Linx
Revue des linguistes de l'université Paris X Nanterre

80 | 2020
L'héritage de Jean Dubois et Françoise Dubois-Charlier

Extraction d'informations sémantiques dans le DEM et le LVF

Extracting semantic information from J. Dubois and F. Dubois-Charlier's DEM and LVF

Elisabeth Godbert

Édition électronique
URL : http://journals.openedition.org/linx/5956
DOI : 10.4000/linx.5956
ISSN : 2118-9692

Éditeur
Presses universitaires de Paris Nanterre

Référence électronique
Elisabeth Godbert, « Extraction d'informations sémantiques dans le DEM et le LVF », *Linx* [En ligne], 80 | 2020, mis en ligne le 10 juillet 2020, consulté le 05 août 2020. URL : http://journals.openedition.org/linx/5956

Département de Sciences du langage, Université Paris Ouest

Extraction d'informations sémantiques dans le DEM et le LVF

Extracting semantic information from J. Dubois and F. Dubois-Charlier's DEM and LVF

Elisabeth Godbert

NOTE DE L'AUTEUR

Ce travail a été financé par l'Agence Nationale pour la Recherche au sein des projets suivants : ANR-15-CE23-0003 (DATCHA), ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) et ANR-11-IDEX-0001-02 (A*MIDEX).

Je suis très reconnaissante à Frédéric Béchet, Benoit Favre et Alexis Nasr pour les conseils et pour l'aide qu'ils m'ont apportée pour ce travail.

1. Introduction

1 Nous nous intéressons ici à l'acquisition de données sémantiques, notre objectif étant de les utiliser pour la détection automatique de relations de coréférence intra- et inter-phrases en français, plus spécifiquement pour le traitement des anaphores pronominales et des coréférences directes.

2 La résolution des anaphores participe à l'interprétation sémantique des textes : elle met en relation les expressions qui font référence à une même entité du discours. Les expressions référentielles, également appelées mentions, sont les groupes nominaux et les pronoms, hors pronoms réfléchis. Dans tout ce qui suit, pour ce qui concerne les pronoms, nous ne nous intéressons qu'aux pronoms à la troisième personne.

3 Considérons les exemples ci-dessous :

> Un navire de croisière est arrivé dans le port ce matin, veux-tu y aller pour le voir ?

C'est un bateau gigantesque. Les passagers vont se régaler, ils partent pour les îles grecques.
J'ai perdu ma carte de bus, que puis-je faire ? J'utilise cette carte tous les jours.

4 Les mentions peuvent être liées par différents types de relations de coréférence. Nous utilisons la terminologie définie dans Muzerelle *et al.* (2013) :

- Anaphore pronominale le navire ... le ; les passagers ... ils
- Relation directe ma carte ... cette carte ; (même tête nominale)
- relation indirecte un navire ... un bateau
- relation associative un bateau ... les passagers

5 On utilise classiquement, pour identifier les éventuelles mentions coréférentes, le genre et le nombre, la position syntaxique, le focus, la distance des mots et des informations sémantiques (Lee *et al.* 2013, Poesio *et al.* 2010).

6 Les systèmes développés pour traiter les anaphores portent en majorité sur l'anglais, on en trouve une description détaillée dans Poesio *et al.* (2010). Dans les systèmes développés pour le français, on peut citer les travaux de F. Trouilleux (2001) qui traite les anaphores pronominales et le système RefGen de L. Longo (Longo *et al.* 2010). La sémantique des entités dont on parle est un trait important à prendre en compte. Considérons les exemples suivants :

(1) Mon fils a eu un problème avec son bus ce matin ; il est passé avec 30 minutes de retard. Il était très ennuyé car il est sérieux.
(2) Mon fils a eu un problème avec son bus ce matin ; il est arrivé en retard au collège, il a été puni.
(3a) J'ai perdu ma carte de bus, que puis-je faire ? J'utilise cette carte tous les jours.
(3b) Allez dans une agence RATP, ils vous la referont comme elle était.
(3c) Oui mais je ne sais pas où elles sont en fait.

7 En (1) on trouve trois anaphores pronominales « *il* ». Si l'on ne considère que les traits de genre et de nombre, on identifie quatre antécédents potentiels : *fils, problème, bus* et *matin*. Il en va de même pour l'exemple (2).

