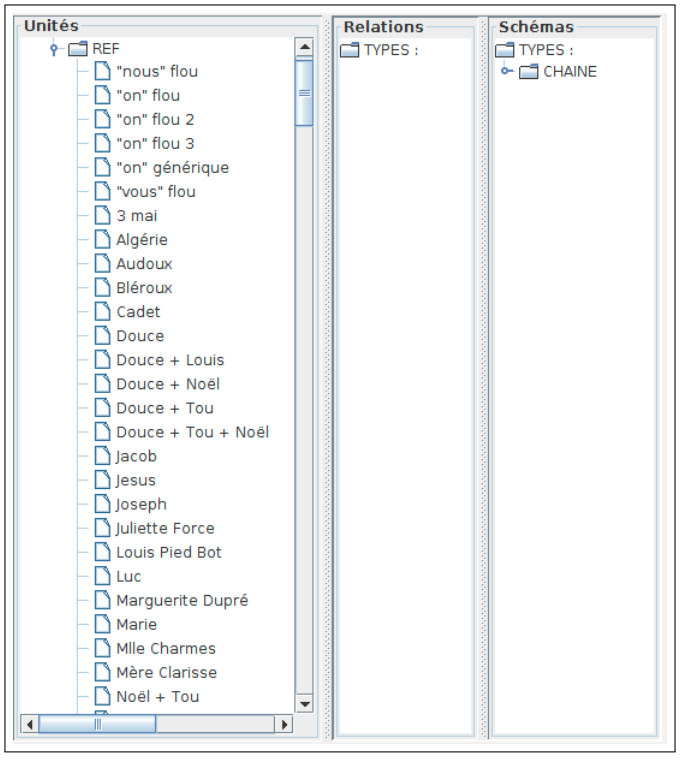


Figure 6.1 – Première phase d'annotation (manuelle) des expressions référentielles dans Democrat dans le texte *Douce Lumière*.

8. Bien que les relations entre unités ne soient pas annotées dans la version du corpus disponible à l'heure actuelle.

À la suite de cette annotation manuelle, une macro TXM permet la création automatique des chaînes (*schéma*) en regroupant les unités possédant le même label. Elle calcule aussi le nombre de maillons de chaque chaîne, comme dans le schéma suivant :

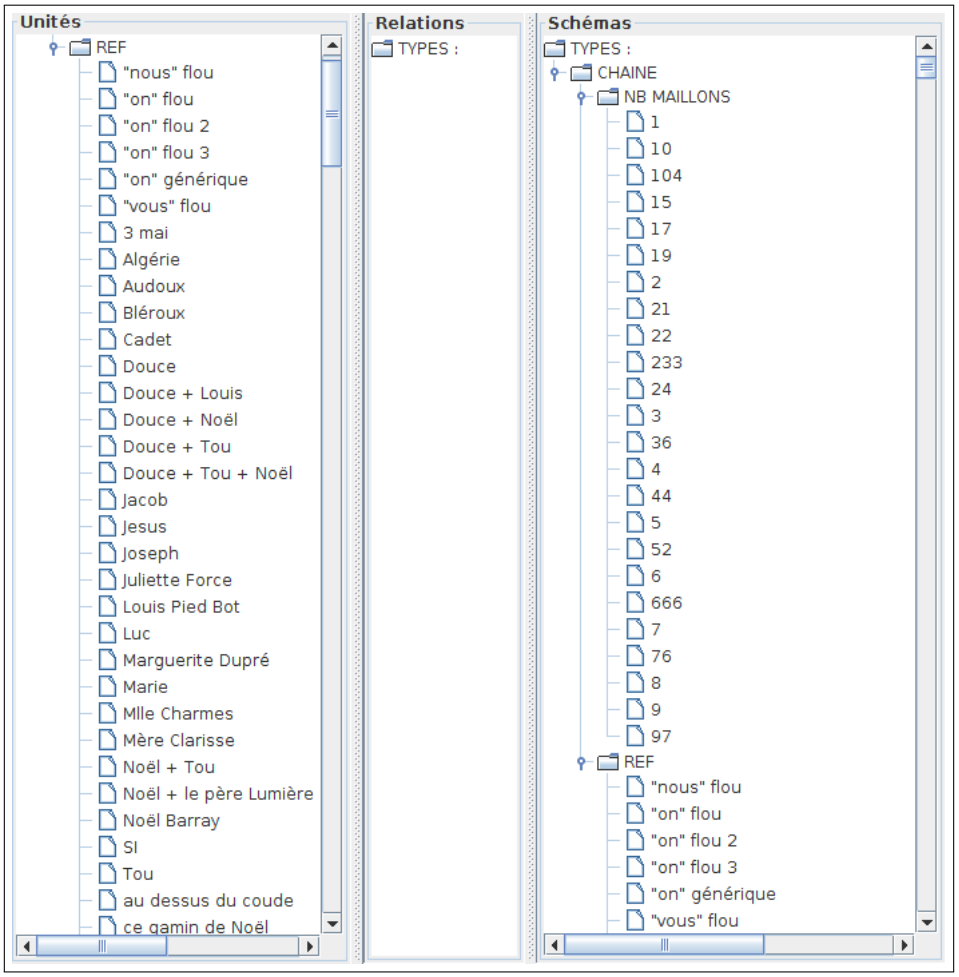


Figure 6.2 – Deuxième phase d'annotation (automatique) des expressions référentielles dans Democrat dans le texte *Douce Lumière*.

Nous souhaitons proposer une annotation du type de coréférence se faisant au niveau des relations pour chaque couple de mentions. Chaque catégorie est représentée par un trait, `stricte` ou `floue` :

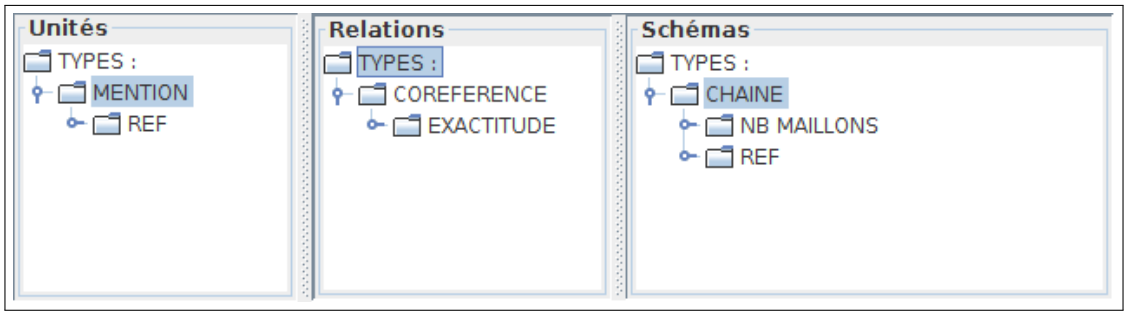


Figure 6.3 – Proposition de schéma d'annotation intégrant la coréférence floue au niveau des relations.

Le schéma de la figure [6.3] possède la même base que celui du projet Democrat auquel un type de relation « COREFERENCE » a été ajouté. Ce type comporte une propriété « EXACTITUDE » qui peut être soit « stricte » soit « floue ». Le logiciel Analec ne permettait pas l'annotation des relations, cette possibilité n'a donc pas été implémentée dans TXM par la suite. Le logiciel GLOZZ possède cette fonctionnalité, des relations anaphoriques ont par exemple été annotées dans le corpus ANCOR avec cet outil (Muzerelle, Schang, Antoine, Eshkol, Maurel, Boyer-Pelletier et al. 2013). Pourtant, la procédure d'annotation des relations dans GLOZZ ne serait pas souhaitable pour le corpus Democrat notamment car elle nécessite des aménagements de données trop coûteux (comme le découpage des blocs de texte par exemple). Les outils évoluent en fonction des besoins des projets dans lesquels ils sont impliqués et du temps disponible pour l'implémentation. Parallèlement, le choix d'un outil d'annotation dépend aussi de la nature des données à annoter. Pour contourner cette problématique, nous pouvons aussi envisager l'annotation de l'exactitude de la coréférence au niveau des schémas (et donc des chaînes), comme dans le schéma suivant :

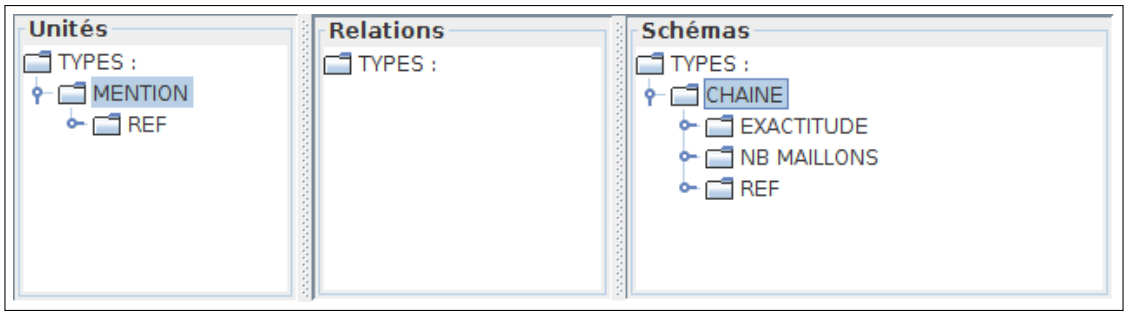


Figure 6.4 – Proposition de schéma d'annotation intégrant la coréférence floue au niveau des schémas.

Le schéma de la figure [6.4] ne possède donc plus le type de relation « COREFERENCE » mais une propriété « EXACTITUDE » a été ajoutée au type de schéma « CHAINE ». La coréférence floue pourrait de cette manière être annotée selon les chaînes. Par exemple, une chaîne stricte peut coréférer de manière floue à une autre

chaîne (stricte ou floue). Il serait aussi possible de relier un schéma possédant une seule unité (une « chaîne » de 1 maillon) à un schéma plus long.

Dans le projet Democrat, les chaînes possèdent deux propriétés (« NB MAILLONS » et « REF ») dont le champ est rempli automatiquement à partir de l'annotation manuelle des unités. Une phase d'annotation supplémentaire des chaînes avait été envisagée, elle aurait été manuelle. Cette annotation aurait eu pour but d'annoter les chaînes avec d'autres propriétés comme le type de référent (humain, animal, objet concret/abstrait, date, lieu, organisation, produit), le sexe du référent ou encore le nombre (groupe strict, groupe flou[9] ou singulier). Cette phase d'annotation n'a pas été appliquée. Cependant, si elle se faisait, l'annotation de l'exactitude des chaînes pourrait se faire à ce moment. Pour un corpus annoté en coréférence stricte[10], toutes les chaînes pourraient avoir une propriété « EXACTITUDE » égale à « stricte » par défaut. Relier les chaînes floues entre elles serait l'étape suivante.

Cette proposition ne prend pas en compte la coréférence proche. Cela pourrait être envisageable, probablement avec un schéma plus simple que celui de la *near-identity* de Recasens : en ajoutant une valeur « proche » possible pour la propriété « EXACTITUDE » des chaînes. De cette manière, ces deux propriétés « proche » et « floue » seraient complémentaires. L'exactitude de la coréférence serait « proche » lorsque les référents sont bien identifiés sans qu'ils correspondent exactement à la même entité. La valeur « floue » de la coréférence s'appliquerait plutôt à des référents qui sont liés sémantiquement mais sur lequel un doute est possible quant à l'exactitude stricte de la relation.

Notre objectif est de faire un premier pas dans la prise en compte de la coréférence non stricte en corpus, en se basant sur le modèle d'annotation de Democrat et en s'inspirant de travaux antérieurs. Pour cela, notre première proposition concerne l'ajout de six précisions dans le manuel d'annotation. Cela concerne en particulier l'annotation de la généricité, de l'indéfinitude et du pronom « on » ainsi que des considérations techniques d'annotation. Notre deuxième proposition consiste à ajouter une propriété dans le schéma d'annotation de Democrat. Elle concerne l'exactitude de la relation de coréférence qui peut être « stricte » ou « floue », voire « proche ». Ce nouveau schéma devrait être validé par une étape de double annotation et un calcul de l'accord inter-annotateur. C'est une étape qui a été réalisée dans Democrat pour la coréférence stricte à posteriori. Elle est parfois même utilisée antérieurement, comme Recasens et al. pour la *near-identity* par exemple, afin de mettre en lumière les cas problématiques et ajuster le schéma.

Annoter les phénomènes de coréférence non stricte permettrait d'obtenir des chaînes

9. Ici, le référent est flou mais pas nécessairement la relation de coréférence.
10. Ce qui est le cas de Democrat en théorie, mais nous avons relevé dans le chapitre précédent quelques « dérives ».

de coréférence qui reflètent mieux le contenu du texte. En effet, elles permettent de stocker l'information correspondant au fait que certaines expressions semblent appartenir à une chaîne sans coréférer à ses maillons de manière stricte. La coréférence stricte correspond à la majeure partie des cas, cependant la prise en compte des subtilités liées à la langue permet de mieux la représenter, notamment lorsqu'il s'agit de traitement automatique des langues.

6.2 Vers une exploitation TAL

L'annotation de la coréférence en corpus a des intérêts en linguistique afin de modéliser le phénomène et d'en tirer une description précise. Elle a aussi de nombreux intérêts en traitement automatique des langues. La prise en compte des phénomènes de coréférence non stricte dans ce domaine pourrait en particulier être bénéfique.

6.2.1 La place de la coréférence floue en corpus

Si les chaînes de coréférence stricte sont les plus nombreuses, est-ce que la coréférence floue est un phénomène marginal pour autant ? Est-il possible de savoir quelle place occupe ce phénomène sans l'avoir annoté explicitement au préalable ?

La répartition des maillons

Les chaînes de coréférence les plus longues correspondent souvent à des entités importantes dans le texte, comme les personnages principaux. En se basant sur les observations du chapitre 4, nous émettons l'hypothèse que la coréférence floue ne possède pas une grande portée et qu'elle correspond plutôt à des chaînes courtes, voire à des singletons.

Pour tenter de vérifier cette hypothèse, nous allons nous pencher sur les « dérives » d'annotation relevées dans le chapitre précédent. En effet, ces différentes approches d'annotation concernent toutes de près ou de loin la coréférence floue, bien que le phénomène annoté soit souvent simplement qualifié de « générique » ou « indéfini ». La longue chaîne annotée selon un unique référent « indéfini », par exemple, englobe plusieurs relations de coréférence floue. La figure suivante présente le diagramme de progression des chaînes[11] du bloc de texte *Bouvard et Pécuchet* :

11. Pour des raisons de lisibilité, nous avons sélectionné uniquement les chaînes à partir de 10 maillons et laissé de côté la chaîne « NO » de 800 maillons, qui sont tous des singletons.

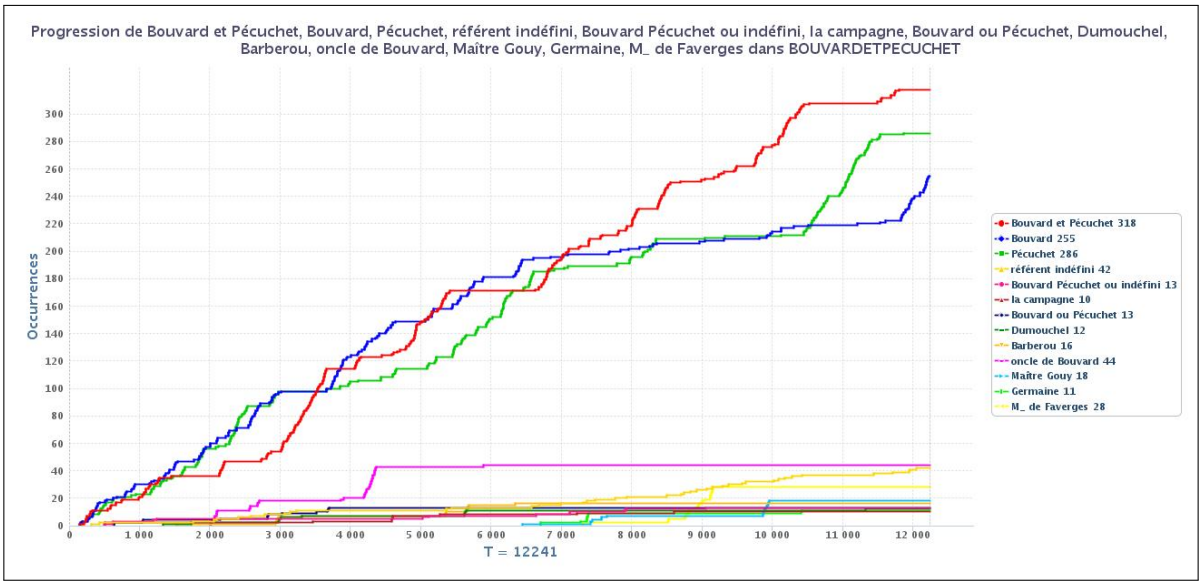


Figure 6.5 – Schéma de progression des chaînes de 10 à 799 maillons dans *Bouvard et Pécuchet*.

