Cahiers de praxématique

54-55 | 2010
Corpus, données, modèles

Un corpus pour optimiser l'identification automatique des chaînes de référence

A Corpus to Optimize Reference Chains Identification

Laurence Longo

Édition électronique
URL : http://journals.openedition.org/praxematique/1172
DOI : 10.4000/praxematique.1172
ISSN : 2111-5044

Éditeur
Presses universitaires de la Méditerranée

Édition imprimée
Date de publication : 1 janvier 2010
Pagination : 246-262
ISBN : 978-2-36781-012-6
ISSN : 0765-4944

Référence électronique
Laurence Longo, « Un corpus pour optimiser l'identification automatique des chaînes de référence », *Cahiers de praxématique* [En ligne], 54-55 | 2010, document 14, mis en ligne le 01 janvier 2013, consulté le 24 septembre 2020. URL : http://journals.openedition.org/praxematique/1172 ; DOI : https://doi.org/10.4000/praxematique.1172

Cahiers de praxématique 54-55, 2010, 249-262

Laurence Longo
LiLPa, université de Strasbourg

Un corpus pour optimiser l'identification automatique des chaînes de référence

1. Contexte et motivation

L'identification automatique des chaînes de référence (désormais CR) est un problème difficile et ce à plusieurs titres. D'abord, de par la variété des expressions linguistiques intervenant dans la construction des CR et ensuite, par la nécessité d'en détenir des connaissances linguistiques et extra-linguistiques approfondies. Les expressions référentielles (telles que les noms propres, les descriptions définies, les démonstratifs, les pronoms) présentes dans les CR ont fait l'objet de nombreuses études qui se sont focalisées sur leur description linguistique (Corblin, 1985 ; Kleiber, 1994), ou sur leur rôle dans l'organisation du texte (Charolles, 1997).

Les CR constituent des marques linguistiques de la cohérence du texte permettant d'identifier la continuité thématique ou la rupture thématique (Cornish, 1995 ; Goutsos, 1997). Les éléments (ou maillons) d'une CR indiquent le même référent, qui constitue l'élément le plus saillant pour une partie de discours. De plus, le type de CR est une caractéristique du genre textuel (Schnedecker, 2005). Par exemple, les CR introduites par des noms propres sont privilégiées pour les portraits journalistiques. Par là même, la structure des CR et le choix des expressions référentielles qui les constituent sont dépendants du genre textuel.

Dans le domaine du Traitement Automatique des Langues, plusieurs travaux proposent des modèles cognitifs pour le calcul de la référence ou exploitent des indices linguistiques de surface pour des tâches telles que le dialogue homme-machine (Salmon-Alt, 2001), la génération automatique (Manuélian, 2003) ou la détection de thèmes (Hernandez, 2006). Cependant, malgré les nombreux travaux sur la référence,

peu de modèles opérationnels rendent compte des propriétés des CR suivant le genre textuel.

L'identification des CR nécessite d'abord l'identification des relations de coréférence entre les expressions anaphoriques (pronoms, groupes nominaux démonstratifs) et leurs antécédents. Puis, la construction des CR s'effectue en exploitant la transitivité de la relation de coréférence (Ailloud et Klenner, 2009). Les méthodes actuelles de résolution de la coréférence appliquent soit des règles heuristiques définies manuellement (qui sélectionnent les candidats les plus pertinents pour une anaphore donnée), soit des règles apprises à partir de corpus annotés.

Si les méthodes d'apprentissage supervisé (Ng et Cardie, 2002 ; Hoste, 2005) s'avèrent efficaces dans le traitement des relations de coréférence, elles nécessitent des corpus d'apprentissage de grande taille, annotés manuellement. Malgré l'existence de quelques corpus français annotés en relations de coréférence (Manuélian, 2003 ; Salmon-Alt, 2001 ; Tutin, 2002), force est de constater qu'il est difficile d'utiliser ces ressources dans un système d'apprentissage automatique. En effet, ces corpus sont insuffisants en nombre, en taille (Salmon-Alt, 2001) et ils se révèlent hétérogènes quant aux schémas d'annotation adoptés ainsi qu'aux choix des phénomènes annotés (pronoms personnels, descriptions définies, coréférence, anaphores associatives). Ainsi, la tâche#1 de la campagne SemEval 2010 *Coreference Resolution in Multiple Languages* propose-t-elle des données d'apprentissage pour plusieurs langues (anglais, espagnol, catalan, italien, allemand, néerlandais) mais aucune pour le français.