8 Si l'on sait identifier quelles expressions dénotent une personne ou un groupe de personnes, et lesquelles dénotent une entité inanimée, certaines anaphores pronominales pourront être résolues facilement : par exemple dans (2) l'objet de *punir* ne peut être qu'un humain ou un animal, et dans (3b) l'objet de *refaire* doit être une entité inanimée ; dans (1) les adjectifs *ennuyé* et *sérieux* ne peuvent se rattacher qu'à un humain. Par contre, dans (1) le verbe *passer* peut aussi bien avoir comme actant sujet le *fils* ou le *bus*.

9 Dans (3b) et (3c), le rattachement pronominal de *ils* et *elles* à *l'agence* ne peut se faire que via la sémantique puisqu'il n'y a pas conservation du genre et du nombre.

10 Il serait donc pertinent de disposer d'un typage sémantique des expressions nominales, des adjectifs et des verbes, et d'un typage de chaque pronom dans le contexte où il apparaît.

11 Le traitement automatique des phrases au niveau sémantique est difficile à mettre en œuvre. Pour acquérir des données sémantiques, la plupart des systèmes existants utilisent WordNet ; d'autres travaux mettent en œuvre des processus d'apprentissage sur des N-grams du Web, ou sur le système de catégories de Wikipedia, ou encore sur des corpus annotés manuellement, mais la fiabilité des résultats d'apprentissage n'est pas totale, et la couverture lexicale des rares corpus annotés est assez faible (Poesio *et al.* 2010).

12 Dans le but d'élaborer pour le français un système de traitement des anaphores qui ait une bonne couverture lexicale, nous faisons le choix de rechercher des données sémantiques dans les ressources élaborées par Jean Dubois et Françoise Dubois-Charlier, « Dictionnaire électronique des mots » (DEM) (Dubois & Dubois-Charlier 2010) et « Les Verbes Français » (LVF) (Dubois & Dubois-Charlier 1997) : en effet, pour le français, ces dictionnaires ont une très large couverture lexicale et les données peuvent en être extraites directement, sans passer par de l'apprentissage. Les corpus sur lesquels nous travaillons sont des corpus de dialogues oraux transcrits ou de dialogues en tchat :

13 1. Le corpus RATP-DECODA est composé de 2100 dialogues oraux enregistrés dans le Centre d'appel de la RATP (environ 560 000 mots) (Bechet *et al.* 2012).

14 2. Le corpus de tchat Orange contient des discussions en tchat entre le Centre de l'opérateur Orange et des clients (environ 95 000 mots)

15 3. Le corpus ANCOR, disponible librement, est constitué d'un ensemble de 460 dialogues. Ce corpus est pour nous une ressource très intéressante car les dialogues transcrits sont annotés manuellement en mentions et en relations de coréférence et relations anaphoriques. Il contient environ 500 000 mots au total et 115 600 mentions (Muzerelle *et al.* 2013).

16 Dans les dialogues, les relations anaphoriques pronominales sont nombreuses. Par exemple, sur le corpus ANCOR, une étude distributionnelle sur les mentions a compté un total de 115 672 mentions tous types confondus, dont 51,1 % de mentions pronominales et 48,9 % de mentions nominales (Désoyer *et al.* 2015).

17 Nous décrivons dans ce qui suit comment, à partir du DEM et du LVF, nous produisons des annotations sémantiques et comment celles-ci sont prises en compte dans le traitement des anaphores. Nous montrons pour finir comment nous avons utilisé le corpus ANCOR pour faire une évaluation automatique de notre système.

2. Annotations préalables en lexique, syntaxe et mentions

18 En préalable à l'annotation en sémantique et à la recherche de coréférences, les corpus sur lesquels nous travaillons sont annotés en lexique, syntaxe et mentions.

19 L'annotation en lexique et syntaxe est effectuée par le système MACAON, chaîne de traitement linguistique séquentielle disponible librement (Nasr *et al.* 2011). Composé de plusieurs modules, MACAON réalise sur une entrée textuelle le découpage en mots, l'étiquetage morpho-syntaxique, la lemmatisation, l'analyse morphologique, et l'analyse syntaxique partielle en dépendances.

20 L'annotation en mentions est ensuite faite par apprentissage sur le corpus ANCOR mentionné ci-dessus, et sur la base des annotations en lexique et syntaxe (Godbert & Favre 2017). L'exemple ci-dessous donne un extrait de dialogue du corpus RATP-DECODA, annoté en lexique et syntaxe par la chaîne MACAON et annoté ensuite en mentions par le modèle d'apprentissage sur les mentions.

