Recherche de relations de coréférence

Elisabeth Godbert

Séminaire TALEP 5 mai 2017

Mentions et relations de coréférence

Mention : expression (nominale ou pronominale) qui dénote une entité

Les relations de coréférence mettent en relation les *mentions* qui font référence à une même entité
permettent le suivi des entités mentionnées dans les textes

[Un paquebot] est arrivé dans [le port] [ce matin], veux-[tu] [y] aller pour [le] voir ?
[c'] est [un bateau gigantesque]

- Identifier les antécédents des pronoms : *un paquebot* ... *le* ... *c'* *le port* ... *y*
  (traitement des ***anaphores*** pronominales)

- Trouver les coréférences : *un paquebot* … *un bateau*

Mentions et relations de coréférence

Mention : expression (nominale ou pronominale) qui dénote une entité

Les relations de coréférence mettent en relation les *mentions* qui font référence à une même entité
permettent le suivi des entités mentionnées dans les textes

[Un paquebot] est arrivé dans [le port] [ce matin], veux-[tu] [y] aller pour [le] voir ?
[c'] est [un bateau gigantesque]

- Identifier les antécédents des pronoms : *un paquebot* ... *le* ... *c'* *le port* ... *y*
  (traitement des ***anaphores*** pronominales)

- Trouver les coréférences : *un paquebot* … *un bateau*

- J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ?
- Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.
- Ce portable (/ cet appareil) m'a coûté cher, j'espère bien le retrouver.

Mai 2005 Jacques Chirac … le président de la république … l'ancien maire de Paris … le successeur de Mitterrand … son mandat

Mentions et relations de coréférence

Mention : expression (nominale ou pronominale) qui dénote une entité

Les relations de coréférence mettent en relation les *mentions* qui font référence à une même entité
permettent le suivi des entités mentionnées dans les textes

[Un paquebot] est arrivé dans [le port] [ce matin], veux-[tu] [y] aller pour [le] voir ?
[c'] est [un bateau gigantesque]

- Identifier les antécédents des pronoms : *un paquebot* ... *le* ... *c'* *le port* ... *y*
  (traitement des ***anaphores*** pronominales)

- Trouver les coréférences : *un paquebot* … *un bateau*

- J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ?
- Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.
- Ce portable (/ cet appareil) m'a coûté cher, j'espère bien le retrouver.

Mai 2005 Jacques Chirac … le président de la république … l'ancien maire de Paris … le successeur de Mitterrand … son mandat

Types de relations
directe : *mon portable … ce portable* (même tête nominale)
indirecte : *mon portable … cet appareil*
pronominale : *mon portable … le*
associative : *Jacques Chirac … son mandat*

Un paquebot est arrivé dans le port ce matin, veux-tu y aller pour le voir ? c'est un bateau gigantesque

- J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ?
- Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.
- Ce portable (/ cet appareil) m'a coûté cher, j'espère bien le retrouver.

Mon fils a eu un problème avec son bus ce matin ; il a perdu plus de 30 minutes à attendre
Mon fils a eu un problème avec son bus ce matin ; il est passé avec 30 minutes de retard
Mon fils a eu un problème avec son bus ce matin ; il est arrivé en retard au collège, il a été puni

Mai 2005 Jacques Chirac … le président de la république … l'ancien maire de Paris … le successeur de Mitterrand … son mandat

On utilise classiquement, pour identifier les éventuelles mentions coréférentes :
le genre et le nombre, la position syntaxique, le focus,
la distance des mots, des informations sémantiques

Un paquebot est arrivé dans le port ce matin, veux-tu y aller pour le voir ? c'est un bateau gigantesque

- J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ?
- Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.
- Ce portable (/ cet appareil) m'a coûté cher, j'espère bien le retrouver.