De ce fait, notre travail se situe dans la lignée des méthodes symboliques (Hartrumpf, 2001 ; Bontcheva *et al.*, 2002), facilement adaptables aux nouveaux domaines ou applications, et qui ont aussi prouvé leur efficacité pour la résolution d'anaphores pronominales (Mitkov, 2001). Pour ces systèmes, la relation de coréférence est établie en appliquant des règles qui repèrent les antécédents possibles, après vérification de critères de compatibilité des propriétés (morphosyntaxiques, syntaxiques, sémantiques) des candidats. Pour chaque candidat anaphorique, le choix des antécédents possibles s'effectue selon un calcul de saillance (ou de référence) (Victorri, 2005), la vérification des propriétés de cohérence locale et globale du discours (Grosz *et al.*, 1995), l'application de contraintes (Beaver, 2004) ou un calcul d'accessibi-

lité des expressions référentielles (Ariel, 1990). Dans notre approche, le calcul de référence des antécédents exploite l'échelle d'accessibilité d'Ariel (1990). Suivant le genre textuel traité, le calcul de la référence de chaque antécédent se modifie selon les propriétés spécifiques des CR : préférence pour une catégorie d'expressions référentielles, longueur moyenne de la CR, distance entre les maillons. Les relations de coréférence entre les anaphores et leurs antécédents possibles sont établies après l'application de filtres morphosyntaxiques, syntaxiques et sémantiques sur des textes lemmatisés, étiquetés et annotés au niveau des *chunks* et des entités nommées (noms de personnes, d'organisations, de lieux, etc.).

Le module d'identification des CR est intégré à un outil de détection de thèmes, paramétrable selon le genre textuel du document. Cet outil est développé dans un cadre industriel [1] pour optimiser les résultats des moteurs de recherche et pour faciliter la navigation dans les documents. Les documents sont indexés par les thèmes qu'ils contiennent. Nous utilisons les CR comme marqueurs linguistiques participant à l'identification des thèmes des documents car elles constituent un indice fiable pour révéler les thèmes (Victorri, 1999).

Dans cet article, nous précisons tout d'abord la notion de CR dans notre approche. Nous présentons ensuite notre corpus multi-genres ainsi que les divers traitements linguistiques nécessaires à l'identification automatique des CR (étiquetage, annotation en *chunks*, identification des entités nommées). Nous nous focalisons sur l'étude manuelle des CR effectuée sur notre corpus, en précisant les différents paramètres pris en compte. Nous discutons les résultats obtenus.

2. Les chaînes de référence

En suivant les travaux de Schnedecker (1997), nous considérons qu'une CR est un marqueur linguistique qui inclut au moins trois expressions référentielles (ou maillons) référant à la même entité du discours. Ainsi, l'exemple suivant comporte une seule CR composée de trois maillons (en italique) :

1. Le projet de détection automatique de thèmes s'effectue en collaboration avec l'entreprise R.B.S. (`www.rbs.fr`).

> Moins virulent, *Patrick Devedjian* a *lui* aussi manifesté *son* soutien à Michèle Alliot-Marie.

Les CR comprennent trois types de constituants à fonction référentielle : les noms propres, les groupes nominaux (*défini, indéfini, possessif ou démonstratif*) et les pronoms. Les noms propres jouent un rôle important dans la structuration du discours car ils se trouvent souvent en tête (première mention) d'une CR (Schnedecker, 2005). En dehors des cas de compétition référentielle où la répétition du nom propre s'effectue pour éliminer une ambigüité potentielle entre deux entités, la redénomination d'un nom propre marque une rupture dans la CR (Schnedecker, 1997) ; même s'il s'agit d'un autre point de vue présenté à propos du référent.

Lorsqu'une expression référentielle est utilisée, elle déclenche un « processus de recrutement » particulier du référent exprimé en première mention (Schnedecker, 1997). De ce fait, le groupe nominal démonstratif (par exemple, « *ce* ministre ») renvoie directement au référent sur la base d'un critère de proximité (Kleiber, 1999) alors que le pronom « il » indique qu'il faut recruter un référent qui soit l'argument d'une proposition saillante (Kleiber, 1994). Ainsi, par le choix des expressions référentielles présentes dans une CR, un indice est donné quant au référent à « garder en mémoire » et qui constitue alors le thème local du discours.