Tableau 1

id	*num*	*mot*	*catég.*	*lemme*	*traits*	*dépendance*	*gouv.*	*mention*
38	11	aller	vinf	aller	W#####	OBJ	37	O
39	12	dans	prep	dans	#####	P_OBJ_LOC	38	O
40	13	une	det	un	##f#s##i	NULL	0	B-m
41	14	agence	nc	agence	##f#s##	NULL	0	I-m
42	15	euh	pres	euh	#####	NULL	0	O
43	16	une	det	un	##f#s##i	DET	44	B-m
44	17	agence	nc	agence	##f#s##	OBJ	39	I-m
45	18	RATP	np	RATP	##f#s##	MOD	44	I-m
46	19	pour	prep	pour	#####	MOD	44	O
47	20	la	clo	cla	#3#f#s##	OBJ	48	B-m
48	21	faire	vinf	faire	W#####	OBJ	46	O
49	22	refaire	vinf	refaire	W#####	OBJ	48	O
50	23	qu'	csu	qu'	#####	OBJ	49	O
51	24	on	cln	cln	#3##s##	SUJ_IMP	54	B-m
52	25	vous	clo	cln	#2##p##	A_OBJ	54	B-m
53	26	en	clo	en	#####	DE_OBJ	54	B-m
54	27	donne	v	donner	PS#13##s##	OBJ	50	O
55	28	une	det	un	##f#s##i	DET	56	B-m
56	29	autre	adj	autre	###s##	OBJ	54	I-m
…	…	…	…	…	…	…	…	…
78	7	ils	cln	cln	#3#m#p##	SUJ	79	B-m
79	8	vont	v	aller	P#3##p##	ROOT	0	O
80	9	vous	clo	cln	#2##p##	A_OBJ	82	B-m
81	10	la	clo	cla	#3#f#s##	OBJ	82	B-m
82	11	refaire	vinf	refaire	W#####	OBJ	79	O
83	12	comme	adv	comme	#####	MOD	82	O
84	13	elle	cln	cln	#3#f#s##	SUJ	85	B-m
…	…	…	…	…	…	…	…	…

3. Typage sémantique des noms à partir du DEM

21 « Le Dictionnaire électronique des Mots » (DEM) est une base de données qui répertorie tous les mots du français et qui est disponible librement. Cette base a été développée par Jean Dubois et Françoise Dubois-Charlier (Dubois & Dubois-Charlier 2010). On y trouve 145 198 entrées, avec pour chaque entrée son sens, des propriétés catégorielles, morphologiques, sémantiques, syntaxiques. Chaque mot peut y avoir plusieurs entrées, qui correspondent à ses différents sens. Pour extraire du DEM des informations sémantiques sur les noms, nous avons procédé ainsi :

22 1. Nous avons extrait du DEM tous les noms communs qui y sont présents, et choisi de ne garder que les deux premiers sens de chaque nom, pour ne pas trop en brouiller la sémantique.

23 2. Pour chaque nom, nous avons extrait dans les champs « Catégorie grammaticale » (CA), « Domaine » (DOM) et « Opérateur » (OP) des informations sémantiques qui donnent :

24 - son appartenance à l'une des cinq classes : *Tout, Animé, NonAnimé, Humain, Animal* ;

25 - des informations sur le ou les domaines sémantiques dans lesquels ce nom est utilisé.

26 3. Ensuite, plusieurs opérations de filtrage et de croisement de ces informations ont permis d'attribuer à chaque nom une entrée unique.

27 Les cinq classes de base sont *Tout, Animé, NonAnimé, Humain, Animal* ; *Tout* est la réunion des classes *Animé* et *NonAnimé*, *Animé* est la réunion des classes *Humain* et *Animal*. Quelques autres classes ont été définies, par exemple *Véhicule*, *Objet-Concret, Quantité*...

Les ensembles de noms qui y ont été intégrés l'ont été à partir d'informations extraites dans le champ DOM.

28 Pour les noms polysémiques, on utilise si nécessaire la réunion de deux classes. Après l'ajout de quelques noms communs spécifiques à nos corpus d'application (dialogues oraux retranscrits ou tchat), nous avons obtenu une table de 87 037 entrées où chaque nom a une entrée unique, avec sa classe et ses domaines de rattachement. Voici ci-dessous un extrait de la table du typage des noms (LOC = locatif/lieu ; LAN = langue ; LIT = littérature).