Mon fils a eu un problème avec son bus ce matin ; il a perdu plus de 30 minutes à attendre
Mon fils a eu un problème avec son bus ce matin ; il est passé avec 30 minutes de retard
Mon fils a eu un problème avec son bus ce matin ; il est arrivé en retard au collège, il a été puni

Mai 2005 Jacques Chirac … le président de la république … l'ancien maire de Paris … le successeur de Mitterrand … son mandat

On utilise classiquement, pour identifier les éventuelles mentions coréférentes :
le genre et le nombre, la position syntaxique, le focus,
la distance des mots, des informations sémantiques

Les systèmes développés pour traiter les anaphores portent en majorité sur l'anglais (quasiment rien sur le français)

Poesio Massimo et al. (2010). Computational Models of Anaphora Resolution: A Survey (112 pages)
Disponible en ligne : http://wwwusers.di.uniroma1.it/~ponzetto/pubs/poesio10a.pdf
Poesio Massimo et al. (2016). Anaphora Resolution: Algorithms, Resources, and Applications (500 pages)
En partie en ligne

Corpus d'application

Ces corpus doivent être **préalablement annotés en lexique et syntaxe (par Macaon)**

1) Corpus RATP-DECODA : dialogues oraux enregistrés dans le Centre d'appel de la RATP (2100 dialogues ; environ 560 000 mots)

2) Corpus de tchat Orange : discussions en tchat entre le Centre de l'opérateur et des clients (environ 95 000 mots)

3) Corpus Ancor, élaboré par l'équipe de J-Y. Antoine (Tours) : corpus de dialogues transcrits, annotés manuellement en - mentions [B-m I-m … I-m …]
- relations de coréférences
(460 dialogues ; environ 500 000 mots)

4) French Tree Bank : corpus arboré d'1 million de mots du journal *Le Monde* (1989-1995)

Dans les dialogues : la majorité des coréférences sont des anaphores pronominales
Celles qui nous intéressent sont les anaphores pronominales à la 3e personne

27	2	**la**	det	-	##f#s##d	DET	3
27	3	**carte**	nc	carte	##f#s##	OBJ	1
27	4	**au**	prep	au	#####	MOD	3
27	5	**mois**	nc	mois	##m#p##	OBJ	4
27	6	**et**	coo	et	#####	COORD	1
27	7	**je**	cln	cln	#1##s##	SUJ	8
27	8	**pensais**	v	penser	I#12##s##	DEP_COORD	6
27	9	**la**	clo	cla	#3#f#s##	OBJ	10
27	10	**reprendre**	vinf	reprendre	W#####	OBJ	8
27	11	**euh**	pres	euh	#####	DISFLINK1	10
27	12	**en**	prep	en	#####	DISFLINK2	11
27	13	**à**	prep	à	#####	MOD	10
27	14	**l'**	det	-	###s##d	DET	15
27	15	**année**	nc	année	##f#s##	OBJ	13
28	1	**euh**	pres	euh	#####	NOLINK	0
28	2	**faire**	vinf	faire	W#####	ROOT	0
28	3	**le**	det	-	##m#s##d	DET	4
28	4	**contrat**	nc	contrat	##m#s##	OBJ	2
28	5	**en**	prep	en	#####	P_OBJ	4
28	6	**décembre**	nc	décembre	##m#s##	OBJ	5
28	7	**pour**	prep	pour	#####	MOD	6
28	8	**la**	clo	cla	#3#f#s##	DISFLINK1	7
28	9	**changer**	vinf	changer	W#####	DISFLINK2	8
28	10	**la**	clo	cla	#3#f#s##	OBJ	11
28	11	**mettre**	vinf	mettre	W#####	OBJ	7
28	12	**à**	prep	à	#####	MOD	7
28	13	**l'**	det	-	###s##d	DET	14
28	14	**année**	nc	année	##f#s##	OBJ	12

L'aspect sémantique

La ***sémantique*** des entités dont on parle est un trait essentiel à prendre en compte

- J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ?
- Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.
- Ce portable (/ cet appareil) m'a coûté cher, j'espère bien le retrouver.

Mon fils a eu un problème avec son bus ce matin ; il a perdu plus de 30 minutes à attendre
Mon fils a eu un problème avec son bus ce matin ; il est passé avec 30 minutes de retard
Mon fils a eu un problème avec son bus ce matin ; il est arrivé en retard au collège, il a été puni

→ définir un typage sémantique des expressions nominales

→ définir un typage sémantique des verbes pour en déduire le type de chaque pronom

Extraction d'informations sémantiques dans le LVF et le DEM

Deux dictionnaires électroniques de J. Dubois et F. Dubois-Charlier, disponibles librement

LVF : Les Verbes Français

25 610 entrées verbales : 12 310 verbes différents, dont 4 188 à plusieurs entrées. 10 champs dont :

- la classe sémantique de cette entrée verbale, ("communication", "mouvement", etc.)
- le nombre d'actants du verbe et la nature de chaque actant : *humain, animal, chose, complétive, …*

Ajout de quelques classes :
Animé-ou-Véhicule,
ObjetConcret, Lieu-Hum, etc.