Dans notre approche, nous travaillons sur les relations mono référentielles s'établissant entre des expressions coréférentes dans un même paragraphe, par exemple : *« Barack Obama... il... il »*. De plus, nous traitons les situations de coréférence directe (Manuélian, 2003) où les groupes nominaux coréférents possèdent la même tête nominale (par exemple, *« le diplomate français/ce diplomate »*).

Parce que nous utilisons une méthode robuste à base de peu de connaissances (Mitkov, 2001), nous ne traitons pas les cas d'anaphores plurielles, comme dans : *« Barack et Michèle Obama... le couple présidentiel... Michèle »* ; où les deux éléments (« Barack » et « Michèle ») sont d'abord instanciés par le groupe nominal « le couple », puis un des deux noms propres est extrait du groupe initial (« Michèle »). Les cas de coréférence indirecte (hyponymes/hyperonymes) et les cas d'anaphores associatives (relation méro-

nymique, fonctionnelle ou partie/tout) seront considérés dans les futures extensions du système.
Pour notre projet, nous utilisons aussi des listes de noms de fonction qui nous permettent d'établir des liens de coréférence entre les noms de personne et leur fonction, comme par exemple :

> *Barack Obama* a promu lundi l'entrepreneuriat, dans le but de doper la création d'emplois. *Le président des États-Unis* a également dit qu'il demanderait au Congrès de rendre permanentes les exemptions d'impôt sur les revenus du capital pour les P.M.E., et de voter d'autres allégements fiscaux pour ce secteur.

3. Étude des chaînes de référence

Les genres textuels peuvent avoir une influence sur le type des expressions référentielles présentes dans un texte ainsi que sur le choix des diverses mentions du même référent. Pour vérifier cette hypothèse, nous avons constitué un corpus composé de cinq genres textuels. Nous présentons le corpus et les traitements linguistiques nécessaires à la mise en place de notre outil, avant de passer à l'étude de corpus à proprement parler.

Présentation du corpus

Vu la variété des genres de documents susceptibles d'être présents dans les archives internes du moteur de recherche, nous avons choisi d'étudier des extraits (50 000 *tokens* [1]) issus de cinq genres textuels différents, répartis de la manière suivante (voir tableau 1) :
Les articles de journaux du *Monde* traitent de la préparation des divers partis politiques à l'élection présidentielle. Une compétition présidentielle est installée et Valéry Giscard D'Estaing réaffirme le principe d'une candidature de l'UDF. De leur côté, les articles du *Monde diplomatique* abordent la création de deux centres d'étude français — le CERMOC et le CEDEJ — au Proche-Orient. L'extrait issu des *Trois Mousquetaires* relate l'arrivée de d'Artagnan au Bourg de Meung. Le

1. Nous avons travaillé sur des extraits de notre corpus de départ (de 500 000 *tokens*) car l'annotation manuelle des CR s'est révélée être une activité fastidieuse. Nous comptons poursuivre l'annotation de ce corpus de référence vu qu'il n'en existe pas encore pour le français.

Tableau 1. — Répartition du corpus multi-genres

GENRE	SOUS-CORPUS	PERIODE	NOMBRE DE MOTS
Articles de journaux	*Le Monde*	2004	110 012
Éditoriaux	*Le Monde diplomatique*	1980-1988	114 037
Roman	*Les trois Mousquetaires (Dumas)*	1844	105 068
Lois européennes	*Acquis communautaire (Steinberger et al.)*	2006	116 702
Rapports publics	*la Documentation française*	2001	106 765
		TOTAL	552 584

portrait du jeune homme et de sa monture y sont décrits de manière détaillée. Dans les lois européennes issues de l'*Acquis communautaire*, sont abordées les relations à établir entre la Commission européenne et les autorités des États Membres pour que chacune des parties puisse prendre les mesures adéquates en temps voulu. Enfin, les rapports publics de la *Documentation française* portent sur une comparaison de la satisfaction des usagers des services publics et privés à l'égard des produits commercialisés.

Le corpus multi-genres ainsi formé comprend des textes narratifs et des textes non narratifs. De plus, les référents présents dans ce corpus sont des référents humains mais aussi des référents non humains. Ainsi pensons-nous que le corpus est adapté à l'étude des CR. Pour pouvoir mettre en place notre module robuste d'identification automatique des CR, nous étudions les types de CR présents dans chaque genre textuel, après avoir soumis notre corpus à divers traitements.