La figure [6.5] présente le schéma de progression des chaînes de plus de 10 maillons dans le bloc de texte *Bouvard et Pécuchet*. Ce schéma représente la présence de chaque référent dans le texte. L'axe des abscisses correspond au nombre de mots (tokens) du texte et représente ainsi la linéarité du texte. La courbe qui représente une chaîne augmente à mesure que le référent qu'elle désigne est mentionné dans le texte. Ce schéma montre que la chaîne correspondant au « référent indéfini »[12] est la troisième chaîne la plus longue de ce bloc de texte. Les trois premières chaînes correspondent aux deux personnages principaux : « Bouvard », « Pécuchet » et « Bouvard et Pécuchet ». Ces chaînes apparaissent nettement plus haut que les autres sur le graphique car elles possèdent plus de maillons et que leur répartition couvre tout le texte. Pour mieux visualiser la chaîne indéfinie et supprimer les trois chaînes principales, nous avons réduit la fourchette du nombre de maillons maximal à 50 dans la figure suivante :

12. Le problème posé par cette chaîne a été abordé dans la section [5.2.2].

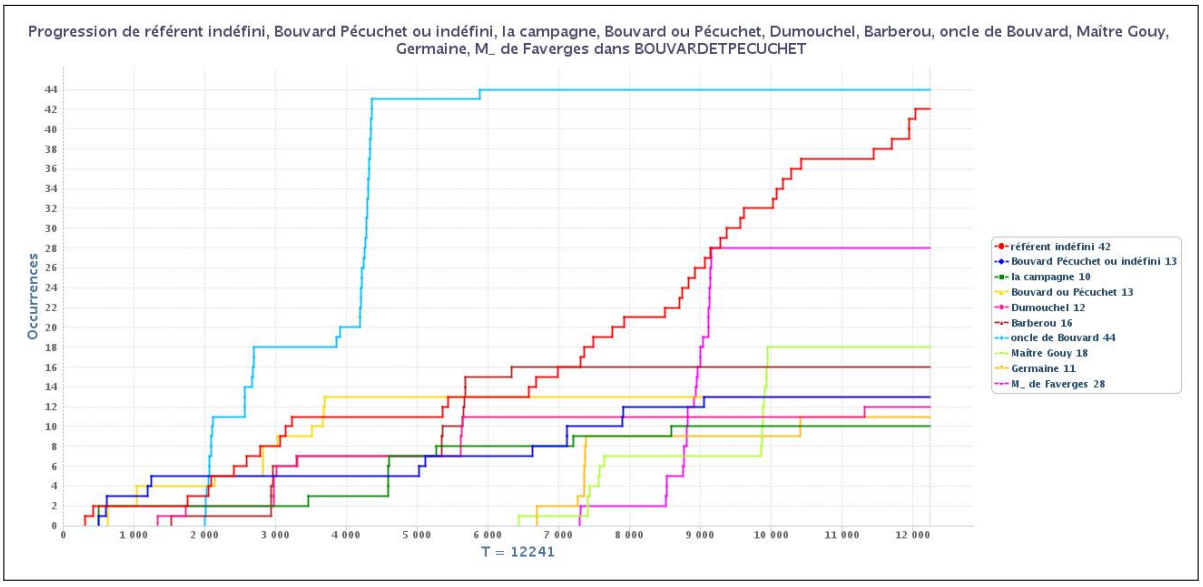


Figure 6.6 – Schéma de progression des chaînes de 10 à 50 maillons dans *Bouvard et Pécuchet*.

La figure [6.6] présente des chaînes dont la répartition est nettement plus hétérogène. En effet, la chaîne dont le référent est l'oncle de Bouvard, par exemple, possède trois paliers principaux, concentrés dans la première moitié du bloc de texte. Cela correspond aux trois passages dans lesquels il est mentionné. La chaîne indéfinie, en rouge sur cette figure, possède moins de paliers aussi nets. Ils sont plus nombreux et moins espacés. C'est-à-dire que cette chaîne est présente de manière continue tout au long du texte, avec de petites coupures. Les nombreux petits paliers que nous pouvons observer sur cette figure correspondent probablement aux réelles chaînes de coréférence stricte que nous avons abordées dans la section 5.2.2. En effet, cette longue chaîne indéfinie correspond en réalité à un regroupement maladroit de maillons qui ne sont pas tous coréférents. Cette « sur-chaîne » devrait donc être découpée en chaînes qui correspondent probablement aux paliers car nous avons pu observer que la coréférence indéfinie possède une portée limitée.

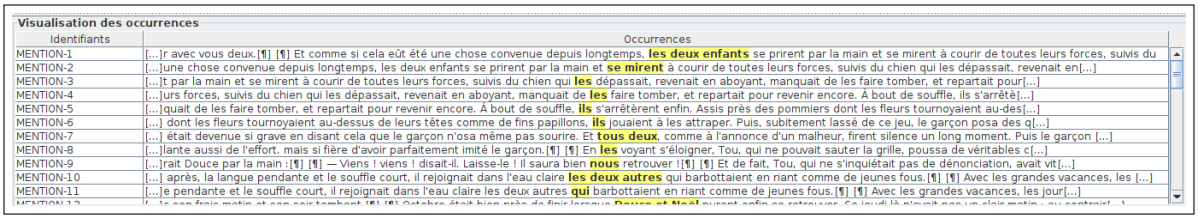
La coréférence floue, un phénomène rare ?

Le texte *Douce Lumière* possède 265 chaînes annotées de plus de deux maillons. Parmi ces chaînes, cinq chaînes désignent des groupes flous et une chaîne est générique. Chacune de ces six chaînes possède seulement deux maillons. Les chaînes désignant des groupes flous peuvent potentiellement être reliées à des chaînes de coréférence stricte. C'est le cas de la chaîne « “nous” flou » de l'exemple suivant :

Exemple [116]

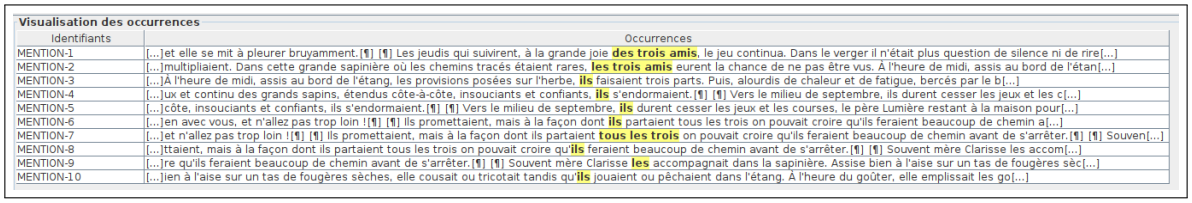
« — Tu n'étais donc pas dehors, ce matin ?
— Non ! Je suis entré chez le père Lumière pour m'abriter. Douce était toute seule avec son chien. [**Nous**]"nous" flou avons joué, et, dès qu'il a fait beau, [**nous**]"nous" flou sommes venus ici ! »
Marguerite AUDOUX, *Douce Lumière*, 1937.

Dans l'exemple [116], la chaîne « "nous" flou » peut coréférer de manière floue à la chaîne « DOUCE + NOËL » (10 maillons) mais aussi à la chaîne « DOUCE + TOU + NOËL » (52 maillons), visibles dans le schéma [6.2] et les figures suivantes :


Visualisation des occurrences

Identifiants	Occurrences
MENTION-1	[...]r avec vous deux.[¶] [¶] Et comme si cela eût été une chose convenue depuis longtemps, **les deux enfants** se prirent par la main et se mirent à courir de toutes leurs forces, suivis du
MENTION-2	[...]une chose convenue depuis longtemps, les deux enfants se prirent par la main et **se mirent** à courir de toutes leurs forces, suivis du chien qui les dépassait, revenait en[...]
MENTION-3	[...]t par la main et se mirent à courir de toutes leurs forces, suivis du chien qui **les** dépassait, revenait en aboyant, manquait de les faire tomber, et repartait pour[...]
MENTION-4	[...]urs forces, suivis du chien qui les dépassait, revenait en aboyant, manquait de **les** faire tomber, et repartait pour revenir encore. À bout de souffle, ils s'arrêtè[...]
MENTION-5	[...]quait de les faire tomber, et repartait pour revenir encore. À bout de souffle, **ils** s'arrêtèrent enfin. Assis près des pommiers dont les fleurs tournoyaient au-des[...]
MENTION-6	[...] dont les fleurs tournoyaient au-dessus de leurs têtes comme de fins papillons, **ils** jouaient à les attraper. Puis, subitement lassé de ce jeu, le garçon posa des q[...]
MENTION-7	[...] était devenue si grave en disant cela que le garçon n'osa même pas sourire. Et **tous deux**, comme à l'annonce d'un malheur, firent silence un long moment. Puis le garçon [...]
MENTION-8	[...]lante aussi de l'effort. mais si fière d'avoir parfaitement imité le garçon.[¶] [¶] En **les** voyant s'éloigner, Tou, qui ne pouvait sauter la grille, poussa de véritables c[...]
MENTION-9	[...]rait Douce par la main :[¶] [¶] — Viens ! viens ! disait-il. Laisse-le ! Il saura bien **nous** retrouver !¶] [¶] Et de fait, Tou, qui ne s'inquiétait pas de dénonciation, avait vit[...]
MENTION-10	[...] après, la langue pendante et le souffle court, il rejoignait dans l'eau claire **les deux autres** qui barbottaient en riant comme de jeunes fous.[¶] [¶] Avec les grandes vacances, les [...]
MENTION-11	[...]e pendante et le souffle court, il rejoignait dans l'eau claire les deux autres **qui** barbottaient en riant comme de jeunes fous.[¶] [¶] Avec les grandes vacances, les jour[...]

Figure 6.7 – Concordance des maillons de la chaîne « DOUCE + NOËL » dans *Douce Lumière*.


Visualisation des occurrences

Identifiants	Occurrences
MENTION-1	[...]et elle se mit à pleurer bruyamment.[¶] [¶] Les jeudis qui suivirent, à la grande joie **des trois amis**, le jeu continua. Dans le verger il n'était plus question de silence ni de rire[...]
MENTION-2	[...]multipliaient. Dans cette grande sapinière où les chemins tracés étaient rares, **les trois amis** eurent la chance de ne pas être vus. À l'heure de midi, assis au bord de l'étan[...]
MENTION-3	[...]À l'heure de midi, assis au bord de l'étang, les provisions posées sur l'herbe, **ils** faisaient trois parts. Puis, alourdis de chaleur et de fatigue, bercés par le b[...]
MENTION-4	[...]ux et continu des grands sapins, étendus côte-à-côte, insouciants et confiants, **ils** s'endormaient.[¶] [¶] Vers le milieu de septembre, ils durent cesser les jeux et les c[...]
MENTION-5	[...]côte, insouciants et confiants, ils s'endormaient.[¶] [¶] Vers le milieu de septembre, **ils** durent cesser les jeux et les courses, le père Lumière restant à la maison pour[...]
MENTION-6	[...]en avec vous, et n'allez pas trop loin ![¶] [¶] Ils promettaient, mais à la façon dont **ils** partaient tous les trois on pouvait croire qu'ils feraient beaucoup de chemin a[...]
MENTION-7	[...]et n'allez pas trop loin ![¶] [¶] Ils promettaient, mais à la façon dont ils partaient **tous les trois** on pouvait croire qu'ils feraient beaucoup de chemin avant de s'arrêter.[¶] [¶] Souven[...]
MENTION-8	[...]ttaient, mais à la façon dont ils partaient tous les trois on pouvait croire qu'**ils** feraient beaucoup de chemin avant de s'arrêter.[¶] [¶] Souvent mère Clarisse les accom[...]
MENTION-9	[...]re qu'ils feraient beaucoup de chemin avant de s'arrêter.[¶] [¶] Souvent mère Clarisse **les** accompagnait dans la sapinière. Assise bien à l'aise sur un tas de fougères sèc[...]
MENTION-10	[...]ien à l'aise sur un tas de fougères sèches, elle cousait ou tricotait tandis qu'**ils** jouaient ou pêchaient dans l'étang. À l'heure du goûter, elle emplissait les go[...]

Figure 6.8 – Concordance des maillons de la chaîne « DOUCE + NOËL+ TOU » dans *Douce Lumière*.

Les maillons des deux chaînes strictes présentées dans les figures [6.7] et [6.8] désignent explicitement les référents associés : les expressions référentielles contiennent à plusieurs reprises le nombre de référents dans les groupes (« les deux enfants », « les trois amis », etc.). Dans le texte, le chien « TOU » est souvent impliqué dans les jeux des enfants et les a suivis lors de leur excursion à la rivière (« ici »). Cependant, au moment de la lecture (et donc de l'annotation) de la phrase de l'exemple [116], il n'est pas évident de savoir si le chien est impliqué dans le « nous ». Dans le doute, en respectant le cadre strict de Democrat, une nouvelle chaîne a été créée.

Les chaînes strictes sont nettement plus nombreuses et plus longues que les chaînes impliquant de la coréférence floue. Cependant, dans Democrat, nous avons relevé des chaînes strictes dont le référent est flou et des chaînes strictes qui pourraient coréférer de manière floue avec d'autres [13]. Ces chaînes sont facilement identifiables notamment en rai-

13. Les deux ne sont pas incompatibles : la chaîne « "nous" flou » de l'exemple [116] possède ces deux

son du fait qu'elles possèdent peu de maillons. Cependant, il existe aussi des singletons qui pourraient entretenir des relations de coréférence floue avec une ou plusieurs chaînes. Ces singletons sont pour le moment difficiles à identifier. Il est néanmoins possible d'appliquer la méthodologie de recherche des expressions qui véhiculent particulièrement bien le flou, comme nous l'avons fait pour l'identification des cas présentés dans le chapitre précédent [5.2]. La prise en compte de ces singletons dans les relations de coréférence floue pourrait augmenter quelque peu le nombre d'expressions totales impliquées dans des relations de coréférence floue.

6.2.2 Intérêt pour le TAL : enjeux et problématique

Par rapport aux méthodes actuelles, la prise en compte des relations de coréférence floue pour l'identification des chaînes aurait des avantages. Le premier correspond au fait de pouvoir prendre en compte différents degrés de coréférence pour récupérer davantage d'informations sémantiques. Cela générerait par la même occasion une meilleure prise en compte du sens du texte et des relations entre les expressions référentielles. En effet, la simple prise en compte de la coréférence stricte ne suffit pas à donner un cadre précis aux annotateurs, tout en laissant de côté des subtilités de la langue qui ne sont pas prises en compte. Comme lorsqu'un « on » générique implique aussi le narrateur ou un personnage de manière floue par exemple.

La coréférence floue semble correspondre à un phénomène rare. Cependant, on la retrouve régulièrement, sous forme de chaînes courtes ou de singletons, associée à des pronoms comme « on », « nous » ou encore « vous ». Ce phénomène n'est donc pas si inhabituel.

Cette prise en compte de différents degrés de coréférence a aussi des conséquences. En linguistique de corpus, cela implique de revoir les procédures d'annotation de corpus. Mais aussi de relâcher les contraintes d'annotation en créant des schémas d'annotation plus permissifs, tel que celui que nous avons proposé dans la section [6.1.3]. Il est aussi nécessaire de relâcher les métriques d'accord inter-annotateurs en raison de la difficulté à identifier précisément les référents : une erreur dans ce cadre devrait avoir moins de poids que pour la coréférence stricte. Il existe déjà plusieurs métriques comme le *kappa* qui est plus ancien et le *gamma* qui est plus récent mais qui a été moins appliqué. Le gamma permet de distinguer deux types d'erreurs : les erreurs de délimitation et d'étiquetage.