Tableau 2

nom	*Classe*	*domaine*
abeille	Tout-Animé-Animal	ENT
acacia	Tout-NonAnimé	SYL
acadien	Tout	LOC, LAN
académicien	Tout-Animé-Humain	LIT

29 Par ailleurs, nous avons établi (en grande partie manuellement) une liste de noms propres, où chacun est associé à une classe sémantique : villes, pays, prénoms, etc. Cette liste contient entre autres les noms propres de nos corpus d'application. Nous obtenons finalement une classification sémantique de ces noms dans une taxinomie d'une quinzaine de classes.

4. Typage sémantique des verbes à partir du LVF

30 « Les Verbes Français » (LVF) est un dictionnaire électronique développé par Jean Dubois et Françoise Dubois-Charlier (Dubois & Dubois-Charlier, 1997) et disponible librement.

31 Il contient 25 610 entrées verbales représentant 12 310 verbes différents, dont 4 188 à plusieurs entrées, et donne pour chaque verbe de nombreuses informations, dont en particulier ses constructions syntaxiques accompagnées de la nature de ses actants, à l'interface syntaxe-sémantique. La classification des verbes du LVF repose sur l'hypothèse « qu'il y a adéquation entre les schèmes syntaxiques de la langue et l'interprétation sémantique qu'en font les locuteurs de cette langue ». Ces informations nous fournissent des éléments très pertinents pour effectuer un typage sémantique des verbes. En particulier, pour chaque entrée verbale :

32 - sa « classe » code la classe sémantique à laquelle appartient cette entrée verbale, par exemple « verbe de communication », « verbe de mouvement », etc. Il existe 54 classes, elles-mêmes découpées en sous-classes et sous-types.

33 - la « syntaxe du verbe », donne le nombre d'actants ou compléments du verbe et la nature de chaque actant : humain, animal, chose, complétive, etc.

34 Pour extraire du LVF des informations sémantiques sur les verbes, nous avons procédé comme nous l'avons fait dans le DEM pour les noms.

35 Nous avons extrait un ensemble d'informations syntactico-sémantiques dans les champs OPERATEUR (OP) et CONSTRUCTION (CONST). Plusieurs opérations de filtrage de ces informations, puis de synthèse, nous ont permis d'obtenir pour chaque verbe un

typage sémantique des actants sujet et complément d'objet direct (Godbert 2014). Nous ne gardons qu'une entrée pour chaque verbe, qui couvre les usages les plus courants de ce verbe. Nous avons ainsi obtenu une table de 12 484 verbes, où chacun est associé au type de ses actants sujet et objet.

36 Les trois types de base des actants sont *Animé, Non Animé, Tout.* Pour les verbes de mouvement et locatifs (*monter, descendre, avancer, passer*, ...) une classe composée a été définie : *Animé-ou-Véhicule* (ces verbes de mouvement sont repérés dans LVF à l'aide du champ CLASSE). Nous avons obtenu une table de typage des verbes qui contient 12 484 entrées. Voici ci-dessous quelques extraits de cette table :

Tableau 3

Verbe	*sujet*	*compl-objet-direct*
atrophier	NonAnimé	
attabler	Animé	Animé
attacher	Animé	Tout
attarder	Animé-ou-Véhicule	
atteindre	Animé-ou-Véhicule	Tout
atteler	Animé	Tout
attendre	Animé-ou-Véhicule	Tout

5. Typage sémantique des adjectifs à partir du DEM

37 De manière similaire, nous avons extrait du DEM des informations sémantiques qui nous ont permis d'établir un typage des adjectifs, pour lesquels nous avons obtenu une table de 33 008 entrées. Voici ci-dessous un extrait de la table de typage des adjectifs.

Tableau 4

adjectif	*classe*	*domaine*
abolitionniste	Tout-Animé-Humain	DRO
abominable	Tout-NonAnimé	PSY
abrogeable	Tout-NonAnimé	DRO
abrupt	Tout	GEG, SOC

6. Détection automatique d'anaphores pronominales

38 Dans chaque texte dans lequel nous cherchons à détecter les coréférences, le premier travail effectué par notre système est l'identification automatique, via les dépendances, des pronoms utilisés dans des formes impersonnelles : *il y a, il faut, il est huit heures...* ou des pronoms qui reprennent un segment phrastique : *- Louis est arrivé. - Oui, je le sais*. Ces pronoms sont ignorés dans les traitements qui suivent.