On obtient :
(environ 12 500 entrées)

verbe	sujet	compl-objet-direct
acquérir	Humain	NonAnimé
attendre	Animé-ou-Véhicule	Tout
atterrir	Tout	
attester	Humain	

Extraction d'informations sémantiques dans le LVF et le DEM

Deux dictionnaires électroniques de J. Dubois et F. Dubois-Charlier, disponibles librement

LVF : Les Verbes Français

25 610 entrées verbales : 12 310 verbes différents, dont 4 188 à plusieurs entrées. 10 champs dont :

- la classe sémantique de cette entrée verbale, ("communication", "mouvement", etc.)
- le nombre d'actants du verbe et la nature de chaque actant : *humain, animal, chose, complétive, …*

Ajout de quelques classes :
Animé-ou-Véhicule,
ObjetConcret, Lieu-Hum, etc.

On obtient :
(environ 12 500 entrées)

verbe	sujet	compl-objet-direct
acquérir	Humain	NonAnimé
attendre	Animé-ou-Véhicule	Tout
atterrir	Tout	
attester	Humain	

1579;attendre 01;TPS;temps;;;P3a;(#) [scrut] qc/qn;rester jusqu'à;On a~ P à la gare, le train trois heures. ;5aZ;T1900;--- -A ---- --;-;-;-;-;-;A;--;--;-;-;3T;1
1580;attendre 02;TPS;temps;;;L1a.1;(#) [lc.qp] en attente;patienter;On a~ à la station en attendant le bus. On n'aime pas a~. ;5aZ;A11;--- -A ---- --;-;-;-;-;-;A;--;--;-;-;3T;5
1581;attendre 03;PSY;psychologie;;;P1i.2;(#) [sent] espoir QUE/infinitif;escompter;On a~ que P~ sorte, de voir la suite. ;5aZ;T14b0;--- -A ---- --;-;-;-;-;-;A;--;--;-;-;3T;5
1582;attendre 04;CUI;cuisine, pâtisserie;;;H3a.2;(qc) [ê.e.état] bien;être prêt;La viande n'a~ pas. Le dîner peut a~. ;5aZ;A36;--- -- ---- --;-;-;-;-;-;--;--;-;-;-;5
1583;attendre 05;PSY;psychologie;;;P3a;(#) [scrut] qn pour qc;compter sur;On a~ P pour ce soir, une lettre pour demain. ;5aZ;T1900;--- -- ---- --;-;-;-;-;-;--;--;-;-;5
1584;attendre 06;PSY;psychologie;;;P1i.3;(#) [sent] espoir DE qn QUE;escompter;On a~ de P des excuses, qu'il s'excuse, une rentrée d'argent. ;5aZ;T15b0;-2- -- ---- --;-;2;-;-;-;--;--;-;-;-;5
1585;attendre 07;PSYf;psychologie;familier;;P3a;(#) [scrut] qn qp abstrait;guetter;On a~ P au tournant. On a~ P quand il abordera le sujet. ;5aZ;T1900;--- -- ---- --;-;-;-;-;-;--;--;-;-;-;5
1586;attendre 08;TPS;temps;;;P1i.2;(#) [sent] espoir QUE;compter que;On a~ que P revienne, son retour. ;5aZ;T14b0;--- -- ---- --;-;-;-;-;-;--;--;-;-;-;5
1587;attendre 09;PSY;psychologie;;;P1d.2;(#) [sent] désir après;compter sur;On a~ après le médecin, après une rentrée d'argent. ;5aZ;N1d;--- -- ---- --;-;-;-;-;-;--;--;-;-;-;5
1588;attendre 10;SOM;physiologie;;;H1c.1;(#) [f.espèce] humaine;être enceinte;Cette femme a~ un bébé pour le mois prochain. ;5aZ;T1100;--- -A ---- --;-;-;-;-;A;--;--;-;-;3T;5
1589;attendre 11 (s');PSY;psychologie;;;P1d.3;(#) [sent] attente A+infinitif/QUE;prévoir, pressentir;On s'a~ (à ce) que la situation s'aggrave, à une aggravation. ;5aZ;P10a0;-2-- ---- --;-;2;-;-;-;--;--;-;-;-;5