La prise en compte de la coréférence floue dans un système de détection automatique de la coréférence implique une évaluation avec métriques adaptées au flou. Il en existe plusieurs pour la coréférence stricte. La métrique MUC a pour unité de base les liens de

caractéristiques par exemple.

coréférence. Il n'y a donc pas de considération des singletons. Les erreurs d'insertion et de suppression sont prises en compte de manière individuelle, c'est pourquoi avec cette métrique, un système qui retournerait une unique chaîne comprenant toutes les expressions référentielles du corpus aurait un très bon score, ce qui n'est pas souhaité. Pour la métrique B3, l'unité de base n'est pas le lien de coréférence mais la mention. Cela évite le biais de la longue chaîne de MUC. C'est aussi une métrique qui permet de prendre en compte les singletons. La métrique CEAF se base sur la référence commune à tous les maillons d'une chaîne. Elle engendre un temps de calcul plus long que les autres car la méthode est plus complexe et consiste à réaliser des mesures de similarité entre les chaînes. La métrique BLANC est la plus récente. Elle est intéressante car elle considère conjointement les liens de coréférence et les liens de non-coréférence.

Dans le cas d'utilisation de systèmes à base de règle, cela implique une plus grande complexité des algorithmes en raison de règles spécifiques à la coréférence floue à prévoir. Ces systèmes obtiennent souvent de meilleurs résultats de précision. Cependant, ils peuvent avoir des scores de rappel plutôt bas, notamment car ils ne sont pas facilement applicables à des textes variés. C'est pourquoi ils ne sont plus réellement utilisés dans les tâches de détection de la coréférence.

Le schéma que nous avons proposé reste simple. En effet, il faut se garder d'aller vers une multiplication des identifications de flou. Et cela même lorsque ce n'est pas pertinent, ce que nous pourrions alors appeler la « surgénération de flou ». Dans le cas d'utilisation de systèmes d'apprentissage, supervisé ou non. Il faut identifier des traits (ou features) qui aident le système à trouver des paramètres pour détecter des situations de flou. Par exemple, la simple forme de surface « on » peut être l'un de ces indices. Cependant, il faut en trouver d'autres pour l'ensemble des expressions référentielles potentiellement floues comme le pronom « ce » et ses dérivés, les pronoms « nous », « vous » et « tu » dans une moindre mesure, les groupes pluriels par exemple. Les pronoms dont l'emploi est générique ou indéfini sont plus difficiles à repérer automatiquement en raison des autres acceptions qu'ils peuvent aussi générer.

Dans les projets de traitement automatique des langues, la tâche de détection de la coréférence correspond à l'identification des relations entre les mentions qui réfèrent à la même entité afin de relier les expressions coréférentes. Le repérage des singletons, qui représentent des maillons potentiels, est une première étape primordiale. Une fois les expressions coréférentes appairées, l'information utile est souvent simplement de savoir qu'elles désignent la même entité pour pouvoir réutiliser cette information, comme une base de connaissances. La notion de chaîne n'est pas toujours prise en compte. Dans la tâche de détection de la coréférence, l'information importante est la relation entre les mentions. Notre proposition de schéma d'annotation selon lequel l'annotation du type de coréférence pourrait se faire au niveau des chaînes serait potentiellement à revoir au

niveau des relations pour la création d'un corpus d'apprentissage.

Afin d'éviter les dérives d'annotations que nous avons relevées au chapitre précédent, nous proposons un cadre plus précis pour la coréférence. La précision de ce cadre passe dans un premier temps par l'ajout d'informations dans le manuel d'annotation de Democrat. Ces informations concernent des définitions précises de notions comme le flou, la généricité ou l'indéfinitude mais aussi des remarques et des consignes sur l'annotation, notamment du pronom « on ». Dans un deuxième temps, nous proposons l'enrichissement du schéma d'annotation afin d'ajouter un trait caractérisant la coréférence pour préciser son exactitude : stricte ou floue[14]. La prise en compte de la coréférence floue en corpus a des conséquences en traitement automatique des langues. Ces conséquences concernent la performance des systèmes de détection de la coréférence comme la qualité des informations sémantiques obtenues ou encore les métriques d'évaluation de l'accord inter-annotateur et des systèmes.

14. Voire proche.

Conclusion de la partie 3

Différents projets ont tenté de prendre en compte des cas de coréférence non stricte en corpus, montrant une volonté de fournir des données plus complètes sémantiquement. Cependant, la coréférence floue n'a jamais été annotée pour la langue française. Nous avons observé les différentes pratiques d'annotation de la coréférence non stricte, et floue en particulier dans le corpus Democrat. Pour cela nous avons observé les cas typiques vecteurs de flou (co)référentiel, comme le pronom « on » en particulier. Cette étude nous a permis d'identifier sept conduites d'annotation différentes pour ces cas problématiques. Nous avons ensuite classé les conduites qui peuvent être considérées comme des dérives d'annotation selon deux catégories : un sous-regroupement de maillons ou un sur-regroupement de maillons. Le sous-regroupement de maillons revient à ne pas annoter ensemble des mentions qui sont coréférentes. Le sur-regroupement de maillons revient à annoter ensemble des maillons qui ne sont pas nécessairement coréférents. Ces deux catégories de conduite d'annotation concernent principalement des cas de flou (co)référentiel, de généricité ou d'indéfinitude. Le cadre de Democrat, pourtant déjà clair, avec un manuel d'annotation fourni et des annotateurs experts, n'a pas suffi à contrer ces dérives d'annotation. C'est pourquoi nous proposons des recommandations pour la prise en compte de ces phénomènes en corpus. Cela passe par l'ajout de précisions dans le manuel d'annotation et par la prise en compte de la coréférence floue dans le schéma d'annotation. Un corpus annoté de cette manière aura nécessairement des conséquences en traitement automatique des langues, comme la performance des systèmes ou leur évaluation. Cela nécessite donc des ajustements, notamment au niveau des métriques.

Conclusion

Bilan

Cette thèse avait pour objectif premier d'effectuer une analyse en corpus de chaînes de coréférence. Dans un premier temps, nous avons donc défini le cadre de notre étude d'un point de vue théorique. Ce cadre concerne donc principalement les chaînes de coréférence. Il s'agit de l'ensemble des expressions référentielles qui désignent le même référent dans le discours. Ces expressions référentielles sont alors coréférentes. Elles peuvent aussi avoir une interprétation anaphorique, bien que la coréférence n'implique pas nécessairement l'anaphore, et inversement. Toutefois, nous ne prenons pas en compte les cas d'anaphores sans coréférence. Une chaîne de coréférence permet le suivi discursif du référent, même lorsqu'elle ne contient que deux maillons. Dans les travaux sur la coréférence et les chaînes de coréférence, il est établi qu'une relation de coréférence implique que le référent soit exactement le même pour plusieurs expressions référentielles. Nous avons passé en revue les différents phénomènes linguistiques impliqués dans les chaînes de coréférence : la référence, les expressions référentielles, la coréférence ou encore l'anaphore. Cela en fait un objet d'étude complexe qui est utilisé dans différents domaines, sous différentes dénominations qui ont nécessité un point terminologique.

Le travail présenté dans cette thèse a pris source au sein du projet Democrat, dont le cadre correspond aux définitions données dans le paragraphe précédent. L'apport de cette thèse pour le projet a débuté par une première expérience d'annotation de corpus en coréférence. Cela correspond à 4 blocs de textes de 10 000 mots pour un corpus qui en comporte 58. Comme tout travail d'annotation, il a été défini par un cadre théorique précis qui a imposé aux membres du projet de faire des choix répertoriés dans le manuel d'annotation. Les annotations réalisées pour le projet sont disponibles et peuvent maintenant être analysées de manière statistique grâce aux avancées de la plateforme TXM de l'équipe de l'ENS de Lyon. Ces avancées ont permis un nouvel élan vers une méthodologie d'analyse des chaînes de coréférence qui est en cours à l'heure actuelle. Le corpus de ce projet est annoté en coréférence stricte. Cependant, comme le souligne la deuxième partie de cette thèse, la coréférence n'est pas toujours exactement stricte. En effet, il est parfois difficile d'attribuer un référent à une chaîne de manière certaine car deux chaînes peuvent par exemple se recouper de manière floue. Cela peut générer des difficultés aussi bien au niveau technique que théorique.

Il existe des cas de coréférence plus subtils, correspondant à la complexité et à la richesse de la langue, qu'il n'est pas toujours aisé de catégoriser. L'ambiguïté, la coréférence proche et les référents évolutifs sont trois phénomènes linguistiques qui attestent de la subtilité du langage. La prise en compte de ces phénomènes dans des travaux ayant des visées de traitement automatique du langage ajoute une complexité nécessaire pour retranscrire la complexité et la richesse du langage. L'ambiguïté référentielle et la coréférence proche impliquent des référents identifiés de manière précise. Pour l'ambiguïté, il s'agit d'un choix entre des référents potentiels, qui ne sont pas proches sémantiquement. La levée d'une ambiguïté devrait pouvoir se faire en fonction de critères de saillance : le référent le plus saillant, celui qui ressort le plus, devrait être le candidat choisi comme référent. Cependant, une ambiguïté effective ne peut pas être levée et les critères de saillance peuvent mener ce choix vers des directions opposées. Pour la coréférence proche, il s'agit d'établir le degré de coréférence que peuvent entretenir deux expressions référentielles ayant des référents bien identifiés qui entretiennent des relations sémantiques. Les référents évolutifs sont à la limite entre l'ambiguïté référentielle, la coréférence proche et le flou coréférentiel car il est souvent difficile d'opérer un choix et de trancher pour décider à quel moment le référent change. Cette difficulté soulève une question d'ordre technique et conceptuel : est-il toujours nécessaire de trancher ?

Plusieurs auteurs avaient déjà abordé les notions de flou ou de vague, principalement en lien avec la dénotation plutôt qu'avec la référence. Le flou référentiel et coréférentiel peut apparaître sous différentes formes. Il existe des cas typiques de ce phénomène qui sont : les groupes pluriels et certains pronoms, comme « on » et « ce » en particulier. Dans chacun de ces cas de figure, le flou se manifeste de différentes manières. Il s'agit souvent d'une superposition de sens potentiels, relevant plutôt de la sur-détermination. Cette superposition implique plusieurs référents, correspondant souvent à des groupes, déjà parfois flous eux-même. Elle peut aussi régulièrement impliquer une référence générique qui se mêle à une référence spécifique. Des expériences psycholinguistiques nous ont montré qu'il n'est pas toujours nécessaire de résoudre ce flou pour comprendre un texte. Une autre manière d'appréhender l'annotation de ce phénomène en corpus est sûrement possible, en prenant en compte le fait que le flou coréférentiel ne doit pas toujours être levé mais plutôt caractérisé comme tel.

Malgré les tentatives d'autres projets pour annoter la coréférence non stricte, le flou des relations de coréférence n'a jamais été pris en compte en corpus alors qu'il génère des conduites d'annotation hétérogènes. Au sein du corpus Democrat, nous avons étudié les cas particuliers vecteurs de flou et leur annotation dans les blocs de texte par les différents annotateurs. Cela nous a permis de relever différentes approches d'annotation. Certaines annotations correspondent à une volonté de regrouper les référents indéfinis, d'autres à regrouper les référents génériques, en les confondant parfois. Une autre approche consiste

à rapprocher des expressions en fonction de leur forme physique : le « on » en particulier. D'autres annotations effectuent un regroupement des référents génériques selon la structure textuelle. Certaines annotations ne regroupent des expressions génériques ou floues que lorsqu'elles sont effectivement coréférentes. Une autre approche prend plutôt le parti de ne pas regrouper les expressions génériques ou indéfinies coréférentes, les considérant parfois comme étant non référentielles. Une dernière approche consiste plutôt à tenter d'identifier un référent précis systématiquement. Certaines de ces approches pourraient être qualifiées de « dérives d'annotation ». En effet, le manuel d'annotation recommande l'annotation des relations de coréférence stricte et ces dérives ne respectent pas ce critère. Le choix de l'annotation en coréférence stricte semble pourtant pertinent car il paraît simple d'annoter uniquement les relations sur lesquelles il n'y a aucun doute. Cependant, l'annotation du pronom « on » en particulier a généré des comportements de sur-regroupements (fréquents) et de sous-regroupements (moins fréquents) de maillons. L'annotation en coréférence stricte n'est finalement pas si évidente. Un cadre plus précis et permettant d'indiquer les relations de coréférence floue serait selon nous nécessaire. De plus, les chaînes de coréférence représentent l'évolution référentielle d'une entité au fil d'un texte et il est regrettable de perdre des informations sur cette entité en mettant la coréférence floue de côté.

Afin d'éviter les dérives d'annotations que nous avons relevées, nous avons proposé un cadre plus précis pour la coréférence. La précision de ce cadre passe dans un premier temps par l'ajout d'informations dans le manuel d'annotation (celui de Democrat). Ces informations concernent des définitions précises de notions comme le flou, la généricité ou l'indéfinitude mais aussi des remarques et des consignes sur l'annotation, notamment du pronom « on » et un rappel sur l'annotation par référent et non par chaîne. Dans un deuxième temps, nous proposons l'enrichissement du schéma d'annotation afin d'ajouter un trait caractérisant la coréférence pour préciser son exactitude : stricte ou floue. La prise en compte de la coréférence floue en corpus a des conséquences en traitement automatique des langues. Ces conséquences concernent la performance des systèmes de détection de la coréférence comme la qualité des informations sémantiques obtenues ou encore les métriques d'évaluation de l'accord inter-annotateur et des systèmes.

Discussion

Pour l'annotation de la coréférence non stricte, il est aussi envisageable d'ajouter une valeur « proche » pour préciser l'exactitude de la coréférence. Dans le corpus Democrat, les chaînes ont été annotées automatiquement (le référent et le nombre de maillons). Leur annotation manuelle avait été envisagée mais n'a pas eu lieu. Nos propositions seraient

applicable dès le début d'un projet comme Democrat (notamment pour les précisions dans le manuel d'annotation). La modification du schéma rajouterait en revanche une étape d'annotation manuelle des chaînes pour caractériser leur lien de coréférence (stricte ou floue).

Nos propositions se basent sur des analyses en corpus de phénomènes repérés lors de l'annotation manuelle de la coréférence dans le corpus Democrat. Il s'agit de recommandations pour la prise en compte de la coréférence floue en corpus, à travers une méthodologie de traitement de ces phénomènes. Ces propositions devraient être validées par une étape de double annotation ainsi qu'un calcul de l'accord inter-annotateur.

La notion de flou reste subjective et l'annotation de ce phénomène au niveau des relations de coréférence risque de générer davantage de désaccords que la coréférence stricte. C'est pourquoi les métriques de calcul de l'accord inter-annotateur devraient être adaptées pour ce travail.

Perspectives

Une expérience psycholinguistique a été montée en collaboration avec Lucie Rousier-Vercruyssen. Cette expérience a pour but de déterminer l'opinion des lecteurs selon diverses interprétations que peut prendre le pronom « on » et les relations de coréférence qu'il peut entretenir avec d'autres expressions référentielles. L'expérience regroupe vingt exemples correspondant à quatre catégories (« "on" générique », « coréférence stricte », « coréférence floue » et « coréférence inclusive floue »). Des difficultés techniques ont empêché la mise en œuvre de cette expérience, reportée à une date ultérieure. Cette mise en œuvre et les résultats qui pourront en être tirés constituent une perspective de recherche à court terme.

À long terme, la prise en compte de la coréférence floue dans un système de traitement automatique des langues est envisageable. Cela ne devrait pas nécessiter un nouveau système à part entière. Il serait intéressant d'observer la réaction d'un système de détection de la coréférence prenant en compte la propriété d'exactitude de la coréférence que nous avons proposé. Cela permettrait de savoir si le système est capable de les détecter seul ou si des traits particuliers sont nécessaires pour aider cette détection.

Bibliographie

[Abn91] Steven P Abney. « Parsing by chunks ». In : *Principle-based parsing*. Springer, 1991, p. 257–278 (Cité page 48).

[Ach98] Guy Achard-Bayle. « Référence, identité, changement : la désignation des référents en contextes évolutifs. Études de cas : les récits de métamorphoses ». In : *L'information grammaticale* 77.1 (1998), p. 50–53 (Cité page 93).

[Ach01] Guy Achard-Bayle. « Grammaire des métamorphoses ». In : (2001) (Cité page 90).

[Ach16] Guy Achard-Bayle. « Les référents évolutifs, objets et objets du discours ». In : *Connexion et Indexation. Colloque d'hommages en l'honneur du professeur Michel Charolles*. ENS éditions. 2016, p. 83–90 (Cité page 90).

[Adg03] David Adger. *Core syntax : A minimalist approach*. T. 20. Oxford University Press Oxford, 2003 (Cité page 18).

[ARR19] Felipe Alfaro, Marta Ruiz Costa-Jussà et José Adrian Rodriguez Fonollosa. « BERT masked language modeling for co-reference resolution ». In : *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 2019, p. 76–81 (Cité page 27).