Étiquetage et segmentation en *chunks*

Pour l'étiquetage de notre corpus brut (en utf-8), nous avons utilisé TTL (Ion, 2007) dans sa version française[1], car il propose un étiquetage fin. En effet, TTL bénéficie du jeu d'étiquettes morphosyntaxiques issu du projet MULTEXT (Ide et Véronis, 1994) qui fournit des informations comme le genre, le nombre, le temps, le mode, la personne. En plus de l'étiquetage morpho-syntaxique (ana), TTL

1. TTL traitait originellement l'anglais et le roumain. Nous avons participé à la constitution des ressources linguistiques nécessaires (corpus d'un million de *tokens*) à l'entraînement de TTL pour le français.

propose une segmentation en *chunks* (ou segments non récursifs) qui permet d'identifier les groupes nominaux simples (`Np`), les groupes adjectivaux (`Ap`) et les groupes prépositionnels (`Pp`). Par exemple, TTL propose le découpage en *chunks* suivant pour « le ministre des affaires étrangères » :

```
<w lemma="le" ana="Da-ms" chunk="Np#1">Le</w>
<w lemma="ministre" ana="Ncms" chunk="Np#1">ministre</w>
<w lemma="de_le" ana="Dg-fp" chunk="Pp#1,Np#2">des</w>
<w lemma="affaire" ana="Ncfp" chunk="Pp#1,Np#2">affaires</w>
<w lemma="étranger" ana="Af-fp" chunk="Pp#1,Np#2,Ap#1">étrangères</w>
```

Figure 1. — Découpage en *chunks* pour « le ministre des affaires étrangères ».

Partant de la sortie XML, nous appliquons une base de patrons symboliques pour identifier des expressions plus complexes susceptibles d'être présentes dans les CR (car plus informatives) : les groupes nominaux complexes (CNp [1]) (par exemple, *« les modifications liées au changement climatique »*, *« la candidate à la primaire socialiste »*) et les entités nommées (nom de fonction, d'organisation ou de personne). Les emplois impersonnels du pronom *« il »* sont aussi annotés (par exemple *« il neige »*), afin d'ignorer ces emplois dans le calcul des CR. Voici un exemple d'annotations comprenant les lemmes, chunks simples et complexes (`CNp`), propriétés morpho-syntaxiques (`ana`), entités nommées (`NER`, avec précision du type : organisation (`org`), personne (`pers`), fonction (`func`)) et il impersonnel (`imp`) (voir figure 2). Nous exploitons ces annotations automatiques pour construire les CR.

Étude en corpus des chaînes de référence

Pour cette étude, nous avons annoté manuellement les CR pour déterminer les propriétés des CR pertinentes pour un genre particulier [2].

1. Un groupe nominal complexe (CNp) est un groupe nominal modifié par deux groupes prépositionnels au plus, ou bien un groupe nominal modifié par une proposition relative (une proposition simple contenant un prédicat et un complément d'objet direct ou indirect).

2. Nous rappelons que nous avons mené notre étude sur des extraits (50 000 *tokens*) de notre corpus multi genres.

```
<w lemma="le" chunk="Np#1" ana="Da-fs">L'</w>
<w lemma ="union" chunk="Np#1" ana="Ncfs" ner="NER#1, org">Union</w>
<w lemma ="européen" chunk="Np#1, Ap#1" ana="Af-fs" ner="NER#1, org">européenne</w>
<w lemma ="avoir" chunk="Vp#1" ana="Vaip3s">a</w>
<w lemma ="adopter" chunk="Vp#1" ana="Vmps-s">adopté</w>
<w lemma ="il" ana="Pp3ms" feat="imp">il</w>
<w lemma ="y" ana="Pp3">y</w>
<w lemma ="avoir" ana="Vaip3s">a</w>
<w lemma ="peu" chunk="Ap#2" ana="R">peu</w>
<w lemma ="de_le" chunk="CNp#5, Pp#1, Np#2" ana="Dg-mp">des</w>
<w lemma ="acte" chunk="CNp#5, Pp#1, Np#2" ana="Ncmp">actes</w>
<w lemma ="législatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" ana="Af-mp">législatifs</w>
<w lemma ="relatif" chunk="CNp#5, Pp#1, Np#2, Ap#3" ana="Af-mp">relatifs</w>
<w lemma ="à+le" chunk="CNp#5, Pp#2, Np#3" ana="Dg-ms">au</w>
<w lemma ="changement" chunk="CNp#5, Pp#2, Np#3" ana="Ncms">changement</w>
<w lemma ="climatique" chunk="CNp#5, Pp#2, Np#3, Ap#4" ana="Af-ms">climatique</w>
```

Figure 2. — Exemple de sortie enrichie en annotations pour « l'Union européenne a adopté, il y a peu, des actes législatifs relatifs au changement climatique ».

