Détection automatique de chaînes de coréférence pour le français écrit

Bruno Oberle

Détection automatique de chaînes de coréférence pour le français écrit: règles et ressources adaptées au repérage de phénomènes linguistiques spécifiques

Bruno Oberle[1]
(1) Linguistique, Langues, Parole (LiLPa, EA-1339), Université de Strasbourg, F-67084 Strasbourg, France
oberleb@unistra.fr

RÉSUMÉ

Nous présentons un système *end-to-end* de détection automatique des chaînes de coréférence, à base de règles, pour le français écrit. Ce système insiste sur la prise en compte de phénomènes linguistiques négligés par d'autres systèmes. Nous avons élaboré des ressources lexicales pour la résolution des anaphores infidèles (*Mon chat... Cet animal...*), notamment lorsqu'elles incluent une entité nommée (*La Seine... Ce fleuve...*). Nous utilisons également des règles pour le repérage de mentions de groupes (*Pierre et Paul*) et d'anaphores zéros (*Pierre boit et ø fume*), ainsi que des règles pour la détection des pronoms de première et deuxième personnes dans les citations (*Paul a dit : "**Je** suis étudiant."*). L'article présente l'élaboration des ressources et règles utilisées pour la gestion de ces phénomènes spécifiques, avant de décrire le système dans son ensemble, et notamment les différentes phases de la résolution de la coréférence.

ABSTRACT

Automatic coreference resolution for written French : rules and resources for specific linguistic phenomena

We introduce a new rule-based coreference resolution system for written French. This system takes into account linguistic phenomena often ignored by other systems. First, we have built lexical resources to improve full NP coreference resolution (*My cat... The animal*), especially when a named entity is involved (*The Seine... The river...*). We have defined rules to detect groups of individuals (*Peter and Paul*) and null anaphora (*Peter drinks and ø smoke*). We have also defined rules to detect first and second person pronouns in quotations (*Paul said : "**I** am a student"*). This paper first presents how we built our lexical resources and how we defined our rules, then it describes how our system works and specifically what are the steps to resolve coreference.

MOTS-CLÉS : expression référentielle, coréférence, détection automatique de la coréférence, entité nommée.

KEYWORDS: referring expression, coreference, automatic coreference detection, named entity.

1 Introduction

Une expression référentielle, ou *mention*, est un segment de texte qui renvoie à une entité extralinguistique : le référent. Lorsque le même référent est repris dans le même texte par une autre mention, une *relation de coréférence* (Corblin, 1995) s'établit entre les mentions, qui font alors partie d'une

même *chaîne de coréférence* (Chastain, 1975; Corblin, 1985; Charolles, 1988; Schnedecker, 1997). Ainsi, dans l'exemple suivant, les termes *Platon, Il, ses* et *il* renvoient tous au même référent :

(1) *Platon* est un philosophe antique de la Grèce classique... *Il* reprit le travail philosophique de certains de *ses* prédécesseurs, notamment Socrate dont *il* fut l'élève. (Wikipédia)

On considère généralement (Charolles, 1988; Schnedecker & Landragin, 2014) que les mentions sont des expressions nominales (noms et pronoms), mais également des déterminants possessifs puisqu'ils varient en personne (*mon vs. ton vs. son*, etc.) et en nombre (*mon vs. notre, son vs. leur*, etc.) en fonction du possesseur, et qu'ils comprennent une référence à une entité.

Un système de détection automatique des chaînes de coréférence assemblent les mentions (expressions référentielles) d'un texte en chaînes de coréférence. On dit qu'il est *end-to-end* (ou *de bout-en-bout*) lorsqu'il repère également les mentions, et qu'il accepte donc en entrée un texte brut, plutôt qu'un texte pré-annoté en mentions.

Plusieurs systèmes de ce type ont été créés pour le français (notre langue cible étant le français, nous n'évoquerons pas les systèmes anglais ; voir cependant (Poesio *et al.*, 2016), notamment la partie 2). Le système de Trouilleux (Trouilleux, 2001), qui ne s'occupe que de certaines anaphores pronominales, celui de Dupont (Dupont, 2003; Victorri, 2005) et RefGen (Longo, 2013) sont des systèmes symboliques, qui résolvent la coréférence à partir de règles prédéfinies. Leurs performances sont généralement moindres que les systèmes suivants, développés à partir du seul corpus francophone pour l'heure annoté en coréférences : ANCOR (Muzerelle *et al.*, 2013). Trois de ces systèmes font appel à l'apprentissage automatique : une adaptation de BART au français dans le cadre du projet SENSEI (Kabadjov & Stepanov, 2015), ainsi que CROC (Désoyer *et al.*, 2015, 2016) et le système décrit dans (Brassier *et al.*, 2018). Ces systèmes ne fonctionnent que sur des mentions gold, c'est-à-dire qu'ils ne détectent pas eux-mêmes les expressions référentielles, mais seulement la coréférence. Un quatrième système, décrit par (Godbert & Favre, 2017), qui se limite aux anaphores pronominales et aux anaphores fidèles (reprise d'un nom par le même nom, par exemple *"le chat... ce chat"*, ce que les concepteurs d'ANCOR appellent "relations directes"), utilise des règles. Tout récemment, des approches à partir des réseaux de neurones ont été développées (Grobol, 2019).