[All09] Keith Allan. *Concise encyclopedia of semantics*. Elsevier, 2009 (Cité page 24).

[Als87] Hiyan Alshawi. *Memory and context for language interpretation*. Cambridge University Press, 1987 (Cité page 72).

[ADR05] Pascal Amsili, Pascal Denis et Laurent Roussarie. « Anaphores abstraites en français : représentation formelle ». In : *TAL. Traitement automatique des langues* 46.1 (2005), p. 15–39 (Cité page 103).

[ACG11] Corinna Anderson, Christophe Cerisara et Claire Gardent. « Vers la détection des dislocations à gauche dans les transcriptions automatiques du Français parlé ». In : *TALN 2011* (2011) (Cité page 77).

[Ans02] Jean-Claude Anscombre. « La nuit, certains chats sont gris, ou la généricité sans syntagme générique ». fr. In : *Linx. Revue des linguistes de l'université Paris X Nanterre* 47 (2002), p. 13–30. issn : 0246-8743. doi : 10.4000/linx.558. url : http://journals.openedition.org/linx/558 (visité le 22/07/2020) (Cité page 101).

[Ans17] Jean-Claude Anscombre. « Génériques et généricités en français ». fre. In : *Cahiers de lexicologie 2017 – 2, n° 111. La sémantique en France : un état des lieux (II)* (2017), p. 29–55. ISSN : 2262-0346. DOI : 10.15122/isbn.978-2-406-07412-0.p.0029 (Cité page 101).

[AB95] Chinatsu Aone et Scott William Bennett. « Evaluating automated and manual acquisition of anaphora resolution strategies ». In : *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 1995, p. 122–129 (Cité page 27).

[Apo02] Denis Apothéloz. *La construction du lexique français : principes de morphologie dérivationnelle*. Editions Ophrys, 2002 (Cité page 11).

[Apo10] Denis Apothéloz. *L'opacité référentielle : paramètres et statuts discursifs*. 2010 (Cité page 100).

[AR95] Denis Apothéloz et Marie-José Reichler-Béguelin. « Construction de la référence et stratégies de désignation ». In : (1995) (Cité page 92).

[AR15] Denis Apotheloz et Marie-Noëlle Roubaud. *Constructions pseudo-clivées*. 2015. URL : http://encyclogram.fr/notx/003/003_Notice.php (Cité page 75).

[Arg17] Anne-Marie Argenti. « Le pluriel dans les chaînes anaphoriques faisant référence à des particuliers ». 2017USPCA166. Thèse de doct. 2017. URL : http://www.theses.fr/2017USPCA166 (Cité page 107).

[Ari88] Mira Ariel. « Referring and accessibility ». In : *Journal of linguistics* 24.1 (1988), p. 65–87 (Cité page 71).

[Ari90] Mira Ariel. *Accessing noun-phrase antecedents*. Routledge, 1990 (Cité pages 22, 63, 72, 91).

[AP08] Ron Artstein et Massimo Poesio. « Inter-coder agreement for computational linguistics ». In : *Computational Linguistics* 34.4 (2008), p. 555–596 (Cité page 44).

[Ash12] Nicholas Asher. *Reference to abstract objects in discourse*. T. 50. Springer Science & Business Media, 2012 (Cité page 103).

[Asn04] Maria Asnes. *Référence nominale et verbale : analogies et interactions*. Presses de l'Université Paris-Sorbonne, 2004 (Cité page 33).

[Atl84] Françoise Atlani. « On l'illusionniste ». In : *La langue au ras du texte* (1984), p. 13–29 (Cité page 114).

[AHG99] Saliha Azzam, Kevin Humphreys et Robert Gaizauskas. « Using coreference chains for text summarization ». In : *Coreference and Its Applications*. 1999 (Cité page 28).

[Bag98] Amit Bagga. « Evaluation of coreferences and coreference resolution systems ». In : *Proceedings of the First Language Resource and Evaluation Conference*. 1998, p. 563–566 (Cité page 28).

[Bal65] Charles Bally. *Linguistique générale et linguistique française*. Franke, 1965 (Cité page 76).

[Bar10] Chris Barker. « Nominals don't provide criteria of identity ». In : *The Semantics of Nominalizations across Languages and Frameworks. Mouton de Gruyter, Berlin* (2010), p. 9–24 (Cité page 86).

[Bar70] Roland Barthes. *S/Z*. Seuil, 1970 (Cité page 7).

[Bau12] Emmanuel Baumer. « Noms propres et anaphores nominales en anglais et en français : étude comparée des chaînes de référence ». Thèse de doct. Paris 7, 2012 (Cité page 25).

[Bea04] David I Beaver. « The optimization of discourse anaphora ». In : *Linguistics and philosophy* 27.1 (2004), p. 3–56 (Cité page 27).

[Ben66] Émile Benveniste. *Problèmes de linguistique générale*. Éditions Gallimard. T. Tome 1. Paris, 1966 (Cité page 130).

[Ber97] Sabine Bergler. « Towards reliable partial anaphora resolution ». In : *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*. 1997 (Cité page 64).

[Ber15] A. Berrendonner. *Constructions disloquées*. 2015. URL : http://encyclogram.fr/notx/001/001_Notice.php (visité le 16/06/2020) (Cité pages 76, 77).

[BNN01] Philippe Bessières, Adeline Nazarenko et Claire Nédellec. « Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques ». fr. In : *Actes du 4e Colloque International sur le Document Electronique*. 2001, p. 1–11 (Cité page 40).

[Bla87] Claire Blanche-Benveniste. *Pronom et syntaxe : l'approche pronominale et son application au français*. T. 1. Peeters Publishers, 1987 (Cité page 77).

[Bla03] Claire Blanche-Benveniste. « Le double jeu du pronom on ». In : *Hadermann, P., A., Van Slijcke et M. Berré (éds). La syntaxe raisonnée. Mélanges de linguistique générale offerts à Annie Boone à l'occasion de son 60e anniversaire. Louvain-la-Neuve : De Boeck Duculot* (2003), p. 43–56 (Cité page 108).

[BC12] Claire Blanche-Benveniste et André Chervel. « Recherches sur le syntagme substantif ». fre. In : *Cahiers de lexicologie 1966 — 2, 9. varia* (août 2012), p. 5–39. ISSN : 2262-0346. DOI : 10.15122/isbn.978-2-8124-4262-9.p.0005 (Cité page 17).

[Bla97] Mylène Blasco. « Pour une approche syntaxique des dislocations ». In : *Journal of French Language Studies* 7.1 (1997), p. 1–21 (Cité pages 76, 77).

[Bla99] Mylène Blasco-Dulbecco. *Les dislocations en français contemporain : étude syntaxique*. T. 1. Honoré Champion, 1999 (Cité page 167).

[Bla06] Mylène Blasco-Dulbecco. « Propositions pour le classement typologique de quelques détachements ». In : *L'Information grammaticale* 109.1 (2006), p. 27–33 (Cité page 78).

[Bon14] Marc Bonhomme. *Pragmatique des figures du discours*. Honoré Champion, 2014 (Cité page 130).

[Bou04] Sylvie Boudreau. « Résolution d'anaphores et identification des chaînes de coréférence selon le type de texte ». Thèse de doct. 2004 (Cité pages 27, 64).

[BK05] Sylvie Boudreau et Richard Kittredge. « Résolution des anaphores et détermination des chaînes de coréférences : différences entre variétés de textes ». In : *TAL. Traitement automatique des langues* 46.1 (2005), p. 41–69 (Cité pages 22, 23, 28).

[Bou99] Tayeb Bouguerra. « L'autre je(u) du on ». In : *L'Autre en discours. Montpellier : Publications de l'Université de Paul Valéry, Montpellier III* (1999) (Cité page 114).

[Bou86] Josiane Boutet. « La référence à la personne en français parlé : le cas de "on" ». In : *Langage & société* 38.1 (1986), p. 19–49 (Cité pages 113, 129).

[BY83] Gillian Brown et George Yule. *Discourse analysis*. Cambridge university press, 1983 (Cité pages 89, 90).

[BR97] Florence Bruneseaux et Laurent Romary. « Codage des références et coréférences dans les DHM ». In : 1997 (Cité pages 38, 46).

[Cab08] Patricia Cabredo Hofherr. « Les pronoms impersonnels humains : syntaxe et interprétation ». In : *Modèles linguistiques* 29.57 (2008) (Cité page 108).

[Cad00] Sandrine Caddéo. « L'apposition : analyse syntaxique de l'apposition nominale détachée dans divers registres de la langue parlée et de l'écrit en français contemporain ». Thèse de doct. Aix-Marseille 1, 2000 (Cité page 108).

[Car82] Greg N. Carlson. « Generic terms and generic sentences ». en. In : *Journal of Philosophical Logic* 11.2 (1982), p. 145–181. issn : 1573-0433. doi : 10.1007/BF00278382. url : https://doi.org/10.1007/BF00278382 (visité le 22/07/2020) (Cité page 101).

[Car56] Rudolf Carnap. « Meaning and Necessity, enlarged edition ». In : *University of Chicago, Chicago* (1956) (Cité page 80).

[Cha76] Wallace L. Chafe. « Givenness, contrastiveness, definiteness, subjects, topics, and point of view ». In : *Subject and Topic*. Sous la dir. de Charles N. Li. New York : Academic Press, 1976, p. 25–55 (Cité page 27).

[CKP09] Jon Chamberlain, Udo Kruschwitz et Massimo Poesio. « Constructing an anaphorically annotated corpus with non-experts : Assessing the quality of collaborative annotations ». In : *Proceedings of the 2009 workshop on the people's web meets NLP : Collaboratively constructed semantic resources*. Association for Computational Linguistics. 2009, p. 57–62 (Cité page 44).

[CPK08] Jon Chamberlain, Massimo Poesio et Udo Kruschwitz. « Phrase detectives : A web-based collaborative annotation game ». In : *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*. 2008, p. 42–49 (Cité page 44).

[CPK16] Jon Chamberlain, Massimo Poesio et Udo Kruschwitz. « Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference ». en. In : *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, 2016, p. 2039–2046 (Cité page 137).

[Cha92] Patrick Charaudeau. *Grammaire du sens et de l'expression*. Hachette Paris, 1992 (Cité page 108).

[CM02] Patrick Charaudeau et Dominique Maingueneau. *Dictionnaire d'analyse du discours*. Seuil, 2002 (Cité page 24).

[Cha83] Marie-Thérèse Charlent. « Notes sur la coréférence ». fre. In : (1983). issn : 0246-8743. doi : 10.3406/linx.1983.974. url : https://www.persee.fr/doc/linx_0246-8743_1983_num_8_1_974 (Cité page 24).

[Cha88] Michel Charolles. « Les plans d'organisation textuelle : périodes, chaînes, portées et séquences ». fre. In : (1988). issn : 0338-2389. doi : 10.3406/prati.1988.1468. url : https://www.persee.fr/doc/prati_0338-2389_1988_num_57_1_1468 (Cité page 28).

[Cha97] Michel Charolles. « IDENTITE, CHANGEMENT ET REFERENCE PRONOMINALE ». In : *La continuité référentielle*. Sous la dir. de G.Kleiber, C.Schnedecker et J.E.Tyvaert éds. Klincksieck, 1997, p. 71–97. url : https://hal.archives-ouvertes.fr/hal-01404638 (Cité page 98).

[Cha01] Michel Charolles. *Référents évolutifs et évolution de la référence*. 2001 (Cité pages 92, 98).

[Cha02] Michel Charolles. *La référence et les expressions référentielles en français*. fr. Ophrys, 2002. isbn : 978-2-7080-1014-7 (Cité pages 7, 8, 10, 13, 68, 90).

[Cha14] Michel Charolles. *Annotation des expressions référentielles et profondeur de traitement*. 2014 (Cité page 131).

[CF98] Michel Charolles et Jacques François. *Les prédicats transformateurs et leurs patient : fondements d'une ontologie naturelle*. Cahiers de Recherche Linguistique, 11, LANDISCO, Université de Nancy 2. 1998. URL : https://hal.archives-ouvertes.fr/hal-01404729 (Cité page 90).

[Cha75] Charles Chastain. « Reference and context ». In : *Language Mind and Knowledge*. K. Gunderson. Minneapolis, 1975, p. 194–269 (Cité page 28).

[Cho87] Noam Chomsky. *La nouvelle syntaxe : concepts et conséquences de la théorie du gouvernement et du liage*. T. 14. Seuil, 1987 (Cité page 24).

[CAT00] Catherine Clouzot, Georges Antoniadis et Agnès Tutin. « Constitution and exploitation of an annotation system of electronic corpora : Toward automatic generation of understandable pronouns in French language ». In : *International Conference on Natural Language Processing*. Springer. 2000, p. 242–251 (Cité page 38).

[Col+12] Maud Colléter et al. *La ressource ANNODIS multi-échelle : guide d'annotation et bonus*. Rapp. tech. 2012. URL : https://hal.archives-ouvertes.fr/hal-00983076 (Cité page 151).

[Com96] Bernard Combettes. « Facteurs textuels et facteurs sémantiques dans la problématique de l'ordre des mots : le cas des constructions détachées ». In : *Langue française* (1996), p. 83–96 (Cité page 70).

[Con05] Anne Condamines. « Anaphore nominale infidèle et hyperonymie : le rôle du genre textuel ». In : *Revue de Sémantique et Pragmatique* (2005), p. 23–42 (Cité page 25).

[CBD97] Dennis Connolly, John D Burger et David S Day. « A machine learning approach to anaphoric reference ». In : *New methods in language processing*. 1997, p. 133–144 (Cité page 27).

[Cor85a] Francis Corblin. « Les chaînes de référence : analyse linguistique et traitement automatique ». fre. In : (1985). ISSN : 0769-4113. DOI : 10.3406/intel.1985.851. URL : https://www.persee.fr/doc/intel_0769-4113_1985_num_1_1_851 (Cité pages 16, 28).

[Cor85b] Francis Corblin. « Remarques sur la notion d'anaphore ». In : *Revue québécoise de linguistique* 15.1 (1985), p. 173–195 (Cité pages 17, 18).

[Cor95] Francis Corblin. *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. fr. Presses Universitaires de Rennes, 1995 (Cité page 30).

[Cor+99] Francis Cornish et al. *Anaphora, discourse, and understanding : Evidence from English and French*. Oxford University Press, 1999 (Cité page 24).

[Cor10] Francis Cornish. « Anaphora : Text-based or discourse-dependent ? : Functionalist vs. formalist accounts ». In : *Functions of language* 17.2 (2010), p. 207–241 (Cité page 9).

[CW05] Richard Craggs et Mary McGee Wood. « Evaluating discourse and dialogue coding schemes ». In : *Computational Linguistics* 31.3 (2005), p. 289–296 (Cité page 44).

[Cry08] David Crystal. *A dictionary of linguistics and phonetics*. Oxford : Wiley-Blackwell, 2008. isbn : 978-1-4051-5297-6 (Cité page 102).

[CWM07] Aron Culotta, Michael Wick et Andrew McCallum. « First-order probabilistic models for coreference resolution ». In : *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Proceedings of the Main Conference*. 2007, p. 81–88 (Cité page 27).

[Dah85] Östen Dahl. « Remarques sur le Générique ». fre. In : (1985). issn : 0458-726X. doi : 10.3406/lgge.1985.2470. url : https://www.persee.fr/doc/lgge_0458-726x_1985_num_20_79_2470 (Cité page 101).

[DM05] Hal Daumé III et Daniel Marcu. « A large-scale exploration of effective global features for a joint entity detection and tracking model ». In : *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2005, p. 97–104 (Cité page 27).

[DL19] Marine Delaborde et Frédéric Landragin. « En quoi le pronom « on » a-t-il une valeur anaphorique ? . Le cas des successions d'occurrences de « on » ». fr. In : *Cahiers de praxématique* 72 (juin 2019). issn : 0765-4944. url : http://journals.openedition.org/praxematique/5464 (visité le 19/12/2019) (Cité page 105).

[Dem11] Annemie Demol. *Les pronoms anaphoriques il et celui-ci*. Recherches. De Boeck, 2011. isbn : 9782801116395 (Cité pages 71, 123).

[DG00] Jean-Pierre Desclés et Zlatka Guentchéva. « Enonciateur, locuteur, médiateur dans l'activité dialogique ». In : *Les rituels du dialogue, Nanterre, Société d'Ethnologie* (2000), p. 79–112 (Cité page 100).

[Dés+15] Adèle Désoyer et al. « Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR ». In : *Traitement Automatique des Langues* 55.2 (2015), p. 97–121 (Cité page 30).