L'étude des CR que nous avons effectuée est basée sur les travaux de (Schnedecker, 2005). Ainsi, pour chaque genre, nous examinons les CR suivant cinq critères :

a) *La longueur moyenne des CR* : est comptabilisé le nombre moyen de maillons contenus dans une CR suivant le genre (ne sont comptabilisées que les CR de trois maillons au moins).

L'étude a révélé quelques différences (voir tableau 2). Par exemple, nous avons constaté que la longueur des CR était de trois maillons en moyenne pour l'*Acquis Communautaire* (« un État Membre... cet État Membre... l'État Membre ») alors qu'elle était trois fois plus élevée en moyenne pour *Les Trois Mousquetaires* ; par exemple : *« un bidet du Béarn... ses (huit lieues)... ce cheval... son (poil étrange)... son (allure incongrue)... du susdit bidet... il... il... son (cavalier) »*.

Cette différence significative entre la longueur des CR de ces deux genres s'explique par le fait que les lois européennes font intervenir de nombreux référents, ce qui crée une compétition référentielle forte. En cas de compétition référentielle, la redénomination d'un nom propre est considérée comme une fermeture du référent en cours (donc une ouverture d'une nouvelle CR) (Schnedecker, 2005), d'où le faible nombre de maillon relevé pour ce type de CR. En revanche, dans l'extrait des *Trois Mousquetaires*, on relève de nombreux passages descriptifs propices aux CR longues.

Tableau 2. — Longueur moyenne (en nombre de maillons) des CR suivant le genre textuel

Corpus	Longueur moyenne
Le Monde	4
Le Monde diplomatique	3,7
Acquis communautaire	3
Les Trois Mousquetaires	9
La Documentation française	3,4

b) *La distance moyenne entre les maillons des CR* : nous avons déterminé ici le nombre de phrases séparant chacun des maillons d'une même CR (0 pour la même phrase, 1 pour la phrase suivante).
Pour ce second critère, nous avons observé que la distance entre les maillons des CR du *Monde* n'excède pas une phrase en moyenne, tandis que pour la *Documentation française*, la distance est supérieure à deux phrases entre le second maillon et le troisième. Pour ce dernier genre, on retrouve une technique séquentielle particulière : introduction du référent, maintien et rappel avant la fermeture de la CR (Goutsos, 1997).

Tableau 3. — Distance moyenne entre les maillons (en nombre de phrases) suivant le genre textuel [1]

CORPUS	RANG DU MAILLON			
	1à 2	2 à 3	3 à 4	4 à 5
Le Monde	0,4	1	1	-
Le Monde diplomatique	0,3	1,3	0	2
Acquis communautaire	0	1,3	-	-
Les Trois Mousquetaires	0	0,3	0	1,3
La Documentation française	0,4	2,4	0,5	-

1. Par souci de lisibilité, ne sont reportés dans ce tableau que les maillons des rangs 1 à 5 (même si les CR des *Trois Mousquetaires* ont une longueur moyenne de neuf maillons).

La catégorie grammaticale privilégiée dans l'ensemble des maillons des CR suivant le genre : On constate (Tableau 4) une part importante de noms propres (30,8%) dans les maillons des CR du *Monde*, alors que cette catégorie n'est pas représentée dans les autres genres textuels [1]. La moitié des maillons des CR du *Monde Diplomatique* sont des groupes nominaux définis (GNdef) alors que 40% des maillons sont des groupes nominaux indéfinis (GNindef) dans l'*Acquis communautaire*. Cette dernière observation pour les lois européennes est corrélée au faible nombre de maillons relevés en moyenne dans les CR (critère 1). En effet, les mesures décidées par la Commission européenne ont un caractère générique qui doit s'appliquer à tout État Membre de la Communauté ; d'où la présence massive des indéfinis (on aura par exemple : « un État Membre », « une décision »).