Le corpus ANCOR, cependant, est un corpus d'oral transcrit (conversations téléphoniques et entretiens) dont les spécificités et l'annotation ne correspondent pas toujours aux caractéristiques de l'écrit.

Notre objectif est d'élaborer, pour le français écrit, un système *end-to-end* de détection automatique de la coréférence, qui s'attache à repérer des phénomènes coréférentiels fréquents dans les textes écrits mais qui sont souvent négligés par les autres systèmes.

L'une des difficultés des systèmes de détection de la coréférence est la résolution des anaphores infidèles (reprise d'un nom par un synonyme, par exemple *Le chat... Cet animal...*), alors qu'il s'agit d'un phénomène très fréquent en français écrit, qui supporte mal la répétition. Notre intérêt s'est donc d'abord porté sur la résolution de ce type d'anaphores, notamment lorsqu'elles impliquent des entités nommées (*Le Rhin... Ce fleuve...*, ou *Paris... La Ville Lumière*).

Ensuite, nous nous sommes intéressé aux phénomènes coréférentiels dans lesquels la détection de la mention elle-même est difficile :

— lorsque le référent est un groupe qui englobe plusieurs autres mentions individuelles : ***Sophie et Marie*** *sortent ensemble.* ***Elles*** *sont très heureuses*. Dans cet exemple, c'est le groupe, qui a des caractéristiques propres (féminin, pluriel), qui est repris ;

— lorsque le référent est repris sans être exprimé (cas de "l'anaphore zéro") : ***Paul*** *boit et* ***ø*** *fume*. Ici, pour être en mesure de comprendre que Paul fait deux actions (boire et fumer), il faut être en mesure de "distribuer" le sujet sur les deux verbes.

Enfin, il est fréquent de trouver dans la presse écrite des citations qui contiennent des pronoms et déterminants possessifs de première et deuxième personnes (*je, tu, nous, vous, mon, ton, notre, votre*). Pour comprendre qui, dans ces citations, exprime une opinion, il faut pouvoir associer le "je" et le "tu" à des référents dont les noms sont donnés dans le texte environnant la citation :

(2) *Lionel Jospin* se livre en revanche à une longue analyse de *son* échec du 21 avril. "*Ma* part de responsabilité dans l'échec existe forcément. *Je* l'ai assumée en quittant la vie politique", explique-t-*il* d'emblée. (L'Est Républicain)

Ces cas complexes demandent des stratégies spécifiques. Pour la résolution des anaphores infidèles, nous avons construit deux ressources lexicales : l'une qui déduit des informations utiles pour la détection de la coréférence à partir d'entités nommées liées à une base de connaissances, l'autre qui est un dictionnaire d'hyperonymes.

Les autres phénomènes sont traités par des règles appliquées sur une analyse syntaxique faite par Talismane (Urieli, 2013).

La suite de cet article insiste d'abord sur la conception des ressources lexicales (section 2) que nous avons créées et des règles que nous avons établies (section 3), avant de décrire le fonctionnement général du système de détection de la coréférence ansi que son évaluation (section 4).

Le système que nous proposons est donc un système qui s'appuie à la fois sur des méthodes d'apprentissage statistiques (Talismane) pour la détection des mentions et l'analyse syntaxique, et sur des règles pour la détection de la coréférence et la gestion de phénomènes linguistiques complexes mais bien définis dans le français écrit.

2 Détection des anaphores infidèles

La reprise d'un référent dans le texte peut se faire de plusieurs façons. Il peut s'agir d'un pronom : *Paul est dehors.* ***Il*** *attend.*

Lorsque la reprise se fait par un nom, on parle d'*anaphore fidèle* quand le nom est le même (cependant, la détermination et les expansions peuvent être différentes) :

(3) *Le chat* que j'ai adopté court partout... Mais j'aime beaucoup *ce chat*.

On parle au contraire d'*anaphore infidèle* lorsque le nom de reprise est différent :

(4) *Mon chat* boit du lait. *Cet animal* est heureux.

C'est à ce dernier cas que nous nous intéressons ici.

2.1 Hyperonymes

Les deux noms peuvent s'incrire dans une relation lexicale telle que la synonymie, ou l'hyperonymie lorsque la relation est hiérarchique (par exemple : chat > félin > mammifère > animal).

Tester une telle relation entre deux noms permet de vérifier qu'ils sont sémantiquement compatibles : *chat* et *animal* pourront (ou non) être coréférents, mais *chat* et *chien* ne le seront jamais.

Il faut noter que les relations lexicales peuvent varier d'un domaine de spécialité à l'autre. Ainsi, alors que dans l'usage courant *chat* est généralement compris comme un animal, dans le domaine technologique, le terme est synonyme de *conversation électronique*, dans le domaine maritime, il désigne un type de *yacht* (qui est alors un hyperonyme de *chat*), dans le domaine militaire, il renvoie à une machine de siège, etc.