39 Ensuite, pour chaque pronom en position d'actant (sujet ou objet) d'un verbe ou d'un adjectif (attribut), le système analyse les dépendances, qui lui permettent de trouver le verbe ou l'adjectif auquel le pronom est lié ; le système en déduit le type sémantique du

pronom, à partir des tables de typage des verbes et des adjectifs décrites dans les parties 4 et 5 ci-dessus.

40 Pour les pronoms en position de complément indirect ou prépositionnel, le type NonAnimé est affecté à *en* et *y*, et le type Animé est affecté à *lui, elle, eux, elles*.

41 Ensuite, pour chaque pronom le système cherche dans le texte qui précède la mention la plus proche qui respecte des contraintes de genre, nombre et type sémantique, ainsi que des éléments de la théorie du liage (Chomsky 1981). La relation de coréférence peut être établie avec un nom qui est l'antécédent du pronom, ou un autre pronom qui fait référence à la même entité du discours.

42 La recherche suit des règles que nous avons définies, en plusieurs passes, chaque passe remontant plus ou moins loin dans le dialogue et étant plus ou moins contrainte : dans les premières passes on impose un accord sur le genre, le nombre et le type sémantique, dans les suivantes on assouplit progressivement ces contraintes. Cette méthode est semblable aux méthodes classiques qui utilisent des filtres linguistiques pour écarter des candidats, et un ordonnanceur qui met en œuvre des préférences. Plus précisément, les principales étapes sont les suivantes :

43 1. On recherche une entité coréférente sous la forme d'un pronom de type sémantique compatible, en ne remontant que très peu dans le dialogue : 20 mots ; si la recherche est fructueuse, les deux pronoms sont notés « coréférents ».

44 2. En cas d'échec de l'étape 1, on recherche un nom de type sémantique compatible, de même genre et de même nombre que le pronom, en ne remontant que 30 mots ; si la recherche est fructueuse, le nom est « l'antécédent » du pronom.

45 3. En cas d'échec du 2, on assouplit peu à peu (sur plusieurs étapes) les contraintes sur le nom, pour finalement ne garder que la contrainte sur le type sémantique (on remonte encore à 30 mots)

46 4. Si l'étape 3 est infructueuse, on remonte beaucoup plus haut (100 mots) en recherchant un nom de même type sémantique, de même genre et de même nombre que le pronom.

47 Si le système n'a trouvé aucune entité acceptable comme antécédent ou coréférent du pronom, il indique l'échec de sa recherche pour ce pronom. En particulier, la recherche échoue :

48 a) lorsque l'antécédent existe mais est trop éloigné dans le dialogue ;

49 b) lorsqu'il n'y a ni antécédent ni pronom coréférent, comme pour le premier *ils* dans :

> - il y a un préavis de grève aujourd'hui ;
> - oh, ils nous cassent les pieds ; je sais que vous n'y êtes pour rien mais ils nous cassent les pieds .

50 Ici, les prédictions du système sont :

51 - le premier *ils* n'a ni antécédent ni coréférent ;

52 - le deuxième *ils* n'a pas d'antécédent non plus, mais il est noté coréférent du premier.

53 Le cas des cataphores est très particulier *(Il a été sympa le chauffeur)* : si l'annotation syntaxique en dépendances est correcte, la relation de coréférence entre le pronom et son coréférent est directement tirée de l'analyse syntaxique en dépendances : les deux mots ont même étiquette et même gouverneur.

7. Détection de coréférences nominales

54 Le système effectue une recherche de coréférents pour chaque expression nominale définie construite sur un nom commun ou un nom propre. Ici encore, on cherche dans le texte qui précède la mention la plus proche qui respecte les contraintes du système.

55 Les expressions définies sont de deux types : celles qui introduisent une nouvelle entité et celles qui font référence à une entité déjà introduite.

56 La recherche ne porte que sur les coréférences directes entre expressions de même tête lexicale (*ma carte ... cette carte*). La recherche se fait en considérant, outre la tête nominale, la présence éventuelle de modifieurs, pour ne pas mettre en relation par exemple *ma nouvelle carte* et *mon ancienne carte*. Mais il faudrait compléter notre travail à ce niveau car il faudrait savoir a priori quels modifieurs sont compatibles et lesquels ne le sont pas. Dans l'état actuel de notre travail, nous considérons que deux expressions peuvent être mises en relation si l'ensemble des modifieurs de l'une est inclus dans l'ensemble des modifieurs de l'autre ; le cas le plus simple étant que l'une des deux expressions n'ait pas de modifieur (*ma carte ... cette nouvelle carte*).

