
Études sur la référence : traits linguistiques pour la résolution automatique

Silvia Federzoni

Thèse codirigée par Cécile Fabre et Lydia-Mai Ho-Dac

Laboratoire CLLE

UE Thématiques actuelles de la recherche en TAL - master LITL

26 octobre 2020

Chaînes de coréférence

Suite d'expressions d'un texte entre lesquelles l'interprétation établit une identité de référence (Corblin, 1995)

Chaîne de coréférence

George H.W. Bush s'inquiéta de la volonté du régime irakien de se doter d'armes de destruction massive, [...] Se sentant dans une situation de plus en plus délicate, **le président** comprit que la guerre était inévitable, et que son issue déterminerait **son** propre avenir politique. **Il** estimait en effet que la mauvaise conduite des opérations pouvait **lui** coûter cher [...]

Chaîne de coréférence

George H.W. Bush s'inquiéta de la volonté du régime irakien de se doter d'armes de destruction massive, [...] Se sentant dans une situation de plus en plus délicate, **le président** comprit que la guerre était inévitable, et que son issue déterminerait **son** propre avenir politique. **Il** estimait en effet que la mauvaise conduite des opérations pouvait **lui** coûter cher [...]

Référent : entité extra-linguistique dont on parle

Types de référent : être humain, entité abstraite ou concrète, événement

Objet d'étude : les chaînes de coréférence

Chaîne de coréférence

George H.W. Bush s'inquiéta de la volonté du régime irakien de se doter d'armes de destruction massive, [...] Se sentant dans une situation de plus en plus délicate, le président comprit que la guerre était inévitable, et que son issue déterminerait son propre avenir politique. Il estimait en effet que la mauvaise conduite des opérations pouvait lui coûter cher [...]

●
Maillons

Référent : entité extra-linguistique dont on parle

Types de référent : être humain, entité abstraite ou concrète, événement

Maillons : expressions linguistiques signalant le **référent** et porteuses d'une **valeur instructionnelle**

Rôle des chaînes de coréférence

Chaîne de coréférence

George H.W. Bush s'inquiéta de la volonté du régime irakien de se doter d'armes de destruction massive, [...] Se sentant dans une situation de plus en plus délicate, **le président** comprit que la guerre était inévitable, et que son issue déterminerait **son** propre avenir politique. **Il** estimait en effet que la mauvaise conduite des opérations pouvait **lui** coûter cher [...]

Maillons

Maillons : leur succession contribue à créer des **liens de cohésion**

Chaînes de coréférence : mécanisme fondamental dans l'organisation et l'interprétation du discours

Rôle des chaînes de coréférence

Cette maison est hantée.

Elle habitait dans cette maison depuis longtemps.

> longtemps.

Il marcha marcha : Il se retourna en entendant ce grand bruit.

grand

bruit. Maintenant ça fait très peur à tout le monde. Dans la rue il y a presque personne qui ose sortir, il y a que des gens courageux qui sortent dehors.

Depuis cette aventure, les enfants ne sortent plus la nuit.

Rôle des chaînes de coréférence

?

Cette maison est hantée.

Elle habitait dans cette maison depuis longtemps.

> longtemps.

Il marcha marcha. Il se retourna en entendant ce grand bruit.

grand

bruit. Maintenant ça fait très peur à tout le monde. Dans la rue il y a presque personne qui ose sortir, il y a que des gens courageux qui sortent dehors.

Depuis cette aventure, les enfants ne sortent plus la nuit.



Rôle des chaînes de coréférence

? Cette maison est hantée.

Elle habitait dans cette maison depuis longtemps.

> longtemps.

Il marcha marcha. Il se retourna en entendant ce grand bruit.

grand

bruit. Maintenant ça fait très peur à tout le monde. Dans la rue il y a presque personne qui ose sortir, il y a que des gens courageux qui sortent dehors.

Depuis cette aventure, les enfants ne sortent plus la nuit.



Rôle des chaînes de coréférence

? Cette maison est hantée.

Elle habitait dans cette maison depuis longtemps.

> longtemps.

Il marcha marcha.

Il se retourna en entendant ce grand bruit.

grand

bruit. Maintenant ça fait très peur à tout le monde. Dans la rue il y a presque personne qui ose sortir, il y a que les gens courageux qui sortent dehors.

Depuis cette aventure, les enfants ne sortent plus la nuit.



Le développement intensif des premières découvertes précipita rapidement une chute du prix du pétrole [...]. Dans cet univers de concurrence "sauvage", la première manifestation des forces organisatrices (ou plus exactement : planificatrices) ne vint pas de l'extérieur - de la puissance publique - mais de l'industrie elle-même.

Un jeune homme de 25 ans, John D. Rockefeller, après avoir brièvement tenté **sa** chance dans l'amont, délaissa ce jeu "où s'épuisent les pauvres gens" pour se concentrer sur le raffinage, à la tête de **la Standard Oil**. Très tôt **il** comprit l'intérêt de l'intégration horizontale, [...]. **Son** ascension fut fulgurante.

En 1873, **la Standard Oil** détenait déjà entre 30 et 40% des capacités de raffinage du pays, et jusqu'à 90% en 1878. Au tournant du siècle, **la S.O.** exportait 50% de **sa** production ; les États-Unis étaient, de très loin, le premier exportateur de pétrole au monde, et l'huile était au second rang des produits d'exportations américains, après le coton.

Les chaînes de coréférence : différentes approches

- En linguistique cognitive : théorie de l'accessibilité (Ariel, 2001) et théorie du centrage (Walker, Joshi, & Prince, 1998)
 - Principes qui régissent l'interprétation et la construction de la continuité référentielle

- En psycholinguistique (Gundel, Hedberg, & Zacharski, 2019 ; Kaiser & Fedele, 2019 ; Roberts, 2019 ; Salazar Orvig, 2019)
 - Facteurs qui influencent le choix des expressions référentielles

Les chaînes de coréférence : différentes approches

- En linguistique descriptive (Charolles, 2002 ; Cornish, 2000 ; Kleiber, 1994, 2002 ; Schnedecker, 2005)
 - Description des variations selon les genres textuels (Schnedecker, 2014 ; Schnedecker & Landragin, 2014 ; Schnedecker & Longo, 2012)

Limite ⇒ Aucune étude à large échelle : petits corpus, prise en compte d'un seul type de référent et peu de paramètres pour caractériser les chaînes

- En traitement automatique du langage (TAL) (Landragin, 2018 ; Longo, 2013 ; Poesio, Pradhan, Recasens, Rodriguez, & Versley, 2016 ; Wilkens, Oberle, Landragin, & Todirascu, 2020)
 - Détection automatique des chaînes de coréférence et des maillons (De Marneffe, Recasens, & Potts, 2015 ; Mitkov, 2014 ; Recasens & Hovy, 2010)

Point positif ⇒ Diffusion de gros corpus annotés en chaînes de coréférence

Objectifs différents : annotations différentes

MUC (MUC consortium 1995c)