Pour connaître la relation entre deux termes, il faut une ressource contenant une liste de synonymes et d'hyperonymes. Le *WOrdnet Libre du Français (WOLF)* (Sagot & Fišer, 2008) a été construit par traduction automatique du WordNet originel de Princeton. Nous avons cependant décidé d'élaborer une nouvelle ressource à partir d'un dictionnaire français, le *Glawi* (Hathout & Sajous, 2016; Sajous & Hathout, 2015), qui a été construit à partir du *Glaff* (Hathout *et al.*, 2014) et du *Wiktionnaire* [1], un dictionnaire "collaboratif" au sens où chaque internaute peut proposer et modifier des gloses (les définitions), des exemples, mais aussi des synonymes, hyponymes et hyperonymes. Ce dictionnaire permet de relier les différents sens d'un mot à des domaines sémantiques (marine, militaire, technologique, etc.). Nous n'avons cependant pas pu utiliser ces relations proposées par les internautes, parce qu'elles sont codées dans le *Wiktionnaire* et donc dans le *Glawi* d'une façon telle qu'on ne peut pas distinguer les domaines sémantiques.

Néanmoins, les gloses du *Wiktionnaire* sont généralement rédigées de façon à faire apparaître un hyperonyme en début de phrase (définitions dites "par le genre prochain"). Ainsi un *chat* est :

- un ***mammifère*** *carnivore félin de taille moyenne...*,
- un ***bélier*** *recouvert par un chat...*,
- un ***navire*** *servant au chargement et au déchargement...*,
- etc.

De même, une *pomme* est le ***fruit*** *comestible du pommier...*, une *table* est un ***meuble*** *composé d'un plateau posé sur un ou plusieurs pieds...*, et ainsi de suite.

Le *Glawi* propose pour chacune des gloses une analyse syntaxique faite avec Talismane (Urieli, 2013) : nous avons donc extrait pour chaque terme un ou plusieurs hyperonymes à partir du premier nom commun de la glose. Par transitivité (un chat est un mammifère, un mammifère est un animal, etc.), nous avons pu constituer pour chaque terme une collection d'hyperonymes sur plusieurs niveaux. De plus, comme chaque glose est associée à un ou plusieurs domaines de spécialité (marine, botanique, etc.), chaque collection d'hyperonymes est associée à un domaine sémantique particulier. Il est donc possible de ne faire usage que de tel ou tel domaine en fonction du texte à annoter automatiquement : *chat* pourra ainsi être reconnu comme un type d'animal, de conversation, de navire ou d'équipement de guerre en indiquant au système le domaine sémantique du texte.

Nous avons annoté de la sorte 160 000 noms communs (puisque, parmi les classes de mots lexicaux, seuls les noms sont considérés comme référentiels).

2.2 Entités nommées

Les anaphores infidèles peuvent impliquer des entités nommées, c'est-à-dire des désignations de personnes, de lieux et d'organisations (Nouvel *et al.*, 2015) mais aussi d'objets (la "Pyramide du

1. fr.wiktionary.org

Louvre") ou d'oeuvres ("Les Misérables").

Dans une relation entre deux expressions, l'une peut être une entité nommée et l'autre un nom commun :

(5) *Paris* est située sur la Seine. *La plus grande ville de France* compte plus de 10 millions d'habitants.

Mais les expressions peuvent aussi être toutes les deux des entités nommées, comme *Paris* et la *Ville Lumière*.

Le seule détection des entités nommées (par exemple : *JRC Names* [2], *CasEN* (Friburger & Maurel, 2004; Maurel *et al.*, 2011), *mXS* (Nouvel *et al.*, 2011), voir (Nouvel *et al.*, 2015, ch. 4)) ne suffit donc pas ici : il faut également obtenir d'une part une liste de noms communs avec lesquels il est possible de reprendre chaque entité (Paris est une *ville*, une *capitale*, le Rhin est un *fleuve*, etc.), d'autre part une liste d'entités nommées (désignations) qui renvoient à la même entité (*Paris* et la *Ville Lumière*, l'*Organisation des Nations Unies* et l'*ONU*, le *Festival de Cannes* et le *Festival International du Film de Cannes* par exemple). C'est pourquoi nous avons cherché à extraire une liste d'entité nommées et à la lier à une base de connaissances.

Parmi les bases disponibles (par exemple *BabelNet* [3], *Wikidata* [4], *DBPedia* [5]), nous avons choisi *Yago* (Mahdisoltani *et al.*, 2014), extrait de *Wikipédia*, *WordNet* et *GeoNames*, qui est téléchargeable, modifiable et redistribuable librement (contrairement à *BabelNet*, par exemple, qui n'est disponible que *via* une API), et qui contient toutes les informations dont nous avons besoin.