8. Résultats

8.1. Un exemple

57 Ci-dessous un exemple de résultats obtenus sur un extrait de dialogue du corpus RATP-DECODA. Pour simplifier l'affichage, seules quelques annotations en lexique et syntaxe figurent ici. Les quatre dernières colonnes sont les annotations produites par notre système, elles donnent pour chaque mention son type sémantique et l'éventuel lien de coréférence trouvé (pour les pronoms, nous ne nous intéressons qu'aux pronoms à la troisième personne).

Tableau 5

62	7	vous	cln	SUJ	8	B-m	-	-	-
63	8	achetez	v	DEP_COORD	6	O	-	-	-
64	9	la	det	DET	10	B-m	-	-	COREF: echec
65	10	carte	nc	OBJ	8	I-m	-	Tout-NonAnime-ObjConcr	-
66	11	Navigo	np	MOD	10	I-m	-	Tout-NonAnime-Prod	-
67	12	Orange	np	MOD	11	I-m	-	Tout-NonAnime-Prod	-
68	13	qui	prorel	SUJ	14	B-m	TYPE: Tout	-	COREF: 64
69	14	est	v	MOD_REL	10	O	-	-	-
…	…	…	…	…	…	…	…	…	…
108	12	vous	cln	SUJ	13	B-m	-	-	-
109	13	achetez	v	ROOT	0	O	-	-	-
110	14	cette	det	DET	15	B-m	-	-	COREF: 64
111	15	carte	nc	OBJ	13	I-m	-	Tout-NonAnime-ObjConcr	-
112	16	cinq	det	DET	17	B-m	-	-	-
113	17	euros	nc	MOD	15	I-m	-	Tout-NonAnime-Abstr-Quantité	-
114	18	après	prep	ROOT	0	O	-	-	-
115	19	cette	det	DET	20	B-m	-	-	COREF: 110
116	20	carte	nc	OBJ	23	I-m	-	Tout-NonAnime-ObjConcr	-
117	21	vous	cln	SUJ	23	B-m	-	-	-
118	22	la	clo	OBJ	23	B-m	TYPE: Tout	-	COREF: 115
119	23	gardez	v	ROOT	0	O	-	-	-
120	24	elle	cln	SUJ	25	B-m	TYPE: Tout-NonAnime	-	COREF: 118
121	25	sera	v	ROOT	0	O	-	-	-
122	26	valable	adj	ATS	25	O	-	Tout-NonAnime	-
123	27	indéfiniment	adv	MOD	26	O	-	-	-
124	1	oui	pres	ROOT	0	O	-	-	-
125	1	et	coo	ROOT	0	O	-	-	-
126	2	elle	cln	NULL	0	O	-	-	COREF: echec
127	3	vous	cln	SUJ	4	B-m	-	-	-
128	4	demandez	v	DEP_COORD	1	O	-	-	-
…	…	…	…	…	…	…	…	…	…

8.2. Évaluation du système

58 L'évaluation et la comparaison des systèmes de recherche de coréférences sont complexes car il faut prendre en compte (Poesio *et al.* 2010) :

59 - les performances des éventuels pré-traitements pour l'annotation en genre, nombre, syntaxe, etc ;

60 - les performances de l'éventuel reconnaisseur d'entités nommées et de l'identification des mentions ;

61 - le type de texte que l'on traite : manuel technique, article de journal, dialogue, *etc*.

62 Différentes métriques ont été définies pour les campagnes d'évaluation. Nous avons choisi d'utiliser la métrique Blanc, qui est la plus récemment définie et dont la vocation est de considérer conjointement les liens de coréférence et de non-coréférence (Recasens & Hovy 2010). Les calculs se font sur les chaînes de coréférences de chaque mention.

63 L'évaluation du système a été faite en considérant le système « end-to-end » : c'est dire que l'on a évalué les performances de la totalité des traitements qui ont permis de passer d'un texte aux chaînes référentielles entre les mentions contenues dans ce texte.

64 Cette évaluation a été faite sur le corpus ANCOR : le principe en a été de comparer les chaînes référentielles établies manuellement dans ANCOR avec les chaînes référentielles que notre système a prédites de façon automatique sur les mêmes textes, et ceci sur toutes les mentions nominales ou pronominales du corpus, référentielles ou non (115 672 mentions).