Extraction d'informations sémantiques dans le LVF et le DEM

Deux dictionnaires électroniques de J. Dubois et F. Dubois-Charlier, disponibles librement

LVF : Les Verbes Français

25 610 entrées verbales : 12 310 verbes différents, dont 4 188 à plusieurs entrées. 10 champs dont :

- la classe sémantique de cette entrée verbale, ("communication", "mouvement", etc.)
- le nombre d'actants du verbe et la nature de chaque actant : *humain, animal, chose, complétive, …*

Ajout de quelques classes :
Animé-ou-Véhicule,
ObjetConcret, Lieu-Hum, etc.

On obtient :
(environ 12 500 entrées)

verbe	sujet	compl-objet-direct
acquérir	Humain	NonAnimé
attendre	Animé-ou-Véhicule	Tout
atterrir	Tout	
attester	Humain	

DEM : Dictionnaire électronique des mots

145 198 entrées. 7 champs, dont : le sens du mot, et ses propriétés sémantiques et syntaxiques (classification par domaine mais pas hiérarchique)

Extraction des noms communs

+ ajout des mots
spécifiques des corpus

On obtient :

(environ 87 000 entrées)

nom	domaine	classe
abeille	ENT	Tout-Animé-Animal
acacia	SYL	Tout-NonAnimé
acadien	LOC-LAN-PERS	Tout
académicien	LIT-PERS	Tout-Animé-Humain

Traitement des anaphores

1) Repérage des mentions : → annotation en mentions *[le nouveau portable de [mon frère]] … [il]*

Par apprentissage à partir d'un modèle appris sur le corpus Ancor

Fait par Benoit, avec un modèle CRF entraîné à partir des annotations lexicales et syntaxiques

→ F-mesure = 88.94%

2) Typage sémantique des noms et pronoms : → annotation sémantique

- Identification des pronoms impersonnels (à ne pas traiter) :

il y a *il faut* *je le sais* *il est huit heures ….*

-Typage des autres pronoms en utilisant les dépendances et le typage des verbes issu du LVF

- Typage sémantique des noms communs en utilisant la classification issue du DEM

- Typage sémantique des noms propres à partir d'un fichier élaboré manuellement

Traitement des anaphores

1) Repérage des mentions : → annotation en mentions *[le nouveau portable de [mon frère]] … [il]*

Par apprentissage à partir d'un modèle appris sur le corpus Ancor

Fait par Benoit, avec un modèle CRF entraîné à partir des annotations lexicales et syntaxiques

→ F-mesure = 88.94%

2) Typage sémantique des noms et pronoms : → annotation sémantique

- Identification des pronoms impersonnels (à ne pas traiter) :

il y a *il faut* *je le sais* *il est huit heures ....*

-Typage des autres pronoms en utilisant les dépendances et le typage des verbes issu du LVF

- Typage sémantique des noms communs en utilisant la classification issue du DEM

- Typage sémantique des noms propres à partir d'un fichier élaboré manuellement

3) Recherche d'un antécédent pour chaque pronom → annotation en coréférences

rechercher plus haut dans le dialogue
(+ ou – haut)

- un nom susceptible d'être antécédent
- un pronom susceptible d'être coréférent

Traitement des anaphores

1) Repérage des mentions à partir d'un modèle appris sur le corpus Ancor → annotation en mentions

[mon portable] … [il] …[le nouveau portable de [mon frère]]

Fait par Benoit, avec un modèle CRF entraîné à partir des annotations lexicales et syntaxiques

F-mesure = 88.94%

2) Typage sémantique des noms et pronoms : → annotation sémantique

- Identification des pronoms impersonnels (à ne pas traiter) :

il y a *il faut* *je le sais* *il est huit heures* ....