La ressource est constituée de plus d'un milliard de triplets de la forme : `(sujet, prédicat, relation)`, par exemple

(6) `<Zinedine_Zidane> <redirectedFrom> "Zizou"@fra`

Cet exemple montre une redirection de "Zizou" vers l'entité `<Zinedine_Zidane>`. En listant toutes les redirections, on obtient la liste de toutes les désignations possibles de Zidane : *El Zid, Zizou, Zinedine Yazid Zidane*, etc. Avec les prédicats `<hasGivenName>` et `<hasFamilyName>`, on ajoute encore des désignations par son nom (*Zidane*) ou son prénom (*Zinédine*). Des règles permettent également de considérer l'usage d'initiales comme *Z. Zidane*.

Par ailleurs, toutes les entités de *Yago* sont reliées à l'ontologie *WordNet* :

(7) `<Zinedine_Zidane> rdf:type <wordnet_football_player_110101634>`

dont les éléments sont associés à des "étiquettes", y compris en français :

(8) `<wordnet_football_player_110101634> rdfs:label "footballeur"@fra`

Ainsi, Zidane est un *footballeur*, mais aussi, en remontant l'ontologie, un *sportif*, un *concurrent*, un *homme*, etc. À chaque niveau se trouvent un ou plusieurs termes synonymes (*sportif, athlète* ; *concurrent, rival, participant, compétiteur...*). Cette liste d'hyperonymes et de synonymes permet de favoriser un lien de coréférence entre, par exemple, *Zidane* et *athlète* plutôt qu'entre *Zidane* et *ministre* ou *Zidane* et *fleuve*.

2. https://ec.europa.eu/jrc/en/language-technologies/jrc-names
3. http://babelnet.org
4. https://wikidata.org
5. https://dbpedia.org

Un autre prédicat (<hasGender>) disponible dans *Yago* est utile pour la coréférence, même s'il concerne les anaphores pronominales, c'est le sexe de la personne : Zinédine Zidane sera repris par *il* et non par *elle* :

(9) <Zinedine_Zidane> <hasGender> <male>

Au total, nous avons extrait de *Yago* une liste de 6,4 millions de désignations pour 2,3 millions d'entités, chacune annotée avec

— les désignations possibles de l'entité,
— son type,
— ses hyperonymes,
— le genre du pronom par lequel l'entité peut être reprise (masculin, féminin).

Nous prévoyons également d'utiliser les "infoboxes" de *Wikipédia*, également présentes dans *Yago*, pour la détection des désignations comprenant des noms communs, comme "le ministre de l'Éducation Nationale".

3 Repérage des mentions de groupes et d'anaphores zéros

Les autres phénomènes coréférentiels annoncés dans l'introduction (groupes, anaphores zéros) sont traités à partir d'une analyse syntaxique automatique faite par Talismane (Urieli, 2013). La sortie est convertie en arbre dont les noeuds peuvent être déplacés ou modifiés en fonction de règles prédéfinies.

L'application de ces règles est sensible aux erreurs de l'analyseur, c'est pourquoi nous appliquons auparavant des règles de correction. Ce sont des règles linguistiques, dont nous donnons deux exemples. D'abord, une série de règles vérifie que seuls des éléments de même type sont coordonnés (par exemple, un verbe ne peut pas être coordonné à un nom) : si ce n'est pas le cas, les noeuds sont déplacés pour obtenir une coordination linguistiquement correcte. Une telle correction est notamment nécessaire pour trouver les groupes et les anaphores zéros. Ensuite, une autre règle vérifie que toutes les propositions relatives sont dépendantes d'un nom, et non d'un verbe ou d'une autre partie du discours, sinon la relative est déplacée et "accrochée" au nom le plus plus probable. Une telle correction est notamment nécessaire pour trouver l'antécédent du pronom relatif, qui ne peut être qu'un nom.

3.1 Les groupes

On appelle ici "groupe" les expressions référentielles coordonnées qui, du fait de leur coordination, forment une nouvelle expression référentielle. Ainsi, dans *Pierre et Marie*, il y a trois expressions référentielles et trois référents : *Pierre*, *Marie*, mais aussi aussi le couple (groupe) *Pierre et Marie*.

Le repérage des groupes est nécessaire pour la détection de la coréférence, car ils peuvent être repris par des pronoms, ou même par un nom collectif :

(10) *Jack et Rose* commencent à faire connaissance. *Ils* s'entendent bien. [...] *Le couple*... (Vikidia)

Nous détectons les groupes en recherchant, dans les arbres de chaque phrase, les juxtapositions ou coordinations de deux ou plusieurs noms. Ces coordinations sont marquées pour signaler une mention, mais elles sont aussi annotées en genre et en nombre. Le nombre est toujours pluriel, mais le genre

varie en fonction des individus qui composent le groupe (*Sophie et Marie* est repris par *elles*, et non par *ils*).

Parfois, les groupes ne sont pas coordonnés (cas des antécédents dits "dispersés") ; ces cas sont pas repérés par notre système :

(11) *Pierre* retrouva *sa femme* au restaurant. *Ils* dînèrent jusqu'à tard dans la nuit.