65 Le corpus ANCOR a été partitionné en *train, dev* et *test*, pour pouvoir développer notre système de recherche des coréférences grâce aux deux premières parties et faire l'évaluation sur la dernière, la partie *test*, qui contient 23 079 mentions. La séquence de traitements a été la suivante sur les textes d'ANCOR :

66 - annotation en lexique et syntaxe par MACAON ;

67 - annotation en mentions prédites par le modèle d'apprentissage ;

68 - annotation en coréférences prédites par notre système.

69 Les étapes, les difficultés rencontrées et les résultats de cette évaluation sont décrites dans Godbert & Favre (2017). Nous avons obtenu sur le *test* d'ANCOR les taux de réussite suivants :

70 - sur les mentions prédites : 0.6570

71 - sur les pronoms prédits : 0.6405

72 Par ailleurs, une autre mesure a été faite avec une approche de base (plus naïve), qui ne recherche que le coréférent le plus proche (sans calculer les chaînes référentielles) : sur le *test* d'ANCOR nous avons obtenu un taux de réussite de 0.680.

73 Nous avons également fait une évaluation manuelle sur le corpus RATP-DECODA-Gold, composé de 102 dialogues, qui a déjà été utilisé comme étalon au cours de la phase d'annotation syntaxique. Cette évaluation a été faite en ne considérant que les pronoms clitiques *il, elle, ils, elles, le, la, l', les, lui, leur ;* dans ce cas, le taux de réussite a été de 0.72.

74 Pour évaluer l'apport du traitement sémantique des phrases, nous avons sur le même corpus effectué la séquence de traitements en ne prenant pas en compte les contraintes sémantiques (dans ce cas le type « Tout » est imposé à tous les noms, pronoms, verbes et adjectifs). Nous avons alors obtenu sur le *test* d'ANCOR les taux de réussite suivants :

75 - sur les mentions prédites : 0.6156

76 - sur les pronoms prédits : 0.5956

77 L'apport de la sémantique est donc relativement peu important, mais néanmoins bien réel.

9. Discussion

78 Dans les prédictions de notre système, la plupart des erreurs apparaissent dans les cas suivants :

79 - les cas où le pronom ne peut pas être typé de façon fine : en particulier s'il est sujet de *avoir* ou *être* ;

80 - les cas où le pronom n'a pas de lien de dépendance vers un verbe, ce qui est souvent dû à une disfluence ; les disfluences sont extrêmement nombreuses dans les dialogues ;

81 - des expressions figées, qui ne sont pas repérées dans les prétraitements ;

82 - des cas d'homonymie sur lesquels une erreur de lemmatisation produit une erreur d'analyse syntaxique.

83 Pour ce qui concerne les expressions figées, nous avons tenté d'en ajouter un traitement pour les identifier en préalable à la recherche de coréférences, notre objectif étant de noter comme « non référençable » un nom qui apparaît en tant que complément d'objet dans une expression polylexicale (formée de plusieurs mots).

84 Nous avons là encore utilisé les données du DEM, en extrayant du DEM les expressions polylexicales et en faisant sur ces expressions plusieurs opérations de croisement. Nous avons alors obtenu une table qui donne pour chaque nom les verbes avec lesquels il peut apparaître comme complément dans une expression polylexicale. Par exemple les mots *attention*, *compagnie* et *compte* sont chacun associé à un ensemble de verbes :

attention : attirer - faire - prêter - retenir
compagnie : fausser - tenir
compte : tenir - rendre - régler

85 Mais pour le moment cette approche n'a pas amélioré le taux de réussite. Il serait nécessaire de reprendre et d'affiner cette étude.

86 Nous avons aussi noté que la prise en compte de la sémantique des adjectifs n'a apporté qu'une faible amélioration des résultats. Il faudrait compléter le typage des adjectifs, car les informations extraites directement du DEM, sans correction manuelle, manquent souvent de pertinence. Nous pensons que, pour la détection automatique de coréférences, notre système est assez général car :

87 - il a une grande couverture lexicale grâce aux dictionnaires que nous y avons utilisés ;

88 - les procédures de recherche que nous avons définies sont classiques et devraient s'appliquer à tous les types de textes.

89 Néanmoins, notre travail s'est appuyé sur l'étude de corpus de dialogues ; ces procédures sont donc probablement plutôt adaptées à ce type de textes.