[Dod+04] George Doddington et al. « The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation ». en. In : *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal, 2004, p. 837–840 (Cité page 136).

[DT72] Oswald Ducrot et Tzvetan Todorov. *Dictionnaire encyclopédique des sciences du langage*. fr. FeniXX, 1972. isbn : 978-2-02-127059-4 (Cité pages 1, 8, 16).

[Dup02] Michel Dupont. « Une approche cognitive pour le calcul des chaînes de référence ». In : *Actes de TALN, Campus Côte de Nacre* (2002) (Cité page 72).

[Dup03] Michel Dupont. « Une approche cognitive du calcul de le référence ». Thèse de doct. Caen, 2003 (Cité page 72).

[Elm19] Daniel Elmiger. « Les genres récrits n° 5 ». In : *GLAD ! [En ligne]* 06 (2019). consulté le 22 mai 2020. url : https://www.revue-glad.org/1541 (Cité page 92).

[EG87] Feride Erku et Jeanette K Gundel. « 28. The pragmatics of indirect anaphors ». In : *The pragmatic perspective*. John Benjamins, 1987, p. 533 (Cité page 17).

[Esh15] Iris Eshkol-Taravella. « La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral ». Thèse de doct. 2015 (Cité page 43).

[Esh+11] Iris Eshkol-Taravella et al. « Un grand corpus oral "disponible" : le corpus d'Orléans 1 1968-2012 ». In : (2011) (Cité page 39).

[Fau74] Gilles Fauconnier. *La coréférence : syntaxe ou sémantique ?* Éditions du Seuil, Paris, 1974 (Cité page 24).

[Fau84] Gilles Fauconnier. « Espaces Mentaux Aspects de la Construction du Sens Dans les Langues Naturelles ». In : (1984) (Cité page 86).

[Fau97] Gilles Fauconnier. *Mappings in thought and language*. Cambridge University Press, 1997 (Cité page 86).

[FT08] Gilles Fauconnier et Mark Turner. *The way we think : Conceptual blending and the mind's hidden complexities*. Basic Books, 2008 (Cité page 86).

[FHR20] Silvia Federzoni, Lydia-Mai Ho-Dac et Josette Rebeyrolle. « Les chaînes topicales dans la ressource ANNODIS ». In : *CMLF2020 : 7e Congrès Mondial de Linguistique Française*. Montpellier, France, juil. 2020. url : https://hal.archives-ouvertes.fr/hal-02890989 (Cité page 151).

[FBF02] Fernanda Ferreira, Karl G.D. Bailey et V. Ferraro. « Good-Enough Representations in Language Comprehension ». In : *Current Directions in Psychological Science* 11.1 (2002), p. 11–15 (Cité pages 130, 131, 167).

[FS00] Fernanda Ferreira et Janis Stacey. « The misinterpretation of passive sentences ». In : *Manuscript submitted for publication* (2000) (Cité page 131).

[Fer93] Stéphane Ferret. « Le philosophe et son scalpel ». In : (1993) (Cité page 91).

[Fin75] Kit Fine. « Vagueness, Truth and Logic ». In : *Synthese* (1975), p. 265–300 (Cité page 102).

[Fli92] Steve Fligelstone. « Developing a scheme for annotating text to show anaphoric relations ». In : *New Directions in English Language Corpora : Methodology, Results, Software Developments* (1992), p. 153–170 (Cité page 45).

[FJN07] Kjersti Fløttum, Kerstin Jonasson et Coco Norén. *On : pronom à facettes*. De Boeck/Duculot, 2007 (Cité page 115).

[FAC11] Karën Fort, Gilles Adda et K Bretonnel Cohen. « Amazon mechanical turk : Gold mine or coal mine ? » In : *Computational Linguistics* 37.2 (2011), p. 413–420 (Cité page 44).

[FEN09] Karën Fort, Maud Ehrmann et Adeline Nazarenko. « Vers une méthodologie d'annotation des entités nommées en corpus ? » In : 2009 (Cité page 48).

[FGS14] Karën Fort, Bruno Guillaume et Valentin Stern. « ZOMBILINGO : eating heads to perform dependency syntax annotation (ZOMBILINGO : manger des têtes pour annoter en syntaxe de dépendances)[in French] ». In : *Proceedings of TALN 2014 (Volume 3 : System Demonstrations)*. 2014, p. 15–16 (Cité page 44).

[FNC11] Karën Fort, Adeline Nazarenko et Ris Claire. « Corpus Linguistics for the Annotation Manager ». In : *Corpus Linguistics 2011*. Birmingham, United Kingdom, juil. 2011. URL : https://hal.archives-ouvertes.fr/hal-00641571 (Cité page 42).

[FNR12] Karën Fort, Adeline Nazarenko et Sophie Rosset. « Modeling the complexity of manual annotation tasks : a grid of analysis ». In : 2012 (Cité page 63).

[FS10] Karën Fort et Benoît Sagot. « Influence of pre-annotation on POS-tagged corpus development ». In : *Proceedings of the fourth linguistic annotation workshop*. Association for Computational Linguistics. 2010, p. 56–63 (Cité page 44).

[Fre92] Gottlob Frege. *Über Sinn und Bedeutung*. 1. Auflage. Zeitschrift für Philosophie und philosophische Kritik, Neue Folge. Leipzig : Pfeffer, 1892 (Cité pages 8, 100, 125).

[Fre71] Gottlob Frege. *Écrits logiques et philosophiques (C. Imbert, trad.)* 1971 (Cité page 12).

[Fuc94] Catherine Fuchs. *Paraphrase et énonciation*. Editions Ophrys, 1994 (Cité page 100).

[Fuc96] Catherine Fuchs. *Les ambiguïtés du français*. Ophrys Editions, 1996 (Cité pages 78–80, 83, 106).

[GCY92] William Gale, Kenneth Ward Church et David Yarowsky. « Estimating upper and lower bounds on the performance of word-sense disambiguation programs ». In : *Proceedings of the 30th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 1992, p. 249–256 (Cité page 83).

[Gal83] Michel Galmiche. « Les ambiguïtés référentielles ou les pièges de la référence ». In : *Langue française* 57 (1983), p. 60–86 (Cité pages 81, 100).

[Gal85] Michel Galmiche. « Phrases, syntagmes et articles génériques ». In : *Langages* 79 (1985), p. 2–39 (Cité page 101).

[Gar12] Laure Gardelle. « 'Anaphora', 'anaphor' and 'antecedent' in nominal anaphora : definitions and theoretical implications ». In : *Cercles : Revue Pluridisciplinaire du Monde Anglophone* 22 (2012), p. 25–40. URL : https://hal-ens-lyon.archives-ouvertes.fr/ensl-00681597 (Cité page 24).

[GM05] Claire Gardent et Hélène Manuélian. « Création d'un corpus annoté pour le traitement des descriptions définies ». In : *Traitement Automatique des Langues* 46.1 (2005), p. 115–140 (Cité page 39).

[Gar97] Roger Garside. « Discourse annotation : Anaphoric relations in corpora. » In : *Corpus Annotation-Linguistic Information from Computer Text Corpora* (1997), p. 66–84 (Cité page 46).

[Gea62] Peter Thomas Geach. « Reference and generality : An examination of some medieval and modern theories ». In : (1962) (Cité page 86).

[Gee93] Dirk Geeraerts. « Vagueness's puzzles, polysemy's vagaries ». In : *Cognitive Linguistics* (1993), p. 223–272 (Cité page 102).

[GL16] Abbas Ghaddar et Philippe Langlais. « WikiCoref : An English Coreference-annotated Corpus of Wikipedia Articles ». In : Portorož, Slovenia : European Language Resources Association (ELRA), 2016 (Cité page 137).

[Giu10] I. Giurgea. *Pronoms, déterminants et ellipse nominale : une approche minimaliste*. Romanica (Bucharest, Romania). Editura Universității din București, 2010. ISBN : 9789737378682. URL : https://books.google.fr/books?id=aBN8ygAACAAJ (Cité page 10).

[Giv83] Talmy Givón. *Topic continuity in discourse*. Amsterdam : John Benjamins, 1983 (Cité pages 22, 23, 39).

[Giv89] Talmy Givón. *Mind, code and context : Essays in pragmatics*. 1989 (Cité page 39).

[Gje08] Anje Müller Gjesdal. *Étude sémantique du pronom ON dans une perspective textuelle et contextuelle*. The University of Bergen, 2008 (Cité pages 108, 109, 115).

[GGO14] Julie Glikman, Céline Guillot-Barbance et Vanessa Obry. « Les chaînes de référence dans un corpus de textes narratifs médiévaux : traits généraux et facteurs de variation ». In : *Langages* 3 (2014), p. 43–60 (Cité page 39).

[GB17] Elisabeth Godbert et Favre Benoit. « Détection de coréférences de bout en bout en français ». In : 2017 (Cité page 28).

[GH98] Peter C Gordon et Randall Hendrick. « The representation and processing of coreference in discourse ». In : *Cognitive science* 22.4 (1998), p. 389–424 (Cité page 18).

[GR19] Mathieu Goux et Nathalie Rossi-Gensane. « Référents évolutifs, anaphores et constructions détachées : étude diachronique de recettes de cuisine ». In : *Cahiers de praxématique* 72 (2019) (Cité page 94).

[GG08] Maurice Grevisse et André Goosse. « Le bon usage (14e éd.) » In : *Bruxelles : De Boeck/Duculot* (2008) (Cité pages 108, 113).

[GS95] Ralph Grishman et Beth Sundheim. « Design of the MUC-6 evaluation ». In : *Proceedings of the 6th conference on Message understanding*. Association for Computational Linguistics. 1995, p. 1–11 (Cité page 50).

[Gro19] Loïc Grobol. « Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French ». In : *Second Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC19)*. Minneapolis, United States, juin 2019. URL : https://hal.inria.fr/hal-02151569 (Cité page 27).

[Gro20] Loïc Grobol. « Coreference resolution for spoken French ». Thèse de doct. Université de la Sorbonne Nouvelle, 2020 (Cité page 41).

[GLH18] Loïc Grobol, Frédéric Landragin et Serge Heiden. « XML-TEI-URS : using a TEI format for annotated linguistic resources ». In : *CLARIN Annual Conference 2018*. Pisa, Italy, oct. 2018. URL : https://hal.archives-ouvertes.fr/hal-01827563 (Cité page 47).

[GWJ95] Barbara J Grosz, Scott Weinstein et Aravind K Joshi. « Centering : A framework for modeling the local coherence of discourse ». In : *Computational linguistics* 21.2 (1995), p. 203–225 (Cité page 71).

[Gué14] Olivia Guérin. « Construction du référent, textualité et genre discursif : les anaphores génériques dans les séquences encyclopédiques des récits de voyage ». In : *SHS Web of Conferences*. T. 8. EDP Sciences. 2014, p. 3091–3110 (Cité page 125).

[Gué79] Jacqueline Guéron. « Relations de coréférence dans la phrase et dans le discours ». In : *Langue Française* 44 (1979), p. 42–79. ISSN : 00238368, 19577982. URL : http://www.jstor.org/stable/41557977 (Cité pages 18, 24).

[Gui79] Paul Guillaume. *La psychologie de la forme*. 1979 (Cité page 70).

[GQ19] Céline Guillot-Barbance et Matthieu Quignard. « Chaînes de référence et structure textuelle dans les Essais sur la peinture de Diderot ». In : *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics* 25 (2019) (Cité page 158).

[Hab00] Benoît Habert. « Détournements d'annotation : armer la main et le regard ». In : *Corpus. Méthodologie et applications linguistiques* 3 (2000), p. 106–120 (Cité page 43).

[Hae91] Liliane Haegeman. *21994 : Introduction to Government and Binding Theory*. 1991 (Cité page 24).

[HH76] Michael Alexander Kirkwood Halliday et Ruqaiya Hasan. *Cohesion in english*. London : Longman, 1976 (Cité page 89).

[Ham11] A Hamm. « Vers une reconnaissance du concept de saillance ». In : *Inkova O., Saillance : aspects linguistiques et communicatifs de la mise en évidence dans un texte* 1 (2011), p. 45–65 (Cité page 69).

[HMP10a] Serge Heiden, Jean-Philippe Magué et Bénédicte Pincemin. « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement ». fr. In : *Proc. of 10th International Conference on the Statistical Analysis of Textual Data*. T. 2. Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy., 2010, p. 1021–1032 (Cité pages 25, 38, 40, 54, 58, 170).

[HMP10b] Serge Heiden, Jean-Philippe Magué et Bénédicte Pincemin. « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement ». In : *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*. Sous la dir. de Sergio Bolasco, Isabella Chiari et Luca Giuliano. T. 2. 3. Rome, Italy : Edizioni Universitarie di Lettere Economia Diritto, juin 2010, p. 1021–1032. URL : https://halshs.archives-ouvertes.fr/halshs-00549779 (Cité page 47).

[Hei82] Irene Heim. « The semantics of definite and indefinite noun phrases ». In : (1982) (Cité page 27).

[HC97] Lynette Hirschman et Nancy Chinchor. *MUC-7 coreference task definition*. 1997 (Cité page 46).

[Hir88] Graeme Hirst. « Semantic interpretation and ambiguity ». In : *Artificial intelligence* 34.2 (1988), p. 131–177 (Cité page 83).

[Hob79] Jerry R Hobbs. « Coherence and coreference ». In : *Cognitive science* 3.1 (1979), p. 67–90 (Cité page 27).

[Hot02] Gilbert Hottois. « Chapitre 6. Le problème de la référence dans la philosophie du langage et de la logique ». fr. In : *Methodes en sciences humaines* 2e éd. (2002), p. 159–190. URL : https://www.cairn.info/penser-la-logique--9782804138356-page-159.htm?contenu=article (Cité page 6).

[HL19] Jiaqi Hou et Frédéric Landragin. « La saillance en français et en chinois ». French. In : *Lingvisticæ Investigationes* 42 :2 (2019), p. 186–233. URL : https://benjamins.com/catalog/li.00034.hou (Cité page 73).

[Hun02] Susan Hunston. *Corpora in applied linguistics*. Ernst Klett Sprachen, 2002 (Cité page 42).

[IV98] Nancy Ide et Jean Véronis. « Introduction to the special issue on word sense disambiguation : the state of the art ». In : *Computational linguistics* 24.1 (1998), p. 2–40 (Cité page 83).

[Jac83] Ray Jackendoff. *Semantics and cognition*. T. 8. MIT press, 1983 (Cité page 86).

[Jac02] Ray Jackendoff. *Foundations of language : Brain, meaning, grammar, evolution*. Oxford University Press, USA, 2002 (Cité page 86).

[Jes71] Otto Jespersen. *La Philosophie de la grammaire : traduit de l'anglais par Anne-Marie Leonard ; preface d'Antoine Culioli*. Editions de Minuit, 1971 (Cité page 101).

[Jon94] Kerstin Jonasson. *Le nom propre*. De Boeck Supérieur, 1994 (Cité page 13).

[KL77] Michèle Kail et Madeleine Léveillé. « Compréhension de la coréférence des pronoms personnels chez l'enfant et l'adulte ». In : *L'année psychologique* 77.1 (1977), p. 79–94 (Cité page 82).

[Kam81] Hans Kamp. « A theory of truth and semantic representation ». In : *Formal semantics-the essential readings* (1981), p. 189–222 (Cité page 27).

[Kar76] Lauri Karttunen. « Discourse Referents ». In : *Syntax and Semantics Vol. 7*. Sous la dir. de J. D. McCawley. Academic Press, 1976, p. 363–386 (Cité pages 8, 11, 27).

[Kem77] Ruth M. Kempson. *Semantic Theory*. Cambridge University Press. 1977 (Cité page 102).

[Ker83] Catherine Kerbrat-Orecchioni. « La connotation. » In : (1983) (Cité page 7).

[Ker09] Catherine Kerbrat-Orecchioni. « L'énonciation. De la subjectivité dans le langage, 1980 ». In : *Paris, Armand Colin, coll.« U, Linguistique* 4 (2009), p. 196 (Cité page 15).

[Kib11] Andrej A Kibrik. *Reference in discourse*. Oxford University Press, 2011 (Cité page 22).

[Kim04] Myong Kim. « Une description des marqueurs évidentiels on dit que et on dirait que ». In : *Travaux de linguistique* 1 (2004), p. 41–52 (Cité page 114).

[Kis95] Laurence Kister. « Accessibilité Pronominale des DéT. N1 de (DéT.) N2 : le Rôle de la Détermination ». In : *Lingvisticae Investigationes* 19.1 (1995), p. 107–121 (Cité page 103).