3.2 Les anaphores zéros

On parle d'anaphore zéro, ou sujet zéro ou non exprimé (indiquée par "ø" dans les exemples), lorsqu'un sujet est mis en "facteur commun" de plusieurs verbes coordonnés :

(12) *Le renard* se glisse, ø se traîne, ø arrive, et ø fait rarement des tentatives inutiles. (Buffon)

Le repérage de ces segments pourtant absents du texte est important car ils donnent des informations sur le référent (Schnedecker & Landragin, 2014) et ils sont donc utiles pour l'interprétation automatique des textes (dans l'exemple, l'action du renard ne se réduit pas à *se glisser*) ou l'étude des chaînes de coréférence (voir par exemple (Schnedecker, 2014)). Il est donc nécessaire de pouvoir "distribuer" ce facteur commun à tous les verbes coordonnés.

Pour ce faire, nous cherchons dans les arbres les verbes juxtaposés ou coordonnés, vérifions que le premier verbe a un sujet mais pas le ou les suivants, et nous dupliquons le sujet du premier verbe sur le deuxième (troisième, etc.) verbe. La mention est marquée sur le verbe (plutôt que d'ajouter le symbole ∅, ce qui modifierait le texte original). Un lien de coréférence est immédiatement établi.

On ne considère ici que les verbes conjugués. On peut aussi parler de sujet zéro pour les participes et les verbes à l'infinitif (Landragin, 2011), mais nous ignorons ces cas car ils n'acceptent pas l'explicitation du sujet. On peut en effet dire *Paul boit et {ø,Paul} fume*, mais pas *Paul pense {ø,*Paul} aller au cinéma ce soir.*

Ce sont parfois des groupes (voir la section précédente) qui sont des anaphores zéros :

(13) *Jack et Rose* sortent des profondeurs du navire et *assistent* au début de la fin du Titanic. (Vikidia)

3.3 Détection de la coréférence des pronoms de première et deuxième personnes dans les citations

Dans les articles de presse, on trouve fréquemment des citations (discours direct) incluant des pronoms (par simplicité, nous parlons de *pronoms*, mais la discussion s'étend aux déterminants possessifs : *mon, ton, son, notre, votre, leur*, etc.) de première et deuxième personnes, dont le référent varie selon la citation, ce qui veut dire qu'on ne peut pas lier tous les "je" du texte ensemble, comme dans l'extrait suivant :

(14) Comme *Solange*, présidente du club des aînés de la commune : "*J'*ai un ordinateur mais *je* ne sais pas *m'*en servir..." *Jacques*, ancien agriculteur, embraie : "L'autre jour, *j'*ai imprimé un document. Toutes les cases se chevauchaient. C'était le bordel, alors *j'*ai débranché l'appareil." *Marivonne*, épouse de gendarme, assise derrière lui, abonde : "*J'*ai un ordinateur à la maison, *je* m'en sers pour jouer gratuitement au solitaire et puis c'est tout." (Le Monde)

Les pronoms de première et deuxième personnes peuvent être considérés comme des déictiques plutôt que des anaphoriques : c'est pourquoi certains corpus (par exemple ANCOR (Muzerelle *et al.*, 2013)) font le choix de ne pas les mettre dans une relation de coréférence. Néanmoins, tant dans la perspective de la fouille de textes que dans l'étude des chaînes de coréférence, il semble important de savoir qui se cache derrière les "je" et les "tu" des citations.

Il convient aussi de repérer les incises (*Paul, **dit-il**, a eu une bonne note.*) et les verbes "de dire" (*dire* et ses synonymes), notamment quand ils sont situés à la fin avec inversion du sujet (*Jacques s'emporta : "Je ne suis pas d'accord avec Paul : et je pense qu'il a tort", **affirma-t-il avec force***), afin d'annoter correctement le sujet de ces verbes, qui est le locuteur du discours direct.

Notre méthode consiste à repérer, avant même l'analyse syntaxique par Talismane, le discours rapporté direct à partir de la présence de guillemets. Dans le discours direct, nous repérons ensuite les incises et les verbes "de dire" avec inversion du sujet afin de les extraire et d'en faire des phrases indépendantes : cela est nécessaire pour que l'analyseur syntaxique fournisse un arbre qui soit ensuite plus aisément exploitable. Une fois l'analyse syntaxique effectuée, les pronoms (et déterminants) de première et deuxième personnes sont repérés et marqués comme mentions. La résolution de la coréference tient compte des différentes annotations : le type de discours (direct ou non), la personne, le nombre, etc.

Notre repérage reste cependant limité : le système détecte des citations telles qu'on les trouve, le plus souvent, dans les articles de presse : elles doivent être encadrées par des guillemets (pas de discours direct libre) et mono-voix (pas de dialogue à l'intérieur de la même citation).

4 Présentation générale du système et de la résolution de la coréférence

Dans cette dernière section, nous décrivons l'architecture de notre système en insistant sur la façon dont nous détectons la coréférence.