Tableau 4. — Répartition des catégories grammaticales des maillons des CR suivant le genre (en %)

CORPUS	CATÉGORIE GRAMMATICALE DES MAILLONS					
	Np	Pr	GNdef	GNindef	poss	dem
Le Monde	30,8	15,4	23,1	0	23,1	7,7
Le Monde diplomatique	0	25	50	0	25	0
Acquis communautaire	0	10	20	40	10	20
Les Trois Mousquetaires	0	35,9	20,5	10,3	28,2	5,1
La Documentation française	0	33,3	33,3	16,7	16,7	0

Les disparités observables entre les fréquences des catégories des maillons présentes dans chaque genre textuel sont, dans une certaine mesure, révélatrices des spécificités des CR suivant le genre.

c) *La classe grammaticale des premiers maillons des CR suivant le genre* : Nous nous sommes intéressés ici à définir la catégorie privilégiée des maillons utilisés en première mention dans chacun des genres. Pour les articles du *Monde*, on relève essentiellement des noms propres en première mention. Ce sont plutôt des descriptions définies qui se

1. Notons que plusieurs noms propres étaient bien présents dans les autres genres textuels étudiés, mais que seuls les noms propres figurant dans des CR de trois maillons au moins (la définition des CR que nous avons adoptée) ont été comptabilisés et reportés dans le tableau 4.

retrouvent souvent en position de premier maillon des CR du *Monde diplomatique*. En effet, parce que les sujets abordés dans les éditoriaux concernent un point de vue à propos des actualités du moment, les auteurs considèrent que les références aux entités présentes dans leurs écrits sont acquises par leurs lecteurs.

Les groupes nominaux indéfinis dominent dans les premiers maillons des CR de l'*Acquis communautaire*. Enfin, de même que pour les éditoriaux, *La Documentation française* compte en majeure partie des GN définis dans les premiers maillons des CR (« la satisfaction des clients », « la mesure de la satisfaction des usagers »).

d) *La correspondance entre le premier maillon d'une CR et le thème phrastique (élément préverbal)* : Pour ce dernier critère, nous souhaitions savoir dans quelle mesure il était possible de regrouper les CR contenant le même thème phrastique. Nous avons donc comptabilisé les cas où le premier maillon des CR coïncidait avec le thème phrastique. On observe ainsi que le premier maillon est le thème phrastique dans 80 % des cas pour les articles du *Monde* par exemple, mais qu'il n'est que de l'ordre de 40 % pour les rapports issus de la *Documentation française*.

Tableau 5. — Correspondance entre le premier maillon des CR et le thème phrastique (en %)

CORPUS	CORRESPONDANCE
Le Monde	80
Le Monde diplomatique	100
Acquis communautaire	60
Les Trois Mousquetaires	60
La Documentation française	40
Moyenne	68

Ainsi, l'étude des CR dans un corpus multi-genres a-t-elle permis de mettre au jour leurs propriétés spécifiques suivant le genre textuel. Ces paramètres sont utilisés pour configurer notre module d'identification des CR selon le genre textuel. Par exemple, si le document à traiter est un article de journal, notre système sera paramétré pour identifier des CR courtes (d'une longueur moyenne de quatre maillons), qui débuteront de préférence par un nom propre et dont les maillons seront compris dans la phrase en cours ou dans la phrase suivante. Le nom propre coïncidera souvent avec le thème phrastique, ce qui signifie que plusieurs CR indiqueront le même thème.

Conclusion et perspectives

L'étude des CR dans un corpus multi-genres a permis d'établir une comparaison de ces CR suivant cinq critères (longueur des CR, distance intermaillonnaire, catégorie grammaticale des premiers maillons, nature des premières mentions, correspondance entre le premier maillon d'une CR et le thème phrastique). À l'issue de cette étude, nous avons pu dégager des spécificités des CR suivant le genre textuel. Ces propriétés spécifiques des CR ont rendue possible l'adaptation du calcul de la référence suivant le genre textuel. Notre module a été évalué dans (Longo et Todirascu, 2010 [1]). La prise en compte des paramètres spécifiques au genre textuel dans le calcul de la référence que nous avons mis en place a amélioré les performances de 12 %. Nous projetons de traiter d'autres genres textuels et d'étendre notre étude aux anaphores plurielles.