[Kle78] Georges Kleiber. « Phrases et valeurs de vérité ». In : *Bulletin des Jeunes Romanistes Strasbourg* (1978) (Cité page 101).

[Kle79] Georges Kleiber. « A propos de l'ambiguïté référentielle : transparence/opacité ». In : *Travaux de Linguistique et de Littérature Strasbourg* 17.1 (1979), p. 233–250 (Cité page 100).

[Kle81] Georges Kleiber. *Problèmes de référence : descriptions définies et noms propres*. Paris : Klincksieck, 1981 (Cité page 13).

[Kle86] Georges Kleiber. « Déictiques, embrayeurs," token-réflexives", symboles indexicaux, etc. : comment les définir ? » In : *L'Information grammaticale* 30.1 (1986), p. 3–22 (Cité page 15).

[Kle88] Georges Kleiber. « Peut-on définir une catégorie générale de l'anaphore ? » In : *Vox Romanica* 47 (1988), p. 1 (Cité page 18).

[Kle91] Georges Kleiber. « Anaphore-deixis : où en sommes-nous ? » fre. In : (1991). ISSN : 0222-9838. DOI : 10.3406/igram.1991.3231. URL : https://www.persee.fr/doc/igram_0222-9838_1991_num_51_1_3231 (Cité pages 17, 34).

[Kle93] Georges Kleiber. « Anaphore associative, pontage et stéréotypie ». In : *Lingvisticae Investigationes* 17.1 (1993), p. 35–82 (Cité page 24).

[Kle94] Georges Kleiber. *Anaphores et pronoms*. Louvain-La-Neuve : Duculot, 1994 (Cité pages 34, 72).

[Kle99] Georges Kleiber. « Problèmes de sémantique. La polysémie en questions ». In : *Presses Universitaires du Septentrion. Villeneuve d'Ascq* (1999) (Cité page 7).

[Kle01] Georges Kleiber. *L'anaphore associative*. fr. Paris : Puf, 2001 (Cité pages 24, 33).

[Kle04] Georges Kleiber. « Peut-on sauver un sens de dénomination pour les noms propres ? » In : *Functions of language* 11.1 (2004), p. 115–145 (Cité page 13).

[KJR82] S. A. Kripke, P. Jacob et F. Recanati. *La logique des noms propres*. 1982 (Cité page 12).

[Lak70] George Lakoff. « A Note on Vagueness and Ambiguity ». In : *Linguistic Inquiry* (1970), p. 357–359 (Cité page 102).

[Lak73] George Lakoff. « Hedges : A Study in Meaning Criteria and the Logic of Fuzzy Concepts ». In : *Journal of Philosophical Logic* 2.4 (1973), p. 458–508. ISSN : 0022-3611. URL : https://www.jstor.org/stable/30226076 (Cité page 102).

[Lam96] Knud Lambrecht. *Information structure and sentence form : Topic, focus, and the mental representations of discourse referents*. T. 71. Cambridge university press, 1996 (Cité page 70).

[Lan85] Ronald Landheer. « L'ambiguïté-Un pommier de discorde dans le verger linguistique. Contribution à une mise au point ». In : *Neophilologus* 69.4 (1985), p. 501 (Cité page 79).

[Lan04a] Frédéric Landragin. « L'utilisation de scores numériques en sémantique computationnelle ». In : *Journées Scientifiques de Sémantique et Modélisation (JSM'04)*. Lyon, France, 2004. URL : https://halshs.archives-ouvertes.fr/halshs-00137045 (Cité page 71).

[Lan04b] Frédéric Landragin. « Saillance physique et saillance cognitive ». In : *Corela. Cognition, représentation, langage* 2-2 (2004) (Cité pages 69, 70, 73).

[Lan05] Frédéric Landragin. « Traitement automatique de la saillance ». In : LIMSI, 2005, p. 263–272. URL : https://halshs.archives-ouvertes.fr/halshs-00137693 (Cité pages 69, 73).

[Lan07] Frédéric Landragin. « L'anaphore à antécédent flou : une caractérisation et ses conséquences sur l'annotation des relations anaphoriques ». In : 2007 (Cité pages 84, 102–104, 131, 167).

[Lan11a] Frédéric Landragin. *De la saillance visuelle à la saillance linguistique*. 2011 (Cité page 70).

[Lan11b] Frédéric Landragin. « Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits ». In : *Corpus* 10 (2011), p. 61–80 (Cité pages 35, 124, 125).

[Lan14] Frédéric Landragin. *Anaphores et coréférences : analyse assistée par ordinateur*. Oct. 2014 (Cité page 104).

[Lan17] Frédéric Landragin. « Analyse, visualisation et identification automatique des chaînes de coréférences : des questions interdépendantes ? » In : *Langue française* 3 (2017), p. 17–34 (Cité page 63).

[Lan18] Frédéric Landragin. « Étude de la référence et de la coréférence : rôle des petits corpus et observations à partir du corpus MC4 ». In : *Corpus* 18 (2018) (Cité page 97).

[Lan19] Frédéric Landragin. *Corpus Democrat*. Licence Creative Commons - Attribution - Partage dans les Mêmes Conditions 4.0 International. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr. 2019. URL : https://hdl.handle.net/11403/democrat/v1 (Cité page 39).

[LO18] Frédéric Landragin et Bruno Oberle. « Identification automatique de chaînes de coréférences : vers une analyse des erreurs pour mieux cibler l'apprentissage ». In : 2018 (Cité page 28).

[LPV12] Frédéric Landragin, Thierry Poibeau et Bernard Victorri. « ANALEC : a New Tool for the Dynamic Annotation of Textual Data. European Language Resources Association (ELRA) ». In : *International Conference on Language Resources and Evaluation*. LREC 2012. Istanbul, Turkey, 2012, p. 357–362 (Cité pages 40, 54, 58, 170).

[LPB17] Frédéric Landragin, Juliette Potier et Meryl Bothua. « Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus ». In : 2017 (Cité pages 45, 48, 50, 55).

[LT14] Frédéric Landragin et Noalig Tanguy. « Référence et coréférence du pronom indéfini on ». fr. In : *Langages* 195.3 (2014), p. 99. ISSN : 0458-726X, 1958-9549. DOI : 10.3917/lang.195.0099. URL : http://www.cairn.info/revue-langages-2014-3-page-99.htm (visité le 05/10/2017) (Cité pages 108, 113).

[LTC15] Frédéric Landragin, Noalig Tanguy et Michel Charolles. « Références aux personnages dans L'Occupation des sols : apport de la linguistique outillée ». In : *Revue Sciences/Lettres* 3 (2015) (Cité page 64).

[LL94] Shalom Lappin et Herbert J Leass. « An algorithm for pronominal anaphora resolution ». In : *Computational linguistics* 20.4 (1994), p. 535–561 (Cité pages 71, 72).

[Le 81] Pierre Le Goffic. « Ambiguïté linguistique et activité de langage : contribution à une étude historique et critique des conceptions sur l'ambiguïté du langage, et à l'élaboration d'une théorie linguistique de l'ambiguïté, avec application au français ». Thèse de doct. Centre de documentation sciences humaines, 1981 (Cité pages 79, 80).

[Lee+17] Kenton Lee, Luheng He, Mike Lewis et al. « End-to-end neural coreference resolution ». In : *arXiv preprint arXiv :1707.07045* (2017) (Cité page 41).

[LHZ18] Kenton Lee, Luheng He et Luke Zettlemoyer. « Higher-order coreference resolution with coarse-to-fine inference ». In : *arXiv preprint arXiv :1804.05392* (2018) (Cité page 27).

[Lee05] Geoffrey Leech. *Adding Linguistic Annotation in Developing Linguistic Corpora : a Guide to Good Practice, ed. M. Wynne*. 2005 (Cité pages 42, 43).

[LK63] Robert B Lees et Edward S Klima. « Rules for English pronominalization ». In : *Language* 39.1 (1963), p. 17–28 (Cité page 24).

[Ler04] Sarah Leroy. *Le nom propre en français*. Editions OPHRYS, 2004 (Cité page 12).

[Les86] Michael Lesk. « Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone ». In : *Proceedings of the 5th annual international conference on Systems documentation*. 1986, p. 24–26 (Cité page 83).

[LDM12] Hector Levesque, Ernest Davis et Leora Morgenstern. « The winograd schema challenge ». In : *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. 2012 (Cité pages 83, 84).

[Lev87] Stephen C. Levinson. « Pragmatics and the grammar of anaphora : a partial pragmatic reduction of Binding and Control phenomena ». In : *Journal of Linguistics* 23.2 (1987), p. 379–434. doi : 10.1017/S0022226700011324 (Cité page 35).

[LG18] Hongyin Luo et Jim Glass. « Learning word representations with cross-sentence dependency for end-to-end co-reference resolution ». In : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, p. 4829–4833 (Cité page 27).

[Luo+04] Xiaoqiang Luo et al. « A mention-synchronous coreference resolution algorithm based on the bell tree ». In : *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 2004, p. 135 (Cité page 27).

[Lus81] Barbara Lust. « Constraints on anaphora in child language : A prediction for a universal ». In : *Language acquisition and linguistic theory* (1981), p. 74–96 (Cité page 22).

[Mal76] Bertil Malmberg. « Phonétique française ». In : (1976) (Cité page 79).

[Mar13] Alda Mari. « Intensional epistemic wholes : A study in the ontology of collectivity ». In : *The compositionality of Meaning and Content* 1 (2013), p. 189–212 (Cité page 107).

[Mar67] André Martinet. « CONNOTATIONS, POÉSIE ET CULTURE ». In : *To honor Roman Jakobson : essays on the occasion of his 70. birthday, 11. October 1966Vol. 2*. Reprint 2018. T. 32. Berlin, Boston : De Gruyter Mouton, 1967. isbn : 978-3-11-134912-1. doi : 10.1515/9783111349121-044 (Cité page 7).

[MW16] Yann Mathet et Antoine Widlöcher. « Évaluation des annotations : ses principes et ses pièges ». fr. In : *TAL* 57.2 (2016), p. 26 (Cité page 44).

[ML95] Joseph F McCarthy et Wendy G Lehnert. « Using decision trees for coreference resolution ». In : *arXiv preprint cmp-lg/9505043* (1995) (Cité page 27).

[McC93] James D. McCawley. *Everything that linguists have always wanted to know about logic... but were ashamed to ask*. University of Chicago Press. 1993 (Cité page 102).

[Men+00] Andreas Mengel et al. « MATE dialogue annotation guidelines ». In : *MATE Deliverable D* 2 (2000), p. 1 (Cité page 46).

[Mil89a] John Stuart Mill. *Système de logique déductive et inductive*. Paris : Félix Alcan, 1889 (Cité pages 7, 12).

[Mil76] Jean-Claude MILNER. « Réflexion sur la référence ». In : *Langue Française* 30 (1976), p. 63–73 (Cité pages 7, 16, 18–20, 33).

[Mil82] Jean-Claude MILNER. *Ordres et raisons de langue*. Collection linguistique. Seuil, 1982. ISBN : 9782020060820 (Cité pages 16, 115).

[Mil89b] Jean-Claude MILNER. *Introduction à une science du langage, éd. du Seuil*. 1989 (Cité page 74).

[Mil07] Eleni MILTSAKAKI. « A rethink of the relationship between salience and anaphora resolution ». In : *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium*. Lagos, Portugal, 2007, p. 91–96 (Cité page 103).

[Mit96] Ruslan MITKOV. « Anaphora and machine translation ». In : *Machine Translation Review* 4 (1996), p. 6–16 (Cité page 28).

[Mit99] Ruslan MITKOV. *Anaphora resolution : the state of the art*. Citeseer, 1999 (Cité pages 28, 71).

[MD95] Lorenza MONDADA et Danièle DUBOIS. « Construction des objets de discours et catégorisation : une approche des processus de référenciation ». In : *Revue Tranel (Travaux neuchâtelois de linguistique)* 23 (1995), p. 273–302 (Cité page 92).

[MUC95] MUC-6. « Appendix D : Coreference Task Definition (v2.3) ». In : *Sixth Message Understanding Conference (MUC-6) : Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. 1995. URL : https://www.aclweb.org/anthology/M95-1025 (Cité pages 26, 28).

[Mul70] Charles MULLER. *Sur les emplois personnels de l'indéfini "on"*. Société de linguistique romane, 1970 (Cité page 109).

[MS06] Christoph MÜLLER et Michael STRUBE. « Multi-level annotation of linguistic data with MMAX2 ». In : *Corpus Technology and Language Pedagogy : New Resources, New Tools, New Methods*. 2006 (Cité pages 40, 137).

[Muz+13a] Judith MUZERELLE, Anaïs LEFEUVRE, Jean-Yves ANTOINE et al. « ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement ». In : *20e conférence sur le Traitement Automatique des Langues Naturelles*. Sous la dir. d'ATALA. Les Sable d'Olonne, France : ATALA, 2013, p. 555–563 (Cité pages 39, 45, 47, 138).

[Muz+14] Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang et al. « ANCOR_Centre, a Large Free Spoken French Coreference Corpus : description of the Resource and Reliability Measures ». In : *LREC'2014, 9th Language Resources and Evaluation Conference*. Sous la dir. d'ELRA. Reyjavik, Iceland, mai 2014, p. 843–847. URL : https://hal.archives-ouvertes.fr/hal-01075679 (Cité page 39).

[Muz+12] Judith Muzerelle, Emmanuel Schang, Jean-Yves Antoine, Iris Eshkol, Denis Maurel, Aurore Boyer et al. « Annotations en chaînes de coréférences et anaphores dans un corpus de discours spontané en français ». In : *SHS Web of Conferences*. T. 1. EDP Sciences. 2012, p. 2497–2516 (Cité pages 28, 52).

[Muz+13b] Judith Muzerelle, Emmanuel Schang, Jean-Yves Antoine, Iris Eshkol, Denis Maurel, Aurore Boyer-Pelletier et al. « Annotation en relations anaphoriques d'un corpus de discours oral spontané en français ». In : *Congrès Mondial de Linguistique Française, CMLF'2012*. Lyon, France, juil. 2013, 15 pp. URL : https://hal.archives-ouvertes.fr/hal-00788164 (Cité page 172).

[Ned+16] Anna Nedoluzhko et al. « Coreference in prague czech-english dependency treebank ». In : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, p. 169–176 (Cité page 47).

[Ng10] Vincent Ng. « Supervised noun phrase coreference research : The first fifteen years ». In : *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics. 2010, p. 1396–1411 (Cité page 27).

[Nic+02] Pascale Nicolas et al. « Towards a large corpus of spoken dialogue in French that will be freely available : the" Parole Publique" project and its first realisations. » In : *LREC*. 2002 (Cité page 39).

[Niz19] Małgorzata Niziołek. « Créer un flou référentiel : l'analyse de la structure pronom indéfini on+ V et de ses équivalents polonais sur l'exemple des textes fantastiques ». In : *Studia Romanica Posnaniensia* 46.1 (2019), p. 153–166 (Cité pages 114, 119).

[Noa96] Michèle Noailly. « Le vide des choses ». fr. In : *Cahiers de praxématique* 27 (jan. 1996), p. 73–90. ISSN : 0765-4944. URL : http://journals.openedition.org/praxematique/2999 (visité le 22/01/2020) (Cité page 15).

[Noa97] Michèle Noailly. « Les mystères de la transitivité invisible ». fre. In : (1997). ISSN : 0458-726X. DOI : 10.3406/lgge.1997.2127. URL : https://www.persee.fr/doc/lgge_0458-726x_1997_num_31_127_2127 (Cité page 35).

[NER15] Damien Nouvel, Maud Ehrmann et Sophie Rosset. *Les entités nommées pour le traitement automatique des langues*. ISTE Group, 2015 (Cité pages 12, 26).

[Obe18] Bruno Oberle. « SACR : A Drag-and-Drop Based Tool for Coreference Annotation ». Anglais. In : *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan : European Language Resources Association (ELRA), mai 2018. isbn : 979-10-95546-00-9 (Cité pages 40, 54).

[Obr+17] Vanessa Obry et al. « Les chaînes de référence dans les récits brefs en français : étude diachronique (xiiie-xvie s.) » In : *Langue française* 3 (2017), p. 91–110 (Cité page 64).

[Øde06] Annelise Ødegaard. « On multiréférentiel : une étude contrastive des valeurs du pronom ON et leurs équivalences norvégiennes ». Mém.de mast. 2006 (Cité pages 114, 132).