Le système étant *end-to-end*, il doit effectuer deux tâches : (1) repérer les mentions dans un texte brut, (2) établir des liens de coréférence entre les expressions référentielles qui renvoient au même référent.

4.1 Repérage des mentions

Le repérage des mentions consiste en plusieurs étapes :

- Repérage des entités nommées et leur annotation (désignations, type, hyperonymes, genre du pronom de reprise, c'est-à-dire le genre du pronom par lequel l'entité peut être reprise : masculin s'il s'agit d'un homme ou d'une entité désignée par un nom masculin, féminin sinon) à partir de la ressources décrite ci-dessus.
- Repérage du discours direct, extraction des incises et des verbes de dire comme décrit ci-dessus.
- Tokenisation pour tenir compte des expressions figées non référentielles (comme *à peine* ou *jouer des coudes*).
- Analyse syntaxique avec Talismane (Urieli, 2013).
- Construction et correction de l'arbre résultant de l'analyse syntaxique.
- Ajout d'informations et d'annotations (groupe, anaphore zéro, etc.).

— Repérage des pronoms non référentiels (adjectifs de météo, verbes impersonnels, etc.).
— Repérage des mentions (expressions référentielles) à partir de l'analyse syntaxique.

4.2 Détection de la coréférence

Les relations de coréférence sont détectées en plusieurs passes, selon le type des relations et d'informations nécessaires (syntaxiques et/ou sémantiques). Faute de place, nous ne donnons qu'une description succincte de ces différentes passes.

Passe 1 : les anaphores liées. On dit qu'il y a anaphore liée lorsque "l'antécédent... est choisi en vertu d'un calcul purement syntaxique" (Corblin, 1995). On obtient alors des "chaînes liées", dont "la construction de la chaîne est exclusivement régie par la syntaxe" (voir aussi (Zribi-Hertz, 1996)).

Le cas le plus évident est celui des *pronoms réfléchis* : leur antécédent est toujours le sujet de la proposition dans laquelle ils apparaissent. Il n'y a besoin d'aucune analyse supplémentaire, ce seul fait syntaxique permet de résoudre l'anaphore dans tous les cas.

Nous résolvons aussi à ce stade les *pronoms relatifs*, qui renvoient toujours au nom dont dépend la proposition relative (éventuellement par l'intermédiaire d'une préposition, puisque le pronom relatif peut être le régime d'une préposition).

Passe 2 : les autres anaphores pronominales. La résolution des autres anaphores (pronoms et déterminants possessifs) requiert à la fois des connaissances syntaxiques et des connaissances sémantiques. Les connaissances syntaxiques permettent de savoir où chercher l'antécédent, et où ne pas le chercher. Par exemple, il ne peut pas y avoir dans la même structure argumentale d'un verbe deux fois le même référent. Dans l'exemple suivant, *lui* ne peut pas désigner Pierre.

(15) Pierre lui donne un cadeau.

La syntaxe permet aussi de détecter les cas de cataphore (lorsque le pronom est *avant* son "antécédent" : *Quand **il** était petit, **Paul** jouait au foot.*). Notre système détecte ces cas de cataphores lorsque le pronom est dans une subordonnée en début de phrase et coréfère au sujet de la proposition principale. Afin d'écarter les faux positifs, nous ne cherchons les cas de cataphores qu'au début du paragraphe.

Dans les autres cas, des candidats sont cherchés dans la phrase courante et la phrase précédente, et, si aucun candidat n'est trouvé, jusqu'au début du paragraphe. Les candidats incompatibles sont éliminés, notamment en fonction du genre et du nombre, mais aussi d'autres règles. Par exemple, le système considère (même si ce n'est pas toujours le cas dans les faits) qu'un pronom démonstratif ne peut avoir pour antécédent qu'un nom dans la phrase précédente. Autre exemple : un pronom de première personne ne peut pas être coréférent à un nom dans une même citation.

Les candidats sont ensuite triés selon plusieurs critères, qui diffèrent selon qu'il s'agit de noms ou de pronoms : la saillance (voir (Mitkov, 2002; Landragin, 2015)) dans le texte et dans le paragraphe en cours (calculée à partir du nombre d'occurrences du terme, de la fonction, du fait que la phrase d'occurrence est une phrase en début de texte ou de paragraphe, etc.), la fonction du candidat et de l'anaphore, la distance en termes de phrases et de mentions, etc. Le meilleur candidat est retenu comme antécédent de l'anaphore.

Passe 3 : les entités nommées. À ce stade, nous avons des chaînes partielles, et tous les pronoms et déterminants possessifs sont censés avoir un antécédent. Les passes suivantes cherchent à fusionner les chaînes partielles qui renvoient au même référent.

Le système commence par les chaînes partielles dont l'une des mentions est une entité nommée : si les entités renvoient au même référent, les chaînes sont fusionnées. Par exemple, deux chaînes contenant l'une *Zinédine* et l'autre *Zizou* (ou l'une *Paris* et l'autre *Ville Lumière*, etc.) seront fusionnées. Nous utilisons la resource lexicale décrite plus haut pour cela.