AnCora (Taulé et al., 2008)

ACE (Dodding et al., 2004)

AnnoDis (Péry-Woodley et al., 2011)

Ontonotes (Pradhan et al., 2012)

① Diversité de phénomènes traités

AnCor (Muzerelle et al., 2014)

② Cadres théoriques différents

③ Diversité des choix d'annotation

ARRAU (Poesio et al., 2013)

Democrat (Landragin et al., 2019)

WikiCoref (Ghaddar & Langlais, 2016)

E-Calm (Garcia-Debanco et al., en cours)

Résultats difficilement comparables

Typologie des chaînes de coréférence

Description **systematique** et **exhaustive** de la **variété** et de la **complexité** des chaînes de référence à la lumière de **corpus diversifiés**

- corpus de **grande taille**
- différents **types** et **genres textuels**
- différents **niveaux d'expertise rédactionnelle**
- plusieurs **types de référents**

Contributions de la thèse

Théorique

Approfondir la connaissance des stratégies de construction de la continuité référentielle grâce à une description systématique des chaînes de coréférence

Méthodologique

Unifier et exploiter des ressources hétérogènes, conçues sur des bases théoriques différentes

Applicative

TAL : Évaluer les erreurs des systèmes de résolution et faire émerger les traits linguistiques discriminants pour améliorer leur performances

Didactique/psycholinguistique : Comprendre le processus d'acquisition des stratégies d'organisation discursive (textes d'experts vs textes d'apprenants)

- ① Contribution théorique
- ② Contribution méthodologique
- ③ Contribution applicative

- ① Contribution théorique
- ② Contribution méthodologique
- ③ Contribution applicative

- Description systématique et exhaustive des chaînes de référence
- Prise en compte de différents genres et types textuels
- Prise en compte de plusieurs types de référents

- Description systématique et exhaustive des chaînes de référence
- Prise en compte de différents genres et types textuels
- Prise en compte de plusieurs types de référents

● **Modèle de description en trois niveaux :**

- ① Maillons
- ② Chaînes
- ③ Structure textuelle

① Maillons

- Nature du maillon : SN_def, SN_dem, pronom, noms propres
- Fonction syntaxique : sujet, objet, autre
- Genre et nombre
- Position dans la chaîne
- Degré d'informativité du maillon : présence ou absence de modifieurs
- Type de référent : humain, non-humain, entité abstraite ou concrète, événement
- Type de reprise :
 - directe** : *La Standard Oil... ...la Standard Oil...*
 - indirecte** : *George H.B. Bush... ...le président...*
 - pronominale** : *Un jeune homme de 25 ans... ...il...*
- Type d'emploi : générique ou spécifique

① Maillons

- Distance en nombre de mots
- Distance en nombre de mentions
- Distance en nombre de phrases
- Distance en nombre de paragraphes

② Chaînes

- Longueur des chaînes en nombre de maillons
- Couverture des chaînes
- Cohabitation des chaînes

Ex. : **Un professeur d'histoire** vit une vie de famille sans histoire, entouré de l'affection des **siens** et aveugle aux problèmes politiques et sociaux engendrés par l'"apartheid", régime ségrégationniste qui sévit en Union sud-africaine. **Cet homme paisible** a **un jardinier noir**, paisible **lui** aussi, et soumis à la fatalité. **Ce jardinier** a **un fils** et **ce fils** [...]

Ex. : **Deux hommes** parurent. **L'un** venait de la Bastille, **l'autre** du Jardin des Plantes. **Le plus grand**, vêtu de toile, marchait le chapeau en arrière, le gilet déboutonné et sa cravate à la main. **Le plus petit**, dont le corps disparaissait dans une redingote marron baissait la tête sous une casquette à visière jaune. Quand **ils** furent arrivés au milieu du boulevard, ils s'assirent à la même minute, sur le même banc.

- Instabilité des chaînes : pourcentage de désignations différentes à l'intérieur d'une même chaîne

Influence de la structure textuelle sur le choix des expressions référentielles (Ariel, 2001 ; Guillot-Barbance & Quignard, 2019 ; Schnedecker, 2005)

③ Structure textuelle

- **Titres** : le référent est-il mentionné dans le titre ?
- **Chapitres, sections, paragraphes** : chaînes locales ou globales

- Nécessité d'avoir des **ressources annotées**
 - annotations à différents niveaux : syntaxique, sémantique

- Nécessité de mettre en place une chaîne de traitement intégrant le parsing

- ① Contribution théorique
- ② Contribution méthodologique
- ③ Contribution applicative

Objectifs différents : annotations différentes

MUC (MUC consortium 1995c)

AnCora (Taulé et al., 2008)

ACE (Doddingtong et al., 2004)

Différents modèles d'annotation :

- ① types de référent
- ② critères de délimitation
- ③ annotation des singletons
- ④ informations annotées

AnnoDis (Péry-Woodley et al., 2011)

Ontonotes (Pradhan et al., 2012)

AnCor (Muzurelle et al., 2014)

ARRAU (Poesio et al., 2013)

Democrat (Landragin et al., 2019)

WikiCoref (Ghaddar & Langlais, 2016)

E-Calm (Garcia-Debanc et al., 2017)

Différents formats :

- ① SGML (in-line/stand-off)
- ② MMAX2
- ③ Glozz
- ④ XML-TEI-URS

Objectifs différents : annotations différentes

MUC (MUC consortium 1995c)

AnCora (Taulé et al., 2008)

ACE (Doddington et al., 2004)

Différents modèles d'annotation :

- ① types de référent
- ② critères de délimitation
- ③ annotation des singletons
- ④ informations annotées

AnnoDis (Péry-Woodley et al., 2011)

Ontonotes (Pradhan et al., 2012)

AnCor (Muzurelle et al., 2014)

ARRAU (Poesio et al., 2013)

Democrat (Landragin et al., 2019)

WikiCoref (Ghaddar & Langlais, 2016)


Différents formats :

- ① SGML (in-line/stand-off)
- ② MMAX2
- ③ Glozz
- ④ XML-TEI-URS


E-Calm (Garcia-Debanc et al., 2017)

- ANNODIS (2010) : textes longs entièrement annotés, non narratifs, 3 genres textuels (rapports et articles scientifiques, textes encyclopédiques)
- DEMOCRAT (2019) : échantillons des textes (10 000 mots), narratifs (en prose et en vers) et non narratifs, 18 genres textuels, du 12^{ème} au 20^{ème} siècle
- ANCOR (2014) : trois corpus de parole conversationnelle (Accueil_UBS, OTG et ESLO)
- E-CALM (2019-2021) : écrits scolaires à différents niveaux de littératie (du primaire à l'Université)

Corpus	Nb de mots	Nb de CR	Nb de maillons
AnnoDis	666 000	581	3 456
DEMOCRAT	688 851	20 410	196 923
ANCOR	488 000	51 337	116 071
E-Calm	en cours de constitution		