[OR23] Kay Odgen Charles et Armstrong Richards Ivor. *The Meaning of Meaning. A study of the influence of language upon throught and of the science of symbolism.* New-York : Harcourt, 1923 (Cité page 6).

[Ogr+14] Maciej Ogrodniczuk, Katarzyna Glowinska et al. *Coreference : Annotation, Resolution and Evaluation in Polish*. en. Walter de Gruyter GmbH & Co KG, 2014. isbn : 978-1-61451-838-9 (Cité pages 47, 137).

[Ogr+13] Maciej Ogrodniczuk, Katarzyna Głowińska et al. « Polish coreference corpus ». In : *Language and Technology Conference*. Springer. 2013, p. 215–226 (Cité page 87).

[ON16] Maciej Ogrodniczuk et Vincent Ng, éds. *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*. San Diego, California : Association for Computational Linguistics, juin 2016. doi : 10.18653/v1/W16-07. url : https://www.aclweb.org/anthology/W16-0700 (Cité page 39).

[ON17] Maciej Ogrodniczuk et Vincent Ng, éds. *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*. Valencia, Spain : Association for Computational Linguistics, avr. 2017. doi : 10.18653/v1/W17-15. url : https://www.aclweb.org/anthology/W17-1500 (Cité page 39).

[Pas97] Rebecca Passonneau. « Instructions for applying discourse reference annotation for multiple applications (DRAMA) ». In : *Unpublished manuscript* (1997) (Cité page 46).

[Pat92] Thiyagarajasarma PATTABHIRAMAN. « Aspects of salience in natural language generation ». Thèse de doct. Theses (School of Computing Science)/Simon Fraser University, 1992 (Cité page 71).

[PC90] Thiyagarajasarma PATTABHIRAMAN et Nick CERCONE. « Selection : Salience, relevance and the coupling between domain-level tasks and text planning ». In : *Proceedings of the Fifth International Workshop on Natural Language Generation*. 1990 (Cité page 70).

[Pec05] Anthony PECQUEUX. « Un témoignage adressé. Parole du rap et parole collective ». In : *Les cahiers de psychologie politique* 7 (2005), ht. URL : https://hal.archives-ouvertes.fr/hal-00349605 (Cité page 130).

[Pei11] C.S. PEIRCE. « Vagueness ». In : *Dictionary of philosophy and psychology*. Macmillan. T. 2. New-York : J.M. Baldwin, 1911 (Cité page 102).

[Per00] Michèle PERRET. « Quelques remarques sur l'anaphore nominale aux XIVe et XVe siècles ». In : *L'Information grammaticale* 87.1 (2000), p. 17–23 (Cité pages 23, 65).

[Pér+11] Marie-Paule PÉRY-WOODLEY et al. « La ressource ANNODIS, un corpus enrichi d'annotations discursives ». In : (2011) (Cité page 151).

[PDB04] Jean-Marie PIERREL, Jacques DENDIEN et Pascale BERNARD. « Le TLFi ou Trésor de la Langue Française informatisé ». In : *Proceedings of the 11th EURALEX International Congress*. Sous la dir. de Geoffrey WILLIAMS et Sandra VESSIER. Lorient, France : Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, juil. 2004, p. 165–170. ISBN : 29-52245-70-3 (Cité pages 9, 99, 108).

[PM14] Bénédicte PINCEMIN et Denise MALRIEU. « Caractérisation quantitative de textes. Application à l'oral représenté, en diachronie ». In : 2014 (Cité page 65).

[Poe00] Massimo POESIO. *Coreference. MATE Dialogue Annotation Guidelines-Deliverable 2.1, January 2000, 126-182*. 2000 (Cité page 46).

[Poe04] Massimo POESIO. « The MATE/GNOME proposals for anaphoric annotation, revisited ». In : *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*. 2004, p. 154–162 (Cité page 47).

[PA08] Massimo POESIO et Ron ARTSTEIN. « Anaphoric Annotation in the ARRAU Corpus. » In : *Proc. of LREC*. Marrakesh, 2008 (Cité pages 47, 137).

[PBR99] Massimo POESIO, Florence BRUNESEAUX et Laurent ROMARY. « The MATE meta-scheme for coreference in dialogues in multiple languages ». In : *Towards Standards and Tools for Discourse Tagging*. 1999 (Cité page 46).

[PRS08] Massimo Poesio, Uwe Reyle et Rosemary Stevenson. « Justified sloppiness in anaphoric reference ». In : *Computing meaning*. Springer, 2008, p. 11–31 (Cité page 167).

[PSV16] Massimo Poesio, Roland Stuckardt et Yannick Versley, éds. *Anaphora resolution : algorithms, resources, and applications*. eng. Theory and applications of natural language processing. Berlin Heidelberg : Springer, 2016 (Cité pages 18, 26, 46).

[Poe+06] Massimo Poesio, Patrick Sturt et al. « Underspecification and anaphora : Theoretical issues and preliminary evidence ». In : *Discourse processes* 42.2 (2006), p. 157–175 (Cité page 167).

[Pop99] Andrei Popescu Belis. « Modelisation multi-agents des echanges langagiers : application au probleme de la reference et a son evaluation ». Thèse de doct. Paris 11, 1999 (Cité page 38).

[Pot92] Bernard Pottier. « Sémantique générale ». In : (1992) (Cité page 70).

[PL17] Céline Poudat et Frédéric Landragin. *Explorer un corpus textuel : Méthodes-pratiques-outils*. De Boeck Superieur, 2017 (Cité pages 43, 63).

[Pra+11] Sameer Pradhan et al. « CoNLL-2011 Shared Task : Modeling Unrestricted Coreference in OntoNotes ». en. In : *Proceedings of the Fifteenth Conference on Computational Natural Language Learning : Shared Task*. Portland, Oregon, USA, 2011, p. 1–27 (Cité pages 47, 50, 137).

[Pra87] Michele Prandi. *Sémantique du contresens : essai sur la forme interne du contenu des phrases*. Les Editions de Minuit, 1987 (Cité page 131).

[Pri81] Ellen F. Prince. « Toward a taxonomy of given-new information ». In : *Syntax and semantics : Vol. 14. Radical Pragmatics*. Sous la dir. de P. Cole. New York : Academic Press, 1981, p. 223–255 (Cité pages 8, 27).

[PP04] Amruta Purandare et Ted Pedersen. « Word sense discrimination by clustering contexts in vector and similarity spaces ». In : *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*. 2004, p. 41–48 (Cité page 83).

[Pus98] James Pustejovsky. *The generative lexicon*. MIT press, 1998 (Cité page 57).

[Qui+18] Matthieu Quignard et al. « Textometric Exploitation of Coreference-annotated Corpora with TXM : Methodological Choices and First Outcomes ». In : 2018 (Cité page 28).

[Qui77] Willard Van Orman Quine. *Le mot et la chose*. Français. Trad. Joseph Dopp et Paul Gochet. Flammarion, 1977 (Cité pages 100, 102, 129).

[Rab98] Alain Rabatel. *La construction textuelle du point de vue*. FeniXX, 1998 (Cité page 109).

[Rab01] Alain Rabatel. « La valeur de on pronom indéfini/pronom personnel dans les perceptions représentées ». In : (2001) (Cité page 109).

[Ras05] François Rastier. « Enjeux épistémologiques de la linguistique de corpus ». In : *La linguistique de corpus* (2005), p. 31–45 (Cité page 43).

[Reb97a] Anne Reboul. « Combien y-a-t-il de poulets ici, les référents évolutifs, identité et désignation ». In : (1997) (Cité page 91).

[Reb97b] Anne Reboul. « What (if anything) is accessibility ? A relevance-oriented criticism of Ariel's Accessibility Theory of referring expressions ». In : *Discourse and pragmatics in functional grammar* 18 (1997), p. 91 (Cité page 72).

[Rec10] Marta Recasens. « Coreference : Theory, Resolution, Annotation and Evaluation ». Thèse de doct. University of Barcelona, 2010 (Cité pages 27, 85, 168).

[RDP13] Marta Recasens, Marie-Catherine De Marneffe et Christopher Potts. « The Life and Death of Discourse Entities : Identifying Singleton Mentions ». In : *Proceedings of NAACL-HLT 2013*. 2013, p. 627–633 (Cité pages 27, 87).

[RHM10] Marta Recasens, Eduard Hovy et Maria Antonia Marti. « A Typology of Near-Identity Relations for Coreference (NIDENT). » In : *LREC*. 2010 (Cité pages 85–87, 137).

[RHM11] Marta Recasens, Eduard Hovy et Maria Antonia Marti. « Identity, non-identity, and near-identity : Addressing the complexity of coreference ». In : *Lingua* 121.6 (2011), p. 1138–1152 (Cité page 85).

[RTM14] Marta Recasens, Liliana Tolchinsky et Maria Antonia Marti. « Coreference is not always either/or : psycholinguistic evidence for near-identity ». In : *Language, Cognition and Neuroscience* 29.7 (2014), p. 844–855. doi : 10.1080/01690965.2013.801503. eprint : https://doi.org/10.1080/01690965.2013.801503. url : https://doi.org/10.1080/01690965.2013.801503 (Cité page 132).

[Rei88] Marie-José Reichler-Béguelin. « Anaphore, cataphore et mémoire discursive ». In : *Pratiques* 57.1 (1988), p. 15–43 (Cité page 16).

[RPR04] M. Riegel, J.-C. Pellat et R. Rioul. *Grammaire méthodique du français*. Français. 3e éd. Quadrige. Paris : Presses Universitaires de France - PUF, 2004 (Cité pages 108, 123).

[Rob97] Craige Roberts. « Anaphora in intensional contexts ». In : (1997) (Cité page 27).

[Rod+10] Kepa Joseba Rodriguez et al. « Anaphoric annotation of wikipedia and blogs in the live memories corpus ». In : *Proceedings of LREC*. 2010, p. 157–163 (Cité page 47).

[Rou18] Magali Rouquier. *Constructions clivées*. 2018. URL : http://encyclogram.fr/notx/018/018_Notice.php (visité le 16/06/2020) (Cité page 74).

[Rus05] Bertrand Russell. « On denoting ». In : *Mind* 14.56 (1905), p. 479–493 (Cité page 12).

[Sai91] Richard Mark Sainsbury. *Concepts without boundaries*. King's College, Department of Philosophy, 1991 (Cité page 102).

[SF09] André Salem et Serge Fleury. « Explorations textométriques ». In : *Paris : Université de la Sorbonne Nouvelle Paris* 3 (2009) (Cité page 63).

[Sal01] Susanne Salmon-Alt. « Entre corpus et théorie : l'annotation (co) référentielle ». In : (2001) (Cité page 38).

[Sal02] Susanne Salmon-Alt. « Le projet ananas : Annotation anaphorique pour l'analyse sémantique de corpus ». In : *TALN 2002, Nancy* (2002), p. 163–172 (Cité page 38).

[San65] Kristian Sandfeld. *Syntaxe du français contemporain : Les pronoms*. T. 1. Paris : Champion, 1965 (Cité page 108).

[Sau+22] Ferdinand de Saussure et al. *Cours de linguistique générale*. French. Paris : Payot, 1922 (Cité page 6).

[SAL17] Emmanuel Schang, Jean-Yves Antoine et Anaïs Lefeuvre-Halftermeyer. « Les chaînes coréférentielles en créole de la Guadeloupe ». In : 2017 (Cité page 28).

[Sch+11] Emmanuel Schang, Aurore Boyer-Pelletier et al. « Coreference and anaphoric annotations for spontaneous speech corpos in French ». In : *DAARC'2011, 8th Discourse Anaphora and Anaphor Resolu1on Colloquium*. Faro, Portugal, oct. 2011, 9 pp. URL : https://hal.archives-ouvertes.fr/hal-00831414 (Cité page 39).

[Sch07] Hans-Jörg Schmid. « Entrenchment, salience, and basic levels ». In : *The Oxford handbook of cognitive linguistics* 117138 (2007) (Cité page 69).

[Sch94] Helmut Schmid. « Probabilistic Part-of-Speech Tagging Using Decision Trees ». In : *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, 1994 (Cité pages 25, 112).

[Sch92] Catherine SCHNEDECKER. « Référence et discours : chaînes de référence et redénomination (essai sur l'emploi en seconde mention du nom propre) ». thèse. Strasbourg 2, 1992 (Cité pages 28, 30).

[Sch11] Catherine SCHNEDECKER. « La notion de "saillance" : problèmes définitoires et avatars ». In : *SAILLANCE* (2011), p. 21 (Cité page 69).

[Sch14] Catherine SCHNEDECKER. « Chaînes de référence et variations selon le genre ». In : *Langages* 3 (2014), p. 23–42 (Cité pages 25, 39).

[Sch17] Catherine SCHNEDECKER. « Les chaînes de référence : une configuration d'indices pour distinguer et identifier les genres textuels ». In : *Langue française* 3 (2017), p. 53–72 (Cité page 65).

[Sch19] Catherine SCHNEDECKER. « De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence ». In : *Les cahiers de praxématique* (2019). URL : https://hal.archives-ouvertes.fr/hal-02317889 (Cité pages 20, 22, 25, 28).

[SC93a] Catherine SCHNEDECKER et Michel CHAROLLES. « Coréférence et identité. Le problème des référents évolutifs ». In : *Langages* (1993), p. 106–126 (Cité pages 90, 91).

[SC93b] Catherine SCHNEDECKER et Michel CHAROLLES. « Les référents évolutifs : points de vue ontologique et phénoménologique ». In : *Cahiers de linguistique française* 14 (1993), p. 197–227 (Cité page 91).

[SGL17] Catherine SCHNEDECKER, Julie GLIKMAN et Frédéric LANDRAGIN. « Les chaînes de référence : annotation, application et questions théoriques ». In : *Langue française* 3 (2017), p. 5–16 (Cité page 63).

[SL14] Catherine SCHNEDECKER et Frédéric LANDRAGIN. « Les chaînes de référence : présentation ». fr. In : *Langages* 195 (2014), p. 3–22. ISSN : 0458-726X. DOI : 10.3917/lang.195.0003 (Cité pages 25, 64).

[Sea72] John R SEARLE. *Les Actes de Langage. Essai de Philosophie du Langage.* Paris : Hermann, 1972 (Cité page 12).

[SG01] Andrea SETZER et Robert GAIZAUSKAS. « A pilot study on annotating temporal relations in text ». In : *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*. 2001 (Cité page 163).

[Sfa08] Inès SFAR. « Traduire les blagues : jouer par/avec les mots ». fre. In : (2008). ISSN : 0751-9532. DOI : 10.3406/equiv.2008.1432. URL : https://www.persee.fr/doc/equiv_0751-9532_2008_num_35_1_1432 (Cité page 79).

[Sid79] Candace Lee Sidner. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse.* Rapp. tech. Massachusetts Inst of Tech Cambridge Artificial Intelligence lab, 1979 (Cité page 70).

[Sin04] John Sinclair. « Trust the Text : Language, Corpus and Discourse, Ronald Carter (Ed ». In : (2004) (Cité page 42).

[SB00] Barry Smith et Berit Brogaard. « A unified theory of truth and reference ». In : *Logique et Analyse* (2000), p. 49–93 (Cité page 167).

[Ste02] Rosemary Stevenson. « The role of salience in the production of referring expressions : A psycholinguistic perspective ». In : *Information sharing : Reference and presupposition in language generation and interpretation* (2002), p. 167–192 (Cité page 71).

[Str00] Christopher Strachey. « Fundamental concepts in programming languages ». In : *Higher-order and symbolic computation* 13.1-2 (2000), p. 11–49 (Cité page 100).

[Str50] Frederick Peter Strawson. « On referring ». In : *Mind* 10.235 (juil. 1950) (Cité page 7).

[Tal00] Leonard Talmy. *Toward a cognitive semantics.* T. 2. MIT press, 2000 (Cité page 70).

[TMR08] Mariona Taulé, Maria Antonia Marti et Marta Recasens. « AnCora : Multilevel Annotated Corpora for Catalan and Spanish. » In : *Lrec.* 2008 (Cité page 47).

[TDC12] Isabelle Tellier, Yoann Dupont et Arnaud Courmet. « Un segmenteur-étiqueteur et un chunker pour le français (A Segmenter-POS Labeller and a Chunker for French)[in French] ». In : *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 5 : Software Demonstrations.* 2012, p. 7–8 (Cité pages 48, 55).

[Tes59] Lucien Tesnière. « Eléments de syntaxe structurale ». In : *Klincksieck, Paris* (1959) (Cité page 16).

[Tod01] Laurence Longo Amalia Todirascu. « RefGen, outil d'identification automatique des chaînes de référence en français ». In : *Traitement Automatique des Langues Naturelles* (2001), p. 329 (Cité page 48).