Passe 4 : les noms communs. Le système assemble ensuite les chaînes partielles en se servant des noms communs. Chaque chaîne partielle peut être fusionnée avec l'une des chaînes précédentes du texte : ce sont des chaînes candidates. Pour chaque chaîne candidate, le ou les noms têtes des syntagmes nominaux de la chaîne candidate et celui ou ceux de la chaîne partielle sont comparés. Les chaînes candidates dont le ou les noms ne sont pas sémantiquement compatibles avec la chaîne courante sont éliminées. Deux noms sont sémantiquement compatibles si l'un est un synonyme ou un hyperonyme de l'autre (y compris pour les entités nommées).

Les chaînes candidates restantes sont triées selon des critères similaires à ceux utilisés pour la phase 2 (saillance, fonction, distance, etc.), ainsi que des critères propres aux syntagmes nominaux, comme la détermination. Par exemple, un défini aura plus de chance de suivre un indéfini que l'inverse (*J'ai vu* ***un chat*** *dans la rue...* ***L'animal*** *m'a suivi tout l'après-midi* sera plus probable que *J'ai vu* ***le chat*** *dans la rue...* ***Un animal*** *m'a suivi tout l'après-midi*).

4.3 Évaluation

L'évaluation des systèmes de détection de la coréference est rendue complexe par le manque de corpus de référence. ANCOR (Muzerelle *et al.*, 2013), pour l'heure le seul corpus annoté en coréférences, est un corpus d'oral transcrit, et dont les choix d'annotation ne correspondent pas aux nôtres.

Nous avons comparé notre système à RefGen (Longo, 2013), dont l'architecture est est similaire à la nôtre :

- les mentions sont détectées à partir d'un étiqueteur en partie du discours (TTL (Ion, 2007)),
- la coréférence est résolue à partir de règles. Une première passe détecte les premières mentions de chaînes, en utilisant des éléments de la théorie de l'accessibilité décrite dans (Ariel, 1990). Une seconde passe rattache les expressions référentielles restantes aux différentes chaînes, en fonction de contraintes syntaxiques et sémantiques (forme de la tête du syntagme, compatibilité en genre et nombre, distance, fonction syntaxique, fréquence de co-occurrence, etc.).

Les choix d'annotation de notre système sont quelque peu différents de ceux de RefGen : par exemple, Refgen considère qu'une chaîne recommence à chaque nom propre, et ne prend en compte ni les anaphores infidèles, ni les anaphores zéros, ni les groupes. Nous avons donc réannoté manuellement un sous-ensemble de 3 250 tokens du corpus d'évaluation de RefGen (composés d'un extrait de texte littéraire, de faits divers et de textes de FLE) afin que chaque système soit évalué sur un corpus adapté à ses propres choix d'annotations.

Nous donnons dans le tableau 1 la mesure pour l'identification des mentions et les F-mesures de chacune des quatre mesures habituellement utilisées pour l'évaluation de la coréférence (calculées

avec l'implémentation officielle du script utilisé pour l'évaluation des campagnes CoNLL-2011/2012), en comparant avec RefGen.

	identification	MUC	B^3	CEAF	BLANC
notre système	83.62	58.1	64.53	69.28	53.08
RefGen	50.2	36	31.9	29.9	24.9

TABLE 1 – Évaluation de notre système et de RegGen (Longo, 2013) (F-mesures).

Parmi les systèmes récents, celui de (Godbert & Favre, 2017) reprend les choix d'annotation d'ANCOR et a été évalué sur une partie de ce corpus (23 079 mentions). Les auteurs ne donnent que le score BLANC qui est de 65.7.

CROC (Désoyer *et al.*, 2016) et le système de (Brassier *et al.*, 2018) ont des scores meilleurs sur ANCOR, mais ne fonctionnent que sur des mentions gold.

5 Conclusion

Nous avons présenté un système de détection de la coréférence en français qui s'attache à traiter des phénomènes négligés par les autres systèmes : détection d'anaphores infidèles, notamment avec entités nommées, repérage de groupes et d'anaphores zéros et prise en compte des pronoms et déterminants possessifs de première et deuxième personnes dans le discours rapporté direct.

Notre système utilise des règles plutôt qu'une approche par apprentissage automatique. Même si cette dernière est plus performante (Désoyer *et al.*, 2016; Brassier *et al.*, 2018) pour la détection de la coréférence en générale, les règles semblent bien adaptées pour le traitement de phénomènes spécifiques comme ceux que nous avons évoqués. C'est pourquoi elles pourraient venir en complément de systèmes par apprentissage pour la gestion de ces aspects. Nous obtiendrions alors un système hybride qui combinerait le meilleur des deux approches.

Remerciements

Ce travail a été réalisé avec le soutien du projet ANR Democrat ("Description et modélisation des chaînes de référence : outils pour l'annotation et le traitement automatique", ANR-15-CE38-0008).