90 Nous retenons de ce travail que les ressources DEM et LVF se sont révélées d'une grande richesse et que les données que nous en avons extraites nous ont permis d'intégrer un niveau sémantique dans la recherche des chaînes de coréférence. Les résultats obtenus montrent que l'apport de la sémantique est relativement peu important, mais néanmoins bien réel.

BIBLIOGRAPHIE

BECHET, F., MAZA, B., BIGOUROUX, N., BAZILLON, T., EL-BEZE, M., DE MORI, R., ARBILLOT, E. (2012). « Decoda : a call-centre human-human spoken conversation corpus », in *LREC*, pp. 1343–1347.

CHOMSKY, N. (1981). *Lectures on Government and Binding*. Foris Publications.

DUBOIS, J., DUBOIS-CHARLIER, F. (1997). *Les Verbes Français*. Paris : Larousse-Bordas.

DUBOIS, J., DUBOIS-CHARLIER, F. (2010). « La combinatoire lexico-syntaxique dans le dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. » in *Langages 3*, pp. 31–56, Paris : Larousse.

DÉSOYER, A., LANDRAGIN, F., TELLIER, I., LEFEUVRE, A., ANTOINE, J. (2015). « Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR », in *TAL*, 55 (2), pp. 97–121.

GODBERT, E. (2014). « Typage sémantique de verbes avec LVF, pour la résolution d'anaphores », in *Atelier FondamenTAL, 21e Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2014),* Marseille, Juillet 2014, pp. 97-102.

GODBERT, E., FAVRE, B., (2017). « Détection de coréférences de bout en bout en français », in *Actes de la 24e Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2017),* Orléans, Juin 2017, pp. 52-59.

KIPPER-SCHULER, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon.* PhD. Thesis. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

LAPPIN, S., LEASS, H. (1994). « An algorithm for pronominal anaphora resolution », in *Computational Linguistics vol. 20 (4)*, pp. 535–561.

LEE, H., CHANG, A., PEIRSMAN, Y., CHAMBERS, N., SURDEANU, M., JURAFSKY, D. (2013). « Deterministic coreference resolution based on entity-centric, precision-ranked rules », in *Computational Linguistics vol. 39 (4)*, pp. 885–916.

LONGO, L., TODIRASCU, A. (2010). « RefGen : a Tool for Reference Chains Identification », in *International Multiconference on Computer Science and Information Technology*, pp. 447–454.

MITKOV, R. (2002). *Anaphora Resolution*. London : Longman.

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 *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pp. 555–563.

NASR, A., BECHET, F., REY, J., LE ROUX, J. (2011). « Macaon : a linguistic tool suite for processing word lattice », in *49th Annual Meeting of the Association for Computational Linguistics : demonstration session.*

POESIO, M., PONZETTO, S., VERSLEY, Y. (2010). « Computational models of anaphora resolution: A survey » *http ://wwwusers.di.uniroma1.it/ ponzetto/pubs/poesio10a.pdf*

RECASENS, M., HOVY, E. (2010). « Blanc: Implementing the rand index for coreference evaluation », in *Natural Language Engineering, 17 (4),* pp. 485–510.

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

RÉSUMÉS

Le travail présenté ici se situe dans le cadre de l'élaboration d'un système de traitement automatique du langage (TAL), pour la détection automatique de relations de coréférence intra- et inter-phrases en français, plus spécifiquement pour le traitement des anaphores pronominales et des coréférences directes. Nous décrivons comment nous avons utilisé les ressources du *Dictionnaire électronique des mots* (DEM) et *Les Verbes Français* (LVF) pour en extraire automatiquement des informations sémantiques sur les noms communs, les verbes et les adjectifs, et comment ces informations ont ensuite été exploitées pour la détection automatique de relations de coréférences.

The work presented here is part of the development of a natural language processing system (NLP), for automatic detection of intra- and inter-sentence co-reference relations in French texts, more specifically to identify pronominal anaphors and direct co-references. We describe how

semantic knowledge has been extracted from the « Dictionnaire électronique des mots » (DEM) and « Les Verbes Français » (LVF) for nouns, verbs and adjectives; and how this information has been used for automatic co-reference/anaphora resolution.

INDEX

Mots-clés : TAL, sémantique, mention, relations de coréférence
Keywords : NLP, semantics, mention, co-reference relations

AUTEUR

ELISABETH GODBERT

Aix-Marseille Université, Université de Toulon, CNRS, LIS, Marseille