[Tou18] Julien Tourille. « Extracting Clinical Event Timelines : Temporal Information Extraction and Coreference Resolution in Electronic Health Records ». Thèse de doct. Université Paris-Saclay, 2018 (Cité page 27).

[Tro01] François Trouilleux. « Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français ». Thèse de doct. 2001 (Cité page 38).

[Tse08] Chiaoi Tseng. « Coherence and cohesive harmony in filmic text ». In : *Multimodal semiotics : Functional analysis in contexts of education* (2008), p. 87–104 (Cité page 28).

[Tur16] Nicolas Turenne. *Analyse de données textuelles sous R*. ISTE Group, 2016 (Cité page 63).

[Tut02] Agnès Tutin. « A corpus-based study of pronominal anaphoric expressions in French ». In : *corpus* 24 (2002), p. 516 (Cité page 25).

[Tut+00] Agnès Tutin et al. « Annotating a large corpus with anaphoric links ». In : 2000 (Cité pages 38, 39).

[VK99] Kees Van Deemter et Rodger Kibble. « What is coreference, and what should coreference annotation be ? » In : *Coreference and Its Applications*. 1999 (Cité page 46).

[Ven19] Zeno Vendler. *Linguistics in philosophy*. Cornell University Press, 2019 (Cité page 28).

[Ver08] Yannick Versley. « Vagueness and referential ambiguity in a large-scale annotated corpus ». In : *Research on Language and Computation* 6.3-4 (2008), p. 333–353 (Cité page 167).

[Wal98] Joshi Prince Walker. *Centering theory in discourse*. Oxford University Press, 1998 (Cité page 27).

[WMA06] Linton Wang, Eric McCready et Nicholas Asher. « Information dependency in quantificational subordination ». In : *Where semantics meets pragmatics* (2006), p. 267–306 (Cité page 27).

[Web78] Bonnie Lynn Webber. « Description Formation and Discourse Model Synthesis ». In : *Theoretical Issues in Natural Language Processing-2*. 1978. url : https://www.aclweb.org/anthology/T78-1006 (Cité page 8).

[WR10] Alfred North Whitehead et Bertrand Russell. *Principia mathematica*. T. 1. Cambridge University Press Cambridge, UK, 1910 (Cité page 100).

[Whi09] Paul Whitla. « Crowdsourcing and its application in marketing activities ». In : *Contemporary Management Research* 5.1 (2009) (Cité page 44).

[WM09] Antoine Widlöcher et Yann Mathet. « La plate-forme Glozz : environnement d'annotation et d'exploration de corpus ». fr. In : *Actes de la 16e Conférence Traitement Automatique des Langues Naturelle*. T. session posters. Senlis, France, 2009, p. 10 (Cité pages 40, 170).

[Wil+20] Rodrigo WILKENS et al. « French coreference for spoken and written language ». In : *Language Resources and Evaluation Conference (LREC 2020)*. Proceedings of the 12th Edition of the Language Resources and Evaluation Conference (LREC 2020). Marseille, France, 2020, p. 80–89. URL : https://hal.archives-ouvertes.fr/hal-02476902 (Cité pages 27, 41).

[WS99] Deirdre WILSON et Dan SPERBER. « Relevance and relevance theory ». In : *MIT Encyclopedia of the Cognitive Sciences*. Citeseer. 1999 (Cité pages 71, 72).

[Win72] Terry WINOGRAD. « Understanding natural language ». In : *Cognitive psychology* 3.1 (1972), p. 1–191 (Cité pages 70, 83).

[Wis+15] Sam Joshua WISEMAN et al. « Learning anaphoricity and antecedent ranking features for coreference resolution ». In : *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*. Association for Computational Linguistics. 2015 (Cité page 27).

[WB01] Maria WOLTERS et David BEAVER. « What does he mean ? » In : *Proceedings of the Annual Meeting of the Cognitive Science Society*. T. 23. 23. 2001 (Cité page 71).

[Yan+04] Xiaofeng YANGY et al. « An NP-cluster based approach to coreference resolution ». In : *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics. 2004, p. 226 (Cité page 27).

[Yap13] Manuela YAPOMO. « Construction de corpus multilingues : état de l'art ». In : *Proceedings of RECITAL 2013* (2013), p. 56–68 (Cité page 25).

[Yar92] David YAROWSKY. « Word-sense disambiguation using statistical models of Roget's categories trained on large corpora ». In : *Proceedings of the 14th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics. 1992, p. 454–460 (Cité page 83).

[Yul82] George YULE. « Interpreting anaphora without identifying reference ». In : *Journal of Semantics* 1 (3-4 1982). DOI : 10.1093/jos/1.3-4.315 (Cité page 90).

[Zad65] Lotfi Asker ZADEH. « Fuzzy sets ». In : *Information and Control* 8.3 (juin 1965), p. 338–353 (Cité pages 102, 167).

[Zha98] Grace ZHANG. « Fuzziness — Vagueness — Generality — Ambiguity ». In : *Journal of Pragmatics* 29 (jan. 1998), p. 13–31. DOI : 10.1016/S0378-2166(97)00014-3 (Cité page 101).

Index

Annexes

0.1 Sous-corpus de Democrat étudié

Titre	Auteur	Siècle	Type textuel	Genre textuel
Aden Arabie	P. Nizan	20e	Non narratif	Pamphlet
Articles	NaN	21e	Non narratif	Articles encyclopédiques
Bouvard et Pécuchet	G. Flaubert	19e	Narratif	Roman
Code Civil 1	NaN	19e	Non narratif	Texte juridique
Code Civil 2	NaN	19e	Non narratif	Texte juridique
Code de procédure pénale	NaN	20e	Non narratif	Texte juridique
Convention univ	NaN	20e	Non narratif	Texte juridique
Convention marin	NaN	20e	Non narratif	Texte juridique
Convention aéro	NaN	20e	Non narratif	Texte juridique
Convention thon	NaN	21e	Non narratif	Texte juridique
Douce Lumière	M. Audoux	20e	Narratif	Roman
De la ville au moulin	M. Audoux	20e	Narratif	Roman
Élisabeth Seton	L. CoN/A	20e	Narratif	Biographie
Est Républicain 1	NaN	21e	Non narratif	Articles de presse
Est Républicain 2	NaN	21e	Non narratif	Articles de presse
Génie du christianisme	F. de Chateaubriand	19e	Non narratif	Traité argumentatif
Jean-Christophe 1	R. Rolland	20e	Narratif	Roman
Jean-Christophe 2	R. Rolland	20e	Narratif	Roman
Madame de Hautefort	V. Cousin	20e	Narratif	Biographie
Mademoiselle Fifi 1	G. de Maupassant	20e	Narratif	Nouvelle
Mademoiselle Fifi 2	G. de Maupassant	20e	Narratif	Nouvelle
Mademoiselle Fifi 3	G. de Maupassant	20e	Narratif	Nouvelle
La morte amoureuse	T. Gautier	19e	Narratif	Nouvelle
Le capitaine Fracasse	T. Gautier	19e	Narratif	Roman
Le Diable au corps	R. Radiguet	20e	Narratif	Roman
Le ventre de Paris	E. Zola	19e	Narratif	Roman
Nemoville	A. Bourgeois	20e	Narratif	Roman
Pauline	G. Sand	19e	Narratif	Roman
Rosalie de Constant	L. Achard	20e	Narratif	Biographie
Sarrasine	H. de Balzac	19e	Narratif	Nouvelle

Annexe 0.1 : Sous-corpus de Democrat. Pour plus de détails : https://www.ortolang.fr/market/corpora/democrat/v1

0.2 Stabilité référentielle

Audoux (2 units) 2 selected units / 1 words = 2
Douce (666 units) 666 selected units / 90 words = 7.4
Tou (76 units) 76 selected units / 25 words = 3.04
le verger (22 units) 22 selected units / 11 words = 2
la très vieille maison (5 units) 5 selected units / 5 words = 1
un clair matin de mai (2 units) 2 selected units / 2 words = 1
les mille et une fleurettes (2 units) 2 selected units / 2 words = 1
une souquenille de grosse toile (2 units) 2 selected units / 2 words = 1
les arbres (2 units) 2 selected units / 2 words = 1
une poignée de fleurettes (3 units) 3 selected units / 3 words = 1
les gamins (4 units) 4 selected units / 4 words = 1
l'entrée du potager (4 units) 4 selected units / 4 words = 1
le potager (6 units) 6 selected units / 2 words = 3
la large et forte grille de l'entrée du potager (19 units) 19 selected units / 8 words = 2.375
les barreaux de la grille (3 units) 3 selected units / 2 words = 1.5
les lances de la grille (5 units) 5 selected units / 5 words = 1
une haute et vaste sapinière (4 units) 4 selected units / 4 words = 1
des charrettes (2 units) 2 selected units / 2 words = 1
deux ornières (2 units) 2 selected units / 2 words = 1
l'homme à besace (2 units) 2 selected units / 2 words = 1
les aboiements furieux du chien (2 units) 2 selected units / 2 words = 1
des jurons et des menaces (2 units) 2 selected units / 2 words = 1
l'instant (2 units) 2 selected units / 2 words = 1
Noël Barray (233 units) 233 selected units / 40 words = 5.825
Douce + Tou (3 units) 3 selected units / 3 words = 1
Douce + Noël (52 units) 52 selected units / 22 words = 2.3636363636
les pommiers (2 units) 2 selected units / 2 words = 1
les fins papillons (2 units) 2 selected units / 2 words = 1
la maison 2 (4 units) 4 selected units / 1 words = 4
les parents de Douce (10 units) 10 selected units / 9 words = 1.1111111111
les réponses (2 units) 2 selected units / 2 words = 1
le père Lumière (97 units) 97 selected units / 34 words = 2.8529411765
Mère Clarisse (44 units) 44 selected units / 14 words = 3.1428571429
Bléroux (7 units) 7 selected units / 3 words = 2.3333333333
une grande ferme (2 units) 2 selected units / 2 words = 1
les trois chiens (2 units) 2 selected units / 2 words = 1
des chiens méchants (3 units) 3 selected units / 3 words = 1
Luc (2 units) 2 selected units / 2 words = 1
la maison (2 units) 2 selected units / 2 words = 1
le sous-bois (2 units) 2 selected units / 2 words = 1
les flèches d'or (2 units) 2 selected units / 2 words = 1
les jeudis qui suivirent (2 units) 2 selected units / 2 words = 1
Douce + Tou + Noël (10 units) 10 selected units / 5 words = 2
le jeu (4 units) 4 selected units / 2 words = 2

le petit âne (2 units) 2 selected units / 2 words = 1
Marie (2 units) 2 selected units / 2 words = 1
Joseph (8 units) 8 selected units / 6 words = 1.3333333333
Jesus (2 units) 2 selected units / 2 words = 1
les bergers (2 units) 2 selected units / 2 words = 1
le nouveau-né (2 units) 2 selected units / 2 words = 1
la bergère (2 units) 2 selected units / 2 words = 1
Jacob (6 units) 6 selected units / 6 words = 1
cette heure (2 units) 2 selected units / 2 words = 1
la table où fume la soupière (3 units) 3 selected units / 2 words = 1.5
un aboiement (2 units) 2 selected units / 2 words = 1
le chemin de la sapinière (4 units) 4 selected units / 4 words = 1
un bout de promenade (2 units) 2 selected units / 2 words = 1
le rayonnement du visage de Douce (2 units) 2 selected units / 2 words = 1
le jeune homme à la poitrine large (2 units) 2 selected units / 2 words = 1
la jeune fille à l'allure distinguée (3 units) 3 selected units / 3 words = 1
la gêne (2 units) 2 selected units / 2 words = 1
le silence qui se prolonge (2 units) 2 selected units / 2 words = 1
la main de Douce (4 units) 4 selected units / 3 words = 1.3333333333
tous les dimanches (3 units) 3 selected units / 1 words = 3
le garçon et la fille (2 units) 2 selected units / 2 words = 1
les joues de Douce (2 units) 2 selected units / 2 words = 1
les couleurs vives (2 units) 2 selected units / 2 words = 1
le retour de Noël (2 units) 2 selected units / 2 words = 1
des roses (3 units) 3 selected units / 3 words = 1
le jeune homme (2 units) 2 selected units / 2 words = 1
ce gamin de Noël (2 units) 2 selected units / 2 words = 1
les amitiés (2 units) 2 selected units / 2 words = 1
les relations de Douce (2 units) 2 selected units / 2 words = 1
les amusements de Douce (2 units) 2 selected units / 2 words = 1
un goût (2 units) 2 selected units / 2 words = 1
l'enfant qu'elle avait emportée (5 units) 5 selected units / 5 words = 1
Noël + le père Lumière (2 units) 2 selected units / 2 words = 1
l'amour (3 units) 3 selected units / 3 words = 1
la haine d'un grand père pour sa petite fille (4 units) 4 selected units / 4 words = 1
un grand-père (2 units) 2 selected units / 2 words = 1
"on" générique (2 units) 2 selected units / 1 words = 2
ces deux lignes (2 units) 2 selected units / 2 words = 1

Annexe [0.2] : Stabilité référentielle de chaque référent (ou chaîne) pour le bloc « Douce Lumière » de Democrat obtenue à l'aide de la macro TXM dédiée.

Table des matières

Liste des tableaux

Table des figures

Analyse en corpus de chaînes de coréférence

Résumé

Une chaîne de coréférence désigne l'ensemble des expressions linguistiques qui réfèrent à la même entité. La relation de coréférence entre les « maillons » d'une chaîne implique que le référent doit être strictement le même pour chaque expression qui la compose. Cependant, il arrive que le référent d'une expression soit difficile à identifier et que la relation de coréférence entre plusieurs expressions ne soit pas stricte de manière certaine. Pour un lecteur, ce manque de précision ne pose pas nécessairement de difficultés. En revanche, lors de l'annotation d'un corpus en coréférences, il est question d'indiquer clairement le référent de chaque expression. Les phénomènes de coréférence non stricte peuvent donc causer des difficultés d'annotation. Cette thèse a débuté au sein du projet ANR Democrat, avec une tâche d'annotation qui a permis de faire émerger des difficultés d'annotation théoriques et techniques liées à la coréférence non stricte. Nous proposons donc de passer en revue les phénomènes linguistiques impliqués dans la coréférence non stricte, notamment le flou (co)référentiel ainsi que les cas typiques relevés en corpus. Dans un second temps, nous proposons une étude de l'annotation de ces phénomènes dans un sous-corpus de Democrat. Cette étude révèle une grande variabilité d'annotation de ces phénomènes dont nous tirons une classification. Pour éviter les difficultés d'annotation liées à ces phénomènes, nous proposons un cadre plus précis pour l'annotation de la coréférence floue. Cela implique des précisions à ajouter au manuel d'annotation ainsi qu'un schéma d'annotation adapté, prenant en compte la coréférence floue.

Mots clés : référence, coréférence, chaînes de coréférence, flou, annotation, corpus, schéma d'annotation.

Corpus analysis of coreference chains

Abstract

A coreference chain designates the set of linguistic expressions that refer to the same entity. The coreference relation between a chain's elements implies that the referent must be strictly the same for each expression that composes it. However, the referent of an expression is sometimes difficult to identify and the coreference relation between several expressions cannot therefore be strict without any doubt. For a reader, this lack of precision does not necessarily pose difficulties. Nevertheless, the coreference annotation task of a corpus consists in unequivocally identifying the referent of each expression. Non-strict coreference phenomena can therefore generate annotation difficulties. This thesis began within the ANR Democrat project, with an annotation task that highlighted the emergence of theoretical and technical annotation difficulties related to non-strict coreference. We thus propose to review the linguistic phenomena involved in non-strict coreference, in particular the (co)referential fuzziness as well as the typical cases found in corpora. In a second step, we propose a study of the annotation of these phenomena in a Democrat sub-corpus. This study reveals a great variability in the annotation of these phenomena from which we derive a classification. To avoid the annotation difficulties related to these phenomena, we propose a more precise framework for the annotation of fuzzy coreference. This implies precisions to be added to the annotation manual as well as an adapted annotation scheme, taking into account the fuzzy coreference.

Keywords : reference, coreference, coreference chains, fuzzyness, annotation, corpus, annotation scheme.

UNIVERSITE SORBONNE NOUVELLE
ED 622 - Sciences du langage : Maison de la recherche — 4, rue des irlandais — 75005 Paris
Lattice - UMR 8094 : École Normale Supérieure — 1, rue Maurice Arnoux — 92120 Montrouge