Corpus	Nb de mots	Nb de CR	Nb de maillons
AnnoDis	666 000	581	3 456
DEMOCRAT	688 851	20 410	196 923
ANCOR 	488 000	51 337	116 071
E-Calm	en cours de constitution		

 Nb de CR \Rightarrow Nb de relations

Corpus	Nb de mots	Nb de CR	Nb de maillons
AnnoDis	666 000	581	3 456
DEMOCRAT	688 851	20 410	196 923
ANCOR 	488 000	51 337	116 071
E-Calm	en cours de constitution		

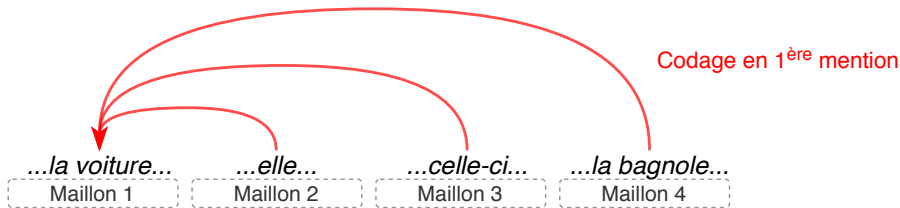


Figure – Adapté de (Antoine et al., 2016)

Corpus	Nb de mots	Nb de CR	Nb de maillons
AnnoDis	666 000	581	3 456
DEMOCRAT	688 851	20 410	196 923
ANCOR ⚠	488 000	51 337	116 071
E-Calm	en cours de constitution		

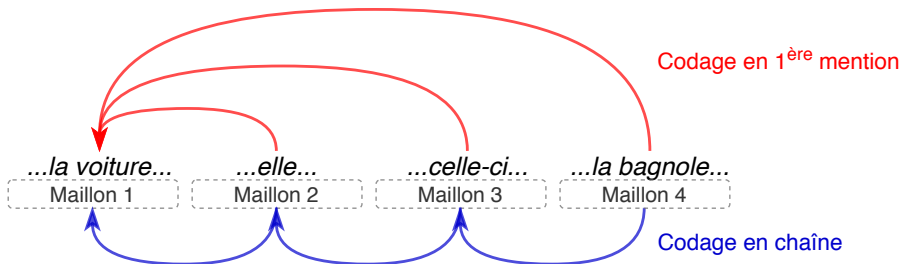


Figure – Adapté de (Antoine et al., 2016)

Corpus	Nb de mots	Nb de CR	Nb de maillons
AnnoDis	666 000	581	3 456
DEMOCRAT	688 851	20 410	196 923
ANCOR	488 000	51 337	116 071
E-Calm	en cours de constitution		

Corpus	Nb de mots	Nb de CR	Nb de maillons
AnnoDis	666 000	581	3 456
DEMOCRAT	688 851	20 410	196 923
ANCOR	488 000	51 337	116 071
E-Calm	en cours de constitution		

• Différents choix d'annotation

AnnoDis

- Annotation par des non-experts
- Guidée par un prémarquage automatique

DEMOCRAT

- Annotation par des experts
- Volonté d'annoter toutes les expressions référentielles

Un jour, **[le père de Léonard]**_{Léo}, ser Piero, « prit plusieurs de **[ses]**_{Léo} dessins et les soumit à son ami Andrea del Verrocchio qu'il pria instamment de lui dire si **[Léonard]**_{Léo} devait se consacrer à l'art du dessin et s'**[il]**_{Léo} pourrait parvenir à quelque chose en cette matière. Andrea s'étonna fort des débuts extraordinaires **[de Léonard]**_{Léo} et exhorta ser Piero à **[lui]**_{Léo} permettre de choisir ce métier, sur quoi, ser Piero résolut que **[Léonard]**_{Léo} entrerait à l'atelier d'Andrea.

AnnoDis

- Annotation par des non-experts
- Guidée par un prémarquage automatique

DEMOCRAT

- Annotation par des experts
- Volonté d'annoter toutes les expressions référentielles

Un jour, [le père de Léonard]_{Léo}, ser Piero, « prit plusieurs de [ses]_{Léo} dessins et les soumit à son ami Andrea del Verrocchio qu'il pria instamment de lui dire si [Léonard]_{Léo} devait se consacrer à l'art du dessin et s'[il]_{Léo} pourrait parvenir à quelque chose en cette matière. Andrea s'étonna fort des débuts extraordinaires [de Léonard]_{Léo} et exhorta ser Piero à [lui]_{Léo} permettre de choisir ce métier, sur quoi, ser Piero résolut que [Léonard]_{Léo} entrerait à l'atelier d'Andrea.

AnnoDis

- Annotation par des non-experts
- Guidée par un prémarquage automatique

Un jour, [le père de [Léonard]_{Léo}]_P, ser Piero, « prit plusieurs de [[ses]_{Léo} dessins]_d et [les]_d soumit à [[son]_P ami Andrea del Verrocchio]_{AV} qu'[il]_P pria instamment de [lui]_P dire si [Léonard]_{Léo} devait se consacrer à [l'art du dessin]_{Ad} et s'[il]_{Léo} pourrait parvenir à quelque chose en [cette matière]_{Ad}. [Andrea]_{AV} s'étonna fort des débuts extraordinaires de [Léonard]_{Léo} et [exhorta]_{AV} [ser Piero]_P à [lui]_{Léo} permettre de choisir [ce métier]_{Ad}, sur quoi, [ser Piero]_P résolut que [Léonard]_{Léo} entrerait à [l'atelier d'[Andrea]_{AV}]_y

DEMOCRAT

- Annotation par des experts
- Volonté d'annoter toutes les expressions référentielles

Corpus	Nb de mots	Nb de CR	Nb de maillons
AnnoDis	666 000	581	3 456
DEMOCRAT	688 851	20 410	196 923
ANCOR	488 000	51 337	116 071
E-Calm	en cours de constitution		

On tire parti des
deux approches

AnnoDis

- Annotation par des non-experts
- Guidée par un prémarquage automatique

DEMOCRAT

- Annotation par des experts
- Volonté d'annoter toutes les expressions référentielles

Tâche d'annotation

- ① Identification des maillons

Contrairement à **une idée répandue**, **les girafes** possèdent **des cordes vocales** mais **elles** n'émettent que très rarement **des sons**, se reposant davantage sur **la vision** que sur **l'audition** pour communiquer via par exemple **des postures et des mouvements du cou et de la tête**.

Tâche d'annotation

- ① Identification des maillons
- ② Délimitation des maillons et attribution des propriétés

Contrairement à **[une idée répandue]**, **[les girafes]** possèdent **[des cordes vocales]** mais **[elles]** n'émettent que très rarement **[des sons]**, se reposant davantage sur **[la vision]** que sur **[l'audition]** pour communiquer via par exemple **[[des postures] et [des mouvements [du cou] et de [la tête]]]**.