Références

ARIEL M. (1990). *Accessing Noun Phrase Antecedents*. Routledge.

BRASSIER M., PURET A., VOISIN-MARRAS A. & GROBOL L. (2018). Classification par paires de mention pour la résolution des coréférences en français parlé interactif. In *Traitement Automatique des Langues Naturelles (TALN'18)*.

CHAROLLES M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques*, **57**, 3–13.

CHASTAIN C. (1975). Reference and context. In K. GUNDERSON, Ed., *Language, mind, and knowledge* : University of Minnesota Press.

CORBLIN F. (1985). Les chaînes de référence : analyse linguistique et traitement automatique. *Intellectica*, **1**(1), 123–143.

CORBLIN F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes.

DÉSOYER A., LANDRAGIN F., TELLIER I., LEFEUVRE A. & ANTOINE J.-Y. (2015). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *Traitement Automatique des Langues*, **55**(2), 97–121.

DÉSOYER A., LANDRAGIN F., TELLIER I., LEFEUVRE A., ANTOINE J.-Y. & DINARELLI M. (2016). Coreference resolution for French oral data : Machine learning experiments with ANCOR. In *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2016)*.

DUPONT M. (2003). *Une approche cognitive du calcul de la référence*. Thèse de doctorat, Université de Caen.

FRIBURGER N. & MAUREL D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, **313**(1), 93–104.

GODBERT E. & FAVRE B. (2017). Détection de coréférences de bout en bout en français. In *Traitement Automatique des Langues Naturelles (TALN'17)*.

GROBOL L. (2019). Neural coreference resolution with limited lexical context and explicit mention detection for oral French. In *Computational Models of Reference, Anaphora and Coreference (CRAC)*.

HATHOUT N. & SAJOUS F. (2016). Wiktionnaire's wikicode glawified : a workable French machine-readable dictionary. In *International Conference on Language Resources and Evaluation (LREC'16)*.

HATHOUT N., SAJOUS F. & CALDERONE B. (2014). GLÀFF, a Large Versatile French Lexicon. In *International Conference on Language Resources and Evaluation (LREC'14)*.

ION R. (2007). *Word sense disambiguation methods applied to English and Romanian*. Thèse de doctorat, Romanian Academy, Bucharest.

KABADJOV M. & STEPANOV E. (2015). *The SENSEI Discourse Analysis Tools, 2*. Rapport interne, University of Essex.

LANDRAGIN F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus*, **10**, 61–80.

LANDRAGIN F. (2015). Sur les aspects multicritères et multidimensionnels de la saillance. In M. BOISSEAU & A. HAMM, Eds., *Saillance. La saillance en langue et en discours, Volume 2*, Annales Littéraires de l'Université de Franche-Comté, p. 15–29. Presses universitaires de Franche-Comté.

LONGO L. (2013). *Vers des moteurs de recherche" intelligents" : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence*. Thèse de doctorat, Université de Strasbourg.

MAHDISOLTANI F., BIEGA J. & SUCHANEK F. (2014). Yago3 : A knowledge base from multilingual wikipedias. In *7th Biennial Conference on Innovative Data Systems Research*.

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL I. & NOUVEL D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues*, **52**(1), 69–96.

MITKOV R. (2002). *Anaphora resolution*. Routledge.

MUZERELLE J., LEFEUVRE A., ANTOINE J.-Y., SCHANG E., MAUREL D., VILLANEAU J. & ESHKOL I. (2013). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *Traitement Automatique des Langues Naturelles (TALN'13)*.

NOUVEL D., ANTOINE J.-Y., FRIBURGER N. & SOULET A. (2011). Recognizing named entities using automatically extracted transduction rules. In *Language & Technology Conference (LTC'11)*.

NOUVEL D., EHRMANN M. & ROSSET S. (2015). *Les entités nommées pour le traitement automatique des langues*. ISTE editions.

POESIO M., STUCKARDT R. & VERSLEY Y. (2016). *Anaphora resolution*. Springer.

SAGOT B. & FIŠER D. (2008). Building a free French wordnet from multilingual resources. OntoLex.

SAJOUS F. & HATHOUT N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, p. 405–426, Herstmonceux, England.

SCHNEDECKER C. (1997). *Nom propre et chaînes de référence*. Metz : Librairie Klincksieck.

SCHNEDECKER C. (2014). Chaînes de référence et variations selon le genre. *Langages*, **195**(3), 23–42.

SCHNEDECKER C. & LANDRAGIN F. (2014). Les chaînes de référence : présentation. *Langages*, **195**(3), 3–22.

TROUILLEUX F. (2001). *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. Thèse de doctorat, Université Blaise-Pascal de Clermont-Ferrand.

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat, Université de Toulouse II le Mirail.

VICTORRI B. (2005). Le calcul de la référence. In P. ENJALBERT, Ed., *Sémantique et traitement automatique des langues* : Hermès.

ZRIBI-HERTZ A. (1996). *L'anaphore et les pronoms : une introduction à la syntaxe générative*. Presses Universitaire du Septentrion.