-Typage des autres pronoms en utilisant les dépendances et le typage des verbes issu du LVF

- Typage sémantique des noms communs en utilisant la classification issue du DEM

- Typage sémantique des noms propres à partir d'un fichier adhoc (manuel)

3) Recherche d'un antécédent pour chaque pronom → annotation en coréférences

rechercher plus haut dans le dialogue *(+ ou – haut)*

- un nom susceptible d'être antécédent
- un pronom susceptible d'être coréférent

4) Recherche d'un coréférent pour chaque expression nominale définie → annotation en coréférences

rechercher plus haut dans le dialogue *(+ ou – haut)*

- un nom susceptible d'être **coréférent direct** (à l'étude : coréférences indirectes)

5) Calcul de la chaîne de coréférences, pour chaque mention

58	7	**ils**	cln	B-m	Tout-Anime-Humain	**COREF : echec**
58	8	**vont**	v			
58	9	**vous**	clo			
58	10	**la**	clo	B-m	Tout-NonAnime	**COREF : echec**
58	11	**refaire**	vinf			
58	12	**comme**	adv			
58	13	**elle**	cln	B-m	Tout	**COREF:la(58.10)**
58	14	**était**	v			
58	15	**avant+que**	csu			
58	16	**vous**	cln			
58	17	**la**	clo	B-m	Tout	**COREF:elle(58.13)**
58	18	**perdiez**	v			
65	2	**d'accord**	adv			
66	1	**bon**	pres			
66	2	**eh+ben**	pres			
66	3	**je**	cln	B-m		
66	4	**vais**	v			
66	5	**voir**	vinf			
66	6	**euh**	pres			
67	1	**et**	coo			
67	2	**je**	cln	B-m		
67	3	**leur**	clo	B-m	Tout-Anime	**COREF:ils(58.7)**
67	4	**demande**	v			

62	7	vous	cln	cln	#2##p##	SUJ	8	B-m	-	-	-
63	8	achetez	acheter	v	P#2##p##	DEP_COORD	6	O	-	-	-
64	9	la	le	det	##m#s##d	DET	10	B-m	-	-	COREF : echec
65	10	carte	carte	np	##f#s##	OBJ	8	I-m	-	Tout-NonAnime-ObjConcr	-
66	11	Navigo	Navigo	np	#####	MOD	10	I-m	-	Tout-NonAnime-Prod	-
67	12	Orange	Orange	np	##m#s##	MOD	11	I-m	-	Tout-NonAnime-Prod	-
68	13	qui	qui	prorel	#####	SUJ	14	B-m	TYPE:Tout	-	COREF : 64
69	14	est	être	v	P#3##s##	MOD_REL	10	O	-	-	-
108	12	vous	cln	cln	#2##p##	SUJ	13	B-m	-	-	-
109	13	achetez	acheter	v	P#2##p##	ROOT	0	O	-	-	-
110	14	cette	cet	det	##f#s##e	DET	15	B-m	-	-	COREF : echec
111	15	carte	carte	nc	##f#s##	OBJ	13	I-m	-	Tout-NonAnime-ObjConcr	-
112	16	cinq	cinq	det	#####	DET	17	B-m	-	-	-
113	17	euros	euro	nc	##m#p##	MOD	15	I-m	-	Tout-NonAnime-Abstr-Quantité	-
114	18	après	après	prep	#####	ROOT	0	O	-	-	-
115	19	cette	cet	det	##f#s##e	DET	20	B-m	-	-	COREF : 110
116	20	carte	carte	nc	##f#s##	OBJ	23	I-m	-	Tout-NonAnime-ObjConcr	-
117	21	vous	cln	cln	#2##p##	SUJ	23	B-m	-	-	-
118	22	la	cla	clo	#3#f#s##	OBJ	23	B-m	TYPE:Tout	-	COREF : 115
119	23	gardez	garder	v	P#2##p##	ROOT	0	O	-	-	-
120	24	elle	cln	cln	#3#f#s##	SUJ	25	B-m	TYPE:Tout	-	COREF : 118
121	25	sera	être	v	F#3##s##	ROOT	0	O	-	-	-
122	26	valable	valable	adj	###s##	ATS	25	O	-	-	-
123	27	indéfiniment	indéfiniment	adv	#####	MOD	26	O	-	-	-
124	1	oui	oui	pres	#####	ROOT	0	O	-	-	-
125	1	et	et	coo	#####	ROOT	0	O	-	-	-
126	2	elle	cln	cln	#3#f#s##	NULL	0	O	-	-	-
127	3	vous	cln	cln	#2##p##	SUJ	4	B-m	-	-	-
128	4	demandez	demander	v	P#2##p##	DEP_COORD	1	O	-	-	-
129	5	le	le	det	##m#s##d	DET	6	B-m	-	-	