Tâche d'annotation

- ① Identification des maillons
- ② Délimitation des maillons et attribution des propriétés
- ③ Création manuelle ou automatique des chaînes

Contrairement à *[une idée répandue]_a*, *[les girafes]_b* possèdent *[des cordes vocales]_c* mais *[elles]_b* n'émettent que très rarement *[des sons]_d*, se reposant davantage sur *[la vision]_e* que sur *[l'audition]_f* pour communiquer via par exemple *[[des postures]_h* et *[des mouvements [du cou]_j et de [la tête]_k]_i]_g*.

Modèles d'annotation

Annotation	AnnoDis	DEMOCRAT	ANCOR	E-Calm
Type de référent	Tout	Tout	Pas spatio-temporel	Humain
Délimitation des maillons Singletons	Non précisé X	Précisé ✓	Précisé ✓	Précisé X
Propriétés annotées				
Nature	✓	X	✓	X
Genre et nombre	X	X	✓	X
Fonction syntaxique	X	X	X	X
Type de reprise	X	X	✓	X
Entité nommée	X	X	✓	X
Type d'emploi	X	X	✓	X

Comparer les annotations - Premières analyses sur AnnoDis

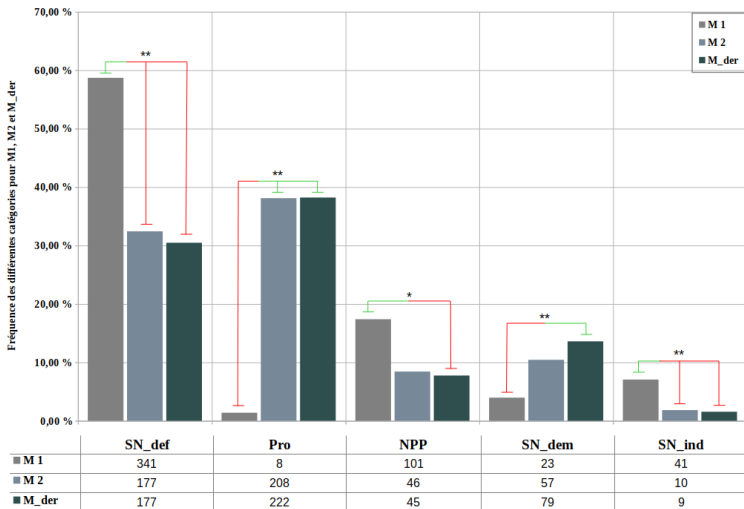
Objectif : évaluer la validité des annotations AnnoDis

Méthode :

1. Préparation des données :
 - Exploration préliminaire et **correction des erreurs de délimitation**
Ex. : Le [pont des embarcations] / [Le père de **Léonard**]
 - **Homogénéisation** des annotations : nature des maillons
2. Comparaison des résultats avec les descriptions existantes (Ariel, 2001 ; Manuélian, 2003 ; Schnedecker, 2014)
 - Description des **maillons** : **nature** et **position** des maillons (Corblin, 1987 ; Cornish, 1998 ; Salles, 2015)
 - Description des **chaînes** : longueur (Longo, 2013 ; Schnedecker, 2005 ; Todirascu et al., 2017)

Nature des maillons

- Des résultats conformes aux théories sur le rapport entre nature et position (Ariel, 2001 ; Corblin, 1987 ; Cornish, 1998)



- Annotations AnnoDis validées
- Méthode de comparaison validée

**Comment mettre au jour des types des chaînes ?
Quels sont les critères de description discriminants ?**

Objectif

Mettre au jour des types des chaînes de référence en observant les enchaînement les plus fréquents

Nature maillons	#	%
SN_def	1198	34,66
SN_dem	272	7,87
SN_ind	126	3,65
SN_sansDET	64	1,85
NPP	442	12,79
Poss	182	5,27
Pro	1026	29,69
Autre	146	4,22

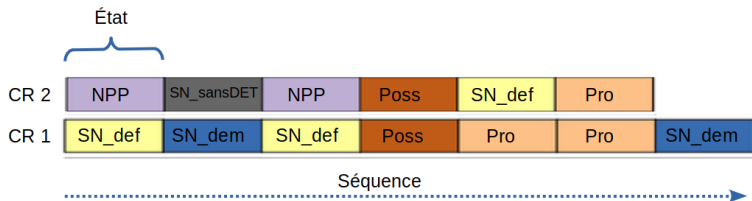
- La fréquence des catégories grammaticales : variations les plus fortes
- Informations les plus riches sur la typologie des chaînes : degré d'homogénéité

(Obry, Glikman, Guillot-Barbance, & Pincemin, 2017)

Vers une typologie des chaînes

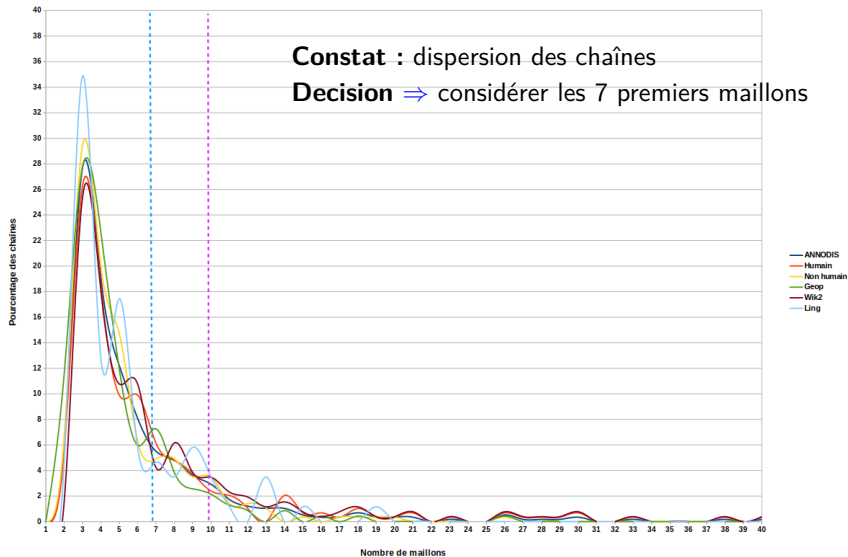
Méthode utilisée

Analyse des séquences (Quiniou, Cellier, Charnois, & Legallois, 2012) à partir de la nature des maillons



TraMineR (Gabadinho, Ritschard, Studer, & Müller, 2009)

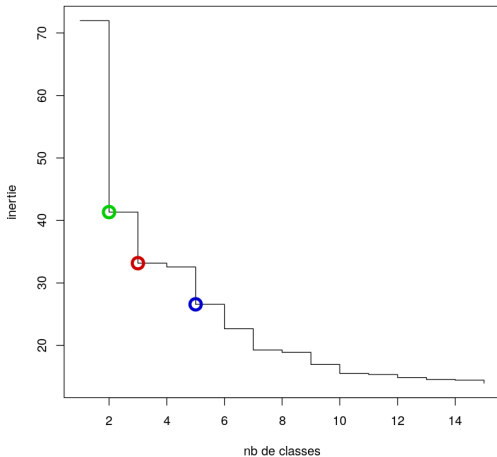
Méthode - limiter la variation dans la taille des séquences