Références bibliographiques

Ailloud É. & Klenner M., 2009, *Towards More Linguistically Constrained Coreference Resolution, Actes de TALN*, Senlis.

Ariel M., 1990, *Accessing Noun-Phrase Antecedents*, London, Routledge.

Beaver D., 2004, « The Optimization of Discourse Anaphora », *Linguistics and Philosophy*, 27 (1), 3-56.

1. L'algorithme et les paramètres pris en compte dans le calcul de la référence y sont détaillés et illustrés.

Bontcheva K., Dimitrov M., Maynard D., Tablan V. & Cunningham H., 2002, *Shallow Methods for Named Entity Coreference Resolution, Actes de TALN*, Nancy, France.

Charolles M., 1997, « L'encadrement du discours : univers, champs, domaines et espaces », *Cahier de Recherche Linguistique 6*, LANDISCO, université Nancy 2, 1-73.

Corblin F., 1995, *Les formes de reprise dans le discours : Anaphores et chaînes de référence*, Presses universitaires de Rennes.

Cornish F., 1995, « Références anaphoriques, références déictiques, et contexte prédicatif et énonciatif », *Sémiotiques* 8, 31-57.

Goutsos D., 1997, *Modeling Discourse Topic : Sequential Relations and Strategies in Expository Text*, Norwood, N.J., Ablex Publishing Corporation.

Grosz B.J., Weinstein S. & Joshi A.K., 1995, « Centering : a framework for modeling the local coherence of discourse », *Computational Linguistics* 21 (2), 203-225.

Hartrumpf S., 2001, « Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics », *Actes de CoNLL (Computational Natural Language Learning Workshop)*, Toulouse, France, 137-144.

Hernandez N., 2004, *Description et détection automatique de structures de texte*, thèse de doctorat, université Paris-Sud XI.

Hoste V., 2005, *Optimization Issues in Machine Learning of Coreference Resolution*, Thèse de doctorat.

Ide N. & Véronis J., 1994, *MULTEXT (Multilingual Tools and Corpora), Actes de ICCL*, Kyoto.

Ion R., 2007, TTL : *A Portable Framework for Tokenization, Tagging and Lemmatization of Large Corpora*, PhD thesis progress report, Research Institute for Artificial Intelligence, Romanian Academy, Bucharest (in Romanian).

Kleiber G., 1994, *Anaphores et pronoms*, Louvain-la-Neuve, Duculot.

Kleiber G., 1999, *Problèmes de sémantique. La polysémie en questions*, Lille, éd. du Septentrion.

Longo L. & Todiraşcu, A., 2010, « Genre-based Reference Chains Identification for French », *Investigationes Linguisticae* XXI, 57-75.

Manuélian H., 2003, *Descriptions définies et démonstratives : analyse de corpus pour la génération de textes*, thèse de doctorat, université de Nancy 2.

Mitkov R., 2001, « Towards a More Consistent and Comprehensive Evaluation of Anaphora Resolution Algorithms and Systems », *Applied Artificial Intelligence : An International Journal* 15, 253-276.

Ng V. & Cardie C., 2002, « Improving Machine Learning Approaches to Coreference Resolution », *Actes de ACL*, Morristown, 104-111.

Salmon-Alt S., 2001, *Référence et Dialogue finalisé : de la linguistique à un modèle opérationnel*, thèse de doctorat, université H. Poincaré, Nancy.

Schnedecker C., 1997, *Nom propre et chaînes de référence, Recherches Linguistiques* 21, Paris, Klincksieck.

Schnedecker C., 2005, « Les chaînes de référence dans les portraits journalistiques : éléments de description », *Travaux de Linguistique*, 51, 85-133.

Steinberger R., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufis D. & Varga D., 2006, « The JRC-Acquis : A Multilingual Aligned Parallel Corpus with 20 + Languages », *Actes de LREC*, 2142-2147.

Tutin A., 2002, « A Corpus-Based Study of Pronominal Anaphoric Expressions in French », *Actes de DAARC (Discourse Anaphora and Anaphora Resolution)*, Lisbonne, Portugal, 18-20 septembre.

Victorri B., 1999, « Traitement automatique des langues et recherche documentaire », *Revue d'interaction Homme-Machine*, 2, 25-36.

Victorri B., 2005, *Le calcul de la référence*, in Enjalbert P. (éd.), *Sémantique et traitement automatique des langues*, Hermès, 133-172.