Evaluation

1) (2015) **Evaluation manuelle sur le DECODA-Gold**

Taux de réussite : 0.72 sur les pronoms clitiques *il*, *elle*, *ils*, *elles*, *le*, *la*, *l'*, *les*, *lui*, *leur*

Evaluation

1) (2015) **Evaluation manuelle sur le DECODA-Gold**

Taux de réussite : 0.72 sur les pronoms clitiques *il, elle, ils, elles, le, la, l', les, lui, leur*

2) (2016-17) **Evaluation par rapport au corpus ANCOR**

ANCOR: annoté manuellement en relations de coréférences directes, indirectes et associatives,
sur toutes les mentions nominales ou pronominales du corpus, référentielles ou non (115 672 mentions)

A été annoté en ***lexique, syntaxe et mentions*** par Macaon et Benoit

Difficultés de l'évaluation :
1- Dans ANCOR, chaque mention référentielle est mise en relation avec sa mention coréférente
la plus éloignée (1e apparition dans le texte, de cette entité dont on parle)
Il faut donc utiliser les chaînes de coréférences, mais celles d'ANCOR passent par des relations *indirectes*

Evaluation

1) (2015) **Evaluation manuelle sur le DECODA-Gold**
Taux de réussite : 0.72 sur les pronoms clitiques *il, elle, ils, elles, le, la, l', les, lui, leur*

2) (2016-17) **Evaluation par rapport au corpus ANCOR**

ANCOR: annoté manuellement en relations de coréférences directes, indirectes et associatives,
Sur toutes les mentions nominales ou pronominales du corpus, référentielles ou non (115 672 mentions)

Ensuite : annoté en lexique, syntaxe et mentions par MACAON et Benoit

Difficultés de l'évaluation :
1- Dans ANCOR, chaque mention référentielle est mise en relation avec sa mention coréférente la plus éloignée (1e apparition dans le texte, de cette entité dont on parle)
Il faut donc utiliser les chaînes de coréférences, mais celles d'ANCOR passent par des relations *indirectes*

2- Quelle métrique utiliser ? Plusieurs métriques définies pour des campagnes d'évaluation (MUC, ACE, SemEval, ConLL)

La métrique BLANC : métrique la plus récente (Recasens (2010),
"dont la vocation est de considérer conjointement les liens de coréférence et de non-coréférence"

On définit la matrice de confusion sur toutes les paires de mentions, d'où on déduit : *rc* , *wc* , *wn* ;
on en déduit *rn* ; puis calculs de précision, rappel et F-mesure sur les coréférences et les non-coréférences
Finalement mes_Blanc = (Fc+Fn) / 2

Résultats actuels sur tout le « test » d'ANCOR, sur les mentions prédites : 0.6564 (sur mentions Ancor : 0.6698)
Sur les pronoms : 0.6392 (sur mentions Ancor : 0.6466) (meilleurs systèmes pour l'anglais : 0.70 à 0.75)

Remarques finales / perspectives

Les erreurs viennent de :

- Verbes au type sémantique très large (classe « Tout »), en particulier le sujet de *avoir* et *être*

- Les expressions figées non identifiées :
  Ex. adverbe *côte à côte* (*nc prep nc*) -- verbe *faire la queue* (*v det nc*) … *elle*

- Erreurs dans les annotations précédentes. (ambiguïtés / polysémie)
  Ex. *le boucher* (*cl v* ou *cl adj* au lieu de *det nc*) … *il*
  que (*csu* au lieu de *prorel* → pas une mention → pas traité)

- Disfluences Ex. *Je l'ai euh je l'ai échangé* *Il me semble euh…* (impers ?)

Remarques finales / perspectives

*** Les erreurs viennent de :**

- Verbes au type sémantique très large (classe « Tout »), en particulier le sujet de *avoir* et *être*

- Les expressions figées non identifiées :
  Ex. adverbe *côte à côte* (*nc prep nc*) -- verbe *faire la queue* (*v det nc*) … *elle*

- Erreurs dans les annotations précédentes. (ambiguïtés / polysémie)
  Ex. *le boucher* (*cl v* ou *cl adj* au lieu de *det nc*) … *il*
  que (*csu* au lieu de *prorel* → pas une mention → pas traité)

- Disfluences Ex. *Je l'ai euh je l'ai échangé* *Il me semble euh…* (impersonnel ?)