Méthode - classification hiérarchique

Établir le nombre de classes : sauts d'inertie (2,3,5) et observations des classes

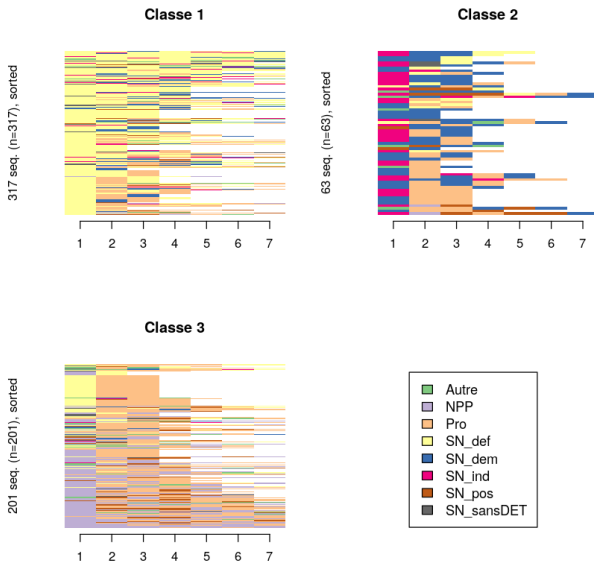
Décision \Rightarrow 3 classes



Méthode - Observer la tendance générale

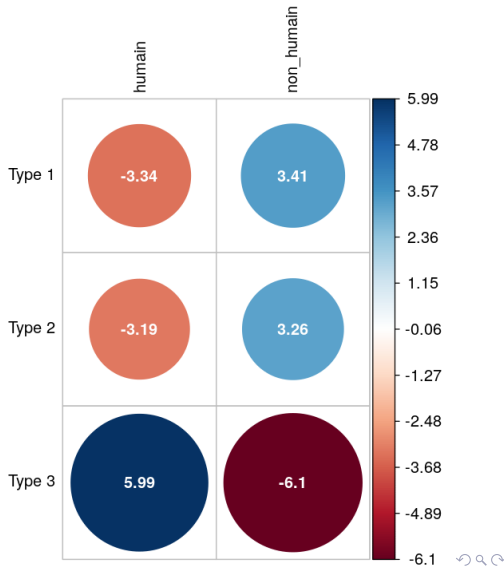
Constat : la classe 3 semble être représentative des chaînes à référent humain

Décision ⇒ vérifier s'il existe une corrélation entre les classes et le type de référent

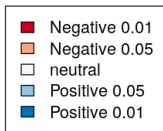
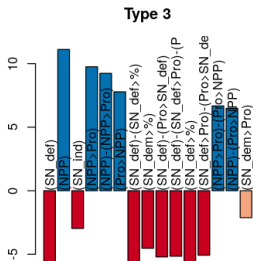
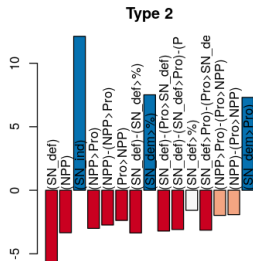
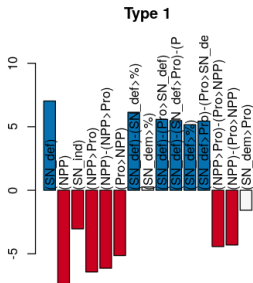


Constat : liaison significative entre les classes et le type de référent (test du khi-deux)

Décision \Rightarrow analyser les séquences pour dégager des **types d'enchaînements** en fonction du **type de référent**



Méthode - récupérer les séquences discriminantes



Analyses des séquences que le système propose

- Tri par fréquence
- Regroupement manuel des séquences ayant les mêmes caractéristiques

Ex. : NPP > Pro **et** NPP > Pro > Pro > Pro
⇒ NPP > Pro{1:3}

Référents humains

- 50% des M_1 = SN_def
- 33% des M_1 = NPP
- 5% des M_1 = SN_ind

Référents non humains

- 69% des M1 = SN_def
- 9% des M1 = SN_ind

Référents humains

NPP{1:2} > Pro {1:7} (15 cas)

Chez **F. de Saussure**, l'analogie [...], **il** pose que les facteurs de trouble [...]. Pour **lui**, cette tendance à l'irrégularité [...]. Comme H. Paul, **il** ramène le concept [...]

Référents humains

- 50% des M_1 = SN_def
- 33% des M_1 = NPP
- 5% des M_1 = SN_ind

Référents non humains

- 69% des M1 = SN_def
- 9% des M1 = SN_ind

Référents humains

NPP{1} > Pro{1:2} > NPP{1:2} > Pro{1:7} (> SN_def|NPP) (16 cas)

En effet, Godefroy Cavaignac, nouveau ministre de la Guerre [...], qu'il tient pour [...] Il est absolument convaincu [...]. Cavaignac a l'honnêteté [...]. Il avait eu la surprise d'apprendre [...] Il décide d'enquêter lui-même, dans son bureau avec ses adjoints [...].

Référents humains

- 50% des M_1 = SN_def
- 33% des M_1 = NPP
- 5% des M_1 = SN_ind

SN_def{1:7} (22 cas)

L'Union européenne

reste le 3e pollueur mondial [...] L'UE a lancé en 2005 le marché de permis européen [...]

La Commission européenne va [...], et publier (prévu en 2007) un « Livre vert » sur l'adaptation de l'UE au changement climatique, [...]

Référents non humains

- 69% des M1 = SN_def
- 9% des M1 = SN_ind

SN_def{1:7} (24 cas)

le vignoble champenois

s'étendait sur quelques [...] le vignoble connaît [...] Après les fléau du phylloxéra et de la Grande guerre, le vignoble s'est réduit à 12 000 hectares. Aujourd'hui, en 2007, le vignoble champenois s'étend sur 32 341 hectares.

① **Ajouter d'autres traits linguistiques et évaluer leur influence sur la classification des chaînes**

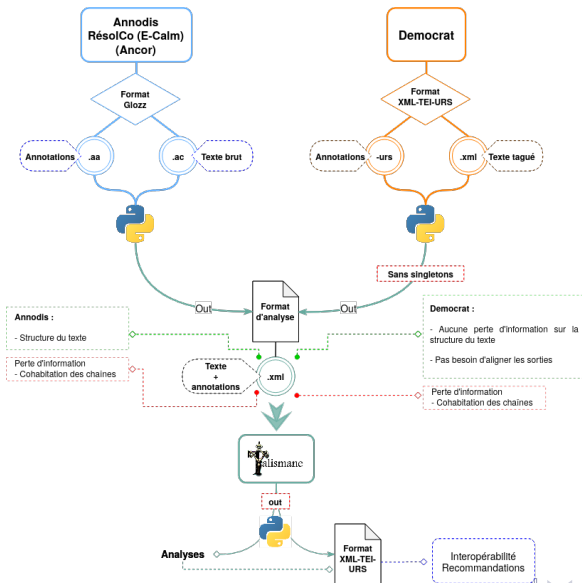
- Prise en compte de la fonction syntaxique
 - ⇒ ne suffit pas pour dégager des types de chaînes
- Combinaison de plusieurs traits

② **Réproduire l'étude sur les corpus unifiés**

- Unifier les ressources : nécessaire pour récupérer les mêmes informations

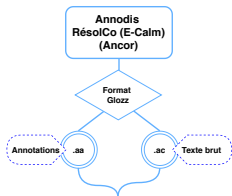
⇒ Aligner les formats d'annotation

Aligner les formats d'annotation



Aligner les formats d'annotation

Point en commun : annotationn déportée



Renvoi au texte : nombre de caractères

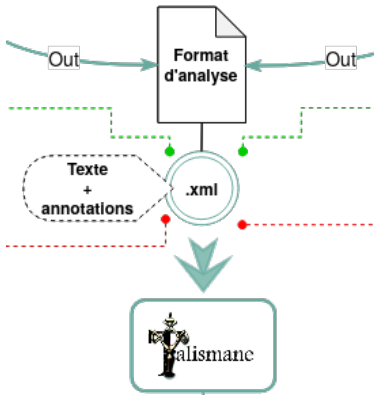
```
</unit>
<unit id="markmacr_1252528266">
  <metadata>
    <author>markmacr</author>
    <creation-date>1252528266</creation-date>
  </metadata>
  <characterisation>
    <type>COREFpropos</type>
    <featureSet/>
  </characterisation>
  <positioning>
    <start>
      <singlePosition index="9146"/>
    </start>
    <end>
      <singlePosition index="9151"/>
    </end>
  </positioning>
```

e le terrorisme : essai de bilan institutionnel François Vergotille de
is vergotille de Chantal, Docteur en Sciences Politiques de l'IEP de Paris,
Conférences en civilisation américaine à l'Université de Bourgogne. Depuis
e d'années, les républicains se sont fait fort de réduire le poids de l'Etat
actuelle lutte contre le terrorisme, menée par une équipe républicaine qui,
re totalement aux critiques contre le Big Government, rennetrait en cause
onservateur en faveur de la décentralisation. Les différentes mesures
is septembre 2001 vont toutes dans le même sens, un considérable renforcement
de l'état fédéral. Comme toutes les guerres menées par les Etats-Unis, celle
le terrorisme risquerait, elle aussi, de renforcer la centralisation. Quels
ts de ce casus est l'Etat central? Comment s'opère la recentralisation, et
onséquences dans l'équilibre fédéral? Finalement, quelles sont les
tirs de cette évolution? En particulier, comment s'articule la lutte contre
avec 'engagement conservateur en faveur des Etats fédérés? Selon nous, la
e terrorisme ne serait pas stérile aux évolutions entraînées par les autres
Débouche en fait sur un activisme tous-azimut, qui concerne aussi bien
que les Etats fédérés et les autorités locales (villes, comtés). Plutôt que
entralisation, il faudrait évoquer le renforcement des fonctions légitimes de
eaux du gouvernement : la défense et la protection des citoyens pour le
; les autorités locales, elles, perent les moyens de réponse immédiats aux
oristes (police, pompiers, santé). L'essentiel des problèmes suscités par la
territoire contre le terrorisme réside dans la coordination entre les
anes. L'administration actuelle s'engage résolument dans cette voie, et
rganisation massive des administrations nationales. Face à l'urgence : les

Aligner les formats d'annotation

Besoin : ne pas perdre d'informations et effectuer des traitements homogènes

Proposition : format intermédiaire



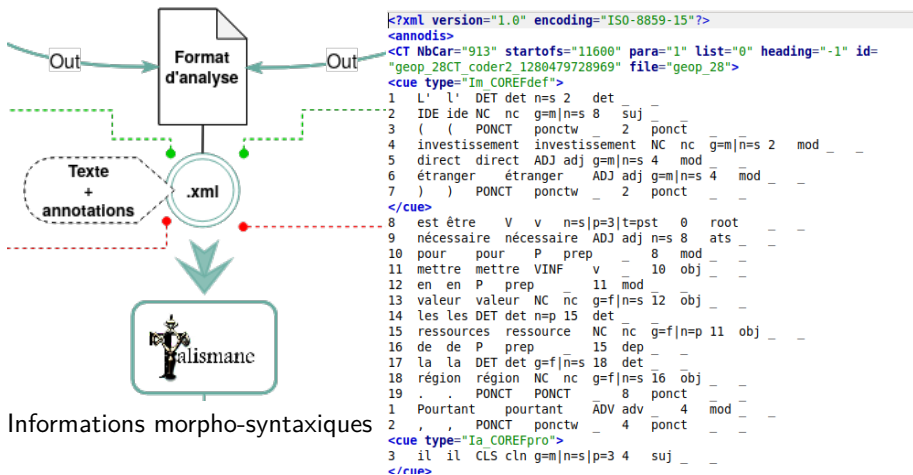
Informations morpho-syntaxiques

```
<?xml version="1.0" encoding="ISO-8859-15"?>
<?xml-stylesheet type="text/xsl" href="ANMODIS.xsl"?>
<annodis>
  <structure>
    <context type="before">... en plus dépendantes de l'aide et de l'investissement étrangers.
    </context>
    <CT NbCar="913" startofs="11600" para="1" list="0" heading="-1" id="
      "geop_2BCT_coder2_1280479728969" file="geop_28">
    <firstCOREf>L'IDE (investissement direct étranger)</firstCOREf>
    <tags>
      <tag nature="Im_COREFdef" start="11600">L'IDE (investissement direct étranger)</tag><tag
        nature="Ia_COREFpro" start="11715">il</tag><tag nature="Im_COREFdef" start="11830">
        l'IDE</tag><tag nature="Im_COREFdef" start="11886">l'IDE par habitant</tag><tag nature=
        "Ia_COREFdem" start="12001">ce ratio</tag><tag nature="Im_COREFdef" start="12127">l'IDE
        </tag><tag nature="Ia_COREFdef_R" start="12151">l'IDE</tag><tag nature="Im_COREFind"
        start="12273">des IDE investis</tag><tag nature="Ia_COREFdem" start="12329">Ce chiffre
        </tag>
    </tags>
    <segment schema="CT_coder2_1280479728969" start="11600" end="12513">
    <fullVersion>
      <cue type="Im_COREFdef">L'IDE (investissement direct étranger)</cue> est nécessaire
      pour mettre en valeur les ressources de la région. Pourtant, <cue type="Ia_COREFpro">il
      </cue> reste encore très faible. Parmi les pays en transition, l'Asie centrale est le
      parent pauvre du point de vue de <cue type="Im_COREFdef">l'IDE</cue>. La BERD a
      calculé, sur la période 1989-1999, que <cue type="Im_COREFdef">l'IDE par habitant</cue>
      avait été de 668 dollars pour les pays d'Europe centrale et orientale. Pour les pays de
      la CEI, <cue type="Ia_COREFdem">ce ratio</cue> était près de cinq fois inférieur,
      s'élevant à 140 dollars. Si on excepte le Kazakhstan qui a attiré près de 80 % de <cue
      type="Im_COREFdef">l'IDE</cue> en Asie centrale, <cue type="Ia_COREFdef_R">l'IDE</cue>
      est inférieur à 50 dollars par habitant. Malgré les hydrocarbures et les métaux, l'Asie
      centrale n'a reçu que 0,3 % <cue type="Im_COREFind">des IDE investis</cue> dans le
      monde sur la période 1998-2000 <cue type="Ia_COREFdem">Ce chiffre</cue> était nul dix
      ans plus tôt mais seuls les pays en développement du Pacifique sud ont attiré moins de
      capitaux que les pays d'Asie centrale sur cette période de trois années.<br><cue type=
      "paragraph"/>
    </fullVersion>
    <shortVersion>L'IDE (investissement direct étranger) est nécessaire ...</shortVersion>
    </segment>
    </CT>
    <context type="after"> L'investissement est faible. Les pays de la région ont ainsi ...</context>
  </structure>
```