*** A faire : ajouter la sémantique des adjectifs**

*** Envisageable (?) : traiter des reprises indirectes** en utilisant plus de données sémantiques
mon portable … cet appareil (un dictionnaire de synonymes ?)

partiellement fait pour DECODA : quelques équivalences :
le 43 *le bus 43* *le procès-verbal* *le PV* *l'amende*

*** Actuellement** Les coréférences indirectes d'Ancor nuisent au taux de réussite
Ex. *tous les milieux … les classes sociales*

*** Hors de portée : anaphores associatives** avec relations de méronymie *Le président … son mandat*

Divers

Pour traiter un nouveau corpus, il faut :

- Une annotation préalable en : lemmes, catégories, dépendances, mentions
- Ajouter aux fichiers existants les ***nouveaux*** verbes et noms (propres ou communs) avec leur sémantique

Divers

Pour traiter un nouveau corpus, il faut :

- Une annotation préalable en : lemmes, catégories, dépendances, mentions
- Ajouter aux fichiers existants les ***nouveaux*** verbes et noms (propres ou communs) avec leur sémantique

Exotismes / termes spécifiques aux corpus

DECODA *:* mots en + *:* *Champ+de+Mars ; Champ-de-Mars ; Château_-_Margaux …*
en+fait ; Service+Clientèle+à+Distance ; Lyon+Bercy ; zéro+un+cinquante+huit+soixante+dix ; ...
(se) situer ; (se) tenir ; (se)+situer ; (se)+tenir ; …

ANCOR : 88 noms communs ont été ajoutés + N noms propres (personnes, lieux, org, …)

Divers

Pour traiter un nouveau corpus, il faut :

- Une annotation préalable en : lemmes, catégories, dépendances, mentions
- Ajouter aux fichiers existants les ***nouveaux*** verbes et noms (propres ou communs) avec leur sémantique

Exotismes / termes spécifiques aux corpus

DECODA *:* mots en + *:* *Champ+de+Mars ; Champ-de-Mars ; Château_-_Margaux …*
en+fait ; Service+Clientèle+à+Distance ; Lyon+Bercy ; zéro+un+cinquante+huit+soixante+dix ; ...
(se) situer ; (se) tenir ; (se)+situer ; (se)+tenir ; …

ANCOR : 88 noms communs ont été ajoutés + N noms propres (personnes, lieux, org, …)

Tchat Orange : 245 noms communs ajoutés dont :
connexion/configuration ; dep_tc_ge ; dyndns ; e-chat ; frutosso ; ipad ; iphone ; livebox2 ; ventilo ; …

* Problème de mots mal orthographiés, donc mal lemmatisés → distance de Hamming
* Problème entre majuscules / noms communs / noms propres

Ex. nom d'un client : FRUTOSSO frutosso nc → typé « Tout »
(tous les noms sont lemmatisés en minuscules et étiquetés nc)

Pronoms traités :

il , ils , elle , elles, le , la , les , l' , lui , leur ,

en , y ,

auquel , auxquels , auxquelles , autres , celle , celle-ci , celle-là , celles , celui , celui-ci , celui-là ,
certains , ceux , ceux-ci , ceux-là , chacun , dont ,
elle , elle-même , elles , elles-mêmes , eux , eux-mêmes ,
laquelle , lequel , lesquels , lesquelles , lui , lui-même ,
où , qu' , que , qui , quelques-unes , quelques-uns

Non traités :

tous , tout , c' , ça , ceci , cela : *très difficile* ! car beaucoup d'emplois impersonnels

Exemples (discutables) tirés d'Ancor, avec c' :

Le choix entre X… et Y… suivant les convictions des parents c'est une question de …
Donner une éducation chrétienne par le catéchisme à la maison j'ai pensé que c'était suffisant
Je fais cuire mon omelette, je la retourne quand elle est dorée d'un côté, c'est vite fait, c'est le menu quand…

Mes enfants aiment bien l'omelette c'est déjà quelque chose
Pour un étranger qui arrive par exemple le pire c'est on peut dire Marseille
Je n'ai qu' une formation élémentaire et c'est assez loin (pas de coréférence ?)