Aligner les formats d'annotation

Besoin : ne pas perdre d'informations et effectuer des traitements homogènes

Proposition : format intermédiaire



Informations morpho-syntactiques

- ① Contribution théorique
- ② Contribution méthodologique
- ③ Contribution applicative**

Tâche de résolution de la coréférence

- ① Détecter les expressions référentielles
- ② Établir les liens de référence

Différents types de systèmes :

- Systèmes traitant seulement les liens de référence
 - Besoin d'avoir un texte déjà annoté en expressions référentielles
- Systèmes qui exécutent les deux étapes : end-to-end (de bout-en-bout)
 - Texte brut

Résolution de la coréférence

Différents types de systèmes (Wilkins et al., 2020)

- **Mention-pair :**

- classification binaire entre deux mentions
 - accord en genre et nombre
 - fonction syntaxique
 - distance entre les mentions
- construction des chaînes par transitivité

- **Easy-first :**

- construction des liens de coréférence, détection des types de relation entre les mentions
 - arbres syntaxiques
 - reconnaissance des entités nommées
 - informations ontologiques

- **Neural-network :** entraînent un modèle pour la classification des mentions

Metriques d'évaluation

⇒ **Aspects à évaluer** : nombre de liens de coréférence corrects, nombre des mentions correctes, nombre des classes ou entités correctes

Chaque metrique se focalise sur un de ces aspects

- MUC score (Vilain, Burger, Aberdeen, Connolly, & Hirschman, 1995) : liens de coréférence
- B³ score (Bagga & Baldwin, 1998) : mentions
- CEAF (Luo, 2005) : mentions ou entités, permet d'évaluer les systèmes end-to-end
- MELA (CONLL) (Denis & Baldrige, 2009) : union des trois métriques précédentes
- BLANC (Recasens & Hovy, 2011) : liens de coréférence et non-coréférence

⇒ **Scores divergents** : difficulté de comparer les performances des systèmes

⇒ **Évaluer qualitativement les erreurs des systèmes de résolution**

- ① **CROC (Désoyer, Landragin, & Tellier, 2015)** : mention-pair
 - texte déjà annoté en mention référentielles
 - distance et accord en genre et nombre entre une mention et l'antécédent candidat
 - développé à partir du corpus AnCor

- ② **DECOFRE (Grobol, 2020)** : end-to-end
 - texte brut
 - développé à partir du corpus AnCor

③ **ODACR (Oberle, 2019)** : système à base de règles

- informations syntaxiques : Talismane
- accord en genre et nombre
- information sémantiques tirées de Wikipédia
- relations sémantiques extraites d'un dictionnaire construit à partir de Glawi (Hatout & Sajous, 2016)

④ **coFR (Wilkins et al., 2020)** : end-to-end

- informations syntaxiques : StanfordNLP (Universal dependencies) (Qi et al., 2018)
- entité nommées identifiées avec Flair (Akbik et al., 2018)
- relations sémantiques extraites de WOLF (sagot and Fiser, 2008)
- utilisation des embeddings : BERT
- développé à partir des corpus AnCor et DEMOCRAT

[Le jour jaunâtre]₃ s' éteint . **[Il]**₃ fait tiède et **[fade]**₃ dans [la chambre] . **[Le nouveau-né]**₄ s' agite dans **[son]**₄ berceau . Bien que **[le vieux]**₅ ait laissé , pour entrer , **[ses]**₅ sabots] à [la porte] , **[son]**₄ pas] a fait craquer [le plancher] : **[l' enfant]**₄ commence à geindre . **[La mère]**₆ se penche hors de **[son]**₆ lit]₇ , afin de **[le]**₄ rassurer ; et **[le grand-père]**₈ allume **[la lampe]**₉ en tâtonnant , pour que **[le petit]**₄ n' ait pas

https://drive.google.com/file/d/1c_C9U1DbQj8i17HVppS-alkK39Tbgch2r/view?usp=sharing

point tout petit , mais infiniment tendre . **[L' enfant]₄** s' éveille et **[pleure]₄** . **[[Son]₄**
regard trouble] s' agite . [Quelle épouvante] ! [Les ténèbres] , [l' éclat brutal de **[la**
lampe]₉] , [les hallucinations d' **[un cerveau à peine dégagé de [le chaos]]₁₂**] , **[la**
nuit étouffante et grouillante]₁₃ **[qui]₁₃** **[l']₁₂** entoure , **[l' ombre sans fond]₁₄** d'
[où]₁₄ se détachent , comme [des jets aveuglants de [lumière]] , [des sensations aiguës] ,
[des douleurs] , [des fantômes] : **[ces figures énormes]₁₅** **[qui]₁₅** se penchent sur
[lui]₁₂ , **[ces yeux]₁₆** **[qui]₁₆** **[le]₁₂** pénètrent , **[qui]₁₆** s' enfoncent en **[lui]₁₂** , et qu'
[il]₁₂ ne comprend pas ! ... **[Il]₁₂** n' a pas [la force de crier] ; [la terreur] **[le]₁₂** cloue
immobile , **[les yeux]₁₆** , [la bouche ouverts] , soufflant de [le fond de [la gorge]] . **[[Sa]₁₂**
grosse tête boursouflée] se plisse de [grimaces lamentables et grotesques] ; [la peau de
[sa]₁₂ figure] et de **[ses]₁₂** mains]] est brune , violacée , avec [des taches jaunâtres] ... -
Bon Dieu ! qu' **[il]₁₂** est laid ! fit **[le vieux]₅** , d' [un ton convaincu] . **[Il]₅** alla reposer **[la**

https://drive.google.com/file/d/1c_C9U1DbQj8i17HVppS-alk39Tbgch2r/view?usp=sharing

- Unifier les formats
- Parser les corpus unifiés : Talismane ?
- Typer sémantiquement les maillons
- Appliquer le modèle de description et proposer une typologie des chaînes

Perspectives :

- **Contribution TAL** : Évaluer quels traits linguistiques sont pertinents pour améliorer les systèmes de résolution
- **Contribution didactique/psycholinguistique** : comparer les chaînes des textes d'experts à celles des textes d'apprenants



Références

- Antoine, J.-Y., Lefeuve, A., & Schang, E. (2016, juillet). Codage enchaîné ou en première mention de la coréférence : approcher la structure des chaînes de référence par comparaison des deux annotations . In 5ème Congrès Mondial de Linguistique Française (CMLF'2016).Tours, France.
- Ariel, M. (2001). Accessibility theory : An overview. Text representation : Linguistic and psycholinguistic aspects, 8, 29–87.
- Bagga, A., & Baldwin, B. (1998). Algorithms for scoring coreference chains. In The first international conference on language resources and evaluation workshop on linguistics coreference (Vol. 1, pp. 563–566).
- Charolles, M. (2002). La référence et les expressions référentielles en français. Editions Ophrys.
- Corblin, F. (1987). Indéfini, défini et démonstratif. Droz. Genève.
- Corblin, F. (1995). Les formes de reprise dans le discours. anaphores et chaînes de référence. Presses Universitaires de Rennes.
- Cornish, F. (1998). Les chaînes topicales : leur rôle dans la gestion et la structuration du discours. Cahiers de grammaire, 23(1), 9–40.
- Cornish, F. (2000). L'accessibilité cognitive des référents, le centrage d'attention, et la structuration du discours : une vue d'ensemble. Verbum, 22(1), 7–30.
- De Marneffe, M.-C., Recasens, M., & Potts, C. (2015). Modeling the lifespan of discourse entities with application to coreference resolution. Journal of Artificial Intelligence Research, 52, 445–475.
- Denis, P., & Baldridge, J. (2009). Global joint models for coreference resolution and named entity classification. Procesamiento del Lenguaje Natural, 42.
- Désoyer, A., Landragin, F., & Tellier, I. (2015, juin). Machine Learning for Coreference Resolution of Transcribed Oral French Data : the CROC System. Vingt-deuxième Conférence sur le Traitement Automatique des Langues 439-445.
- Gabardin, A., Ritschard, G., Studer, M., & Müller, N. S. (2009). Mining sequence data in r with the traminer package : A user's guide. Geneva : Department of Econometrics and Laboratory of Demography, University of Geneva.
- Guillot-Barbance, C., & Quignard, M. (2019). Chaînes de référence et structure textuelle dans les essais sur la peinture de diderot. Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics(25).
- Gundel, J. K., Hedberg, N., & Zacharski, R. (2019). Cognitive status and the form of referring expressions in discourse. In J. Gundel B. Abbott (Eds.), The oxford handbook of reference (pp. 67–99). New York : Oxford University Press.
- Kaiser, E., & Fedele, E. (2019). Reference resolution : A psycholinguistic perspective. In J. Gundel B. Abbott (Eds.), The oxford handbook of reference (pp. 309–336). New York : Oxford University Press.
- Kleiber, G. (1994). Anaphores et pronoms. Louvain-la-Neuve, Duculot.
- Kleiber, G. (2002). Marqueurs référentiels et théorie du centrage. Linx. Revue des linguistes de l'université Paris X Nanterre(47), 107–119.
- Landragin, F. (2018). Étude de la référence et de la coréférence : rôle des petits corpus et observations à partir du corpus mc4.

Références

- Landragin, F. (2015). Description, modélisation et détection automatique des chaînes de référence (democrat). *Bulletin de l'AFIA*(92), 11–15.
- 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).
- Luo, X. (2005, octobre). On coreference resolution performance metrics. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 25–32). Vancouver, British Columbia, Canada : Association for Computational Linguistics.
- Manuélian, H. (2003). Descriptions définies et démonstratives : analyses de corpus pour la génération de textes (Thèse de doctorat non publiée). (Thèse de doctorat dirigée par Riley, Philip et Pierrel, Jean-Marie Sciences du langage Nancy 2 2003)
- Mélanie-Becquet, F., & Landragin, F. (2014, septembre). Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques. *Langages*(195), 117-137.
- Mitkov, R. (2014). *Anaphora resolution*. Routledge.
- Oberle, B. (2019). Détection automatique de chaînes de coréférence pour le français écrit. *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*
- Oby, V., Glikman, J., Guillot-Barbance, C., & Pincemin, B. (2017). Les chaînes de référence dans les récits brefs en français : étude diachronique (xiii^e-xvii^e s.). *Langue française*(3), 91–110.
- Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. (2012). Fuille de données pour la stylistique : cas des motifs séquentiels émergents. *Journées Internationales d'Analyse Statistique des Données Textuelles* (821-833).
- Recasens, M., & Hovy, E. (2010). Coreference resolution across corpora : Languages, coding schemes, and preprocessing information. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1423–1432).
- Recasens, M., & Hovy, E. (2011). Blanc : Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4), 485–510.
- Roberts, C. (2019). Contextual influences on reference. In J. Gundel & B. Abbott (Eds.), *The oxford handbook of reference* (pp. 260–280). New York : Oxford University Press.
- Salazar Orvig, A. (2019). Reference and referring expressions in first language acquisition. In J. Gundel B. Abbott (Eds.), *The oxford handbook of reference* (pp. 283–308). New York : Oxford University Press.
- Salles, M. (2015). Chaînes de référence : la deuxième mention. l'exemple des entités inanimées dans les narrations littéraires. *Travaux de linguistique*(2), 111–133.
- Schneidecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. *Travaux de linguistique*(2), 85–133.
- Schneidecker, C. (2014). Chaînes de référence et variations selon le genre. *Langages*(3), 23–42.
- Schneidecker, C., & Landragin, F. (2014). Les chaînes de référence : présentation. *Langages*(195), 13–22.

Références

- Schnedecker, C., & Longo, L. (2012). Impact des genres sur la composition des chaînes de référence : le cas des faits divers. In 3ième congrès mondial de linguistique française (p. 1957-1972). Lyon, France.
- Todirascu, A., François, T., Bernhard, D., Gala, N., Ligozat, A.-L., & Khobzi, R. (2017, septembre). Chaînes de référence et lisibilité des textes : Le projet allusif. *Langue française*, 195(3), 35-52.
- Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In Sixth message understanding conference (MUC-6) : Proceedings of a conference held in Columbia, Maryland, november 6-8, 1995.
- Walker, M. A., Joshi, A. K., & Prince, E. F. (1998). *Centering theory in discourse*. Oxford University Press.
- Wilkens, R., Oberle, B., Landragin, F., & Todirascu, A. (2020). French coreference for spoken and written language. In *Language Resources and Evaluation Conference (LREC 2020)* (p. 80-89). Marseille, France.