Corpus analysis of coreference chains

Annotation and good-enough representations (in a french corpus)

Marine DELABORDE

CogSci Seminar

Everywhere, 17/02/21

Lattice

ANR

@demoCrat

Sorbonne Nouvelle

Presentation

- **Marine Delaborde**
  - Ph.D. in language sciences : defended on December 15, 2020
    - **Subject** : « Corpus analysis of coreference chains : non-strict coreference to the test of tool-based linguistics »
    - **Advisor** : Frédéric Landragin
    - **University** : Université de la Sorbonne Nouvelle (Paris, France)
    - **Laboratory** : Lattice (Montrouge)
    - **Project** : ANR Democrat
    - Available soon on TEL (in french) : https://tel.archives-ouvertes.fr
  - Now : ATER in NLP at Université de la Sorbonne Nouvelle
    - One-year contract involving teaching and research

Definitions

- **Reference** = link between a **referring expression** and the **entity** being designated by it

  → Discourse referent (Karttunen 1976)

- **Coreference** = link between **referring expressions** that designates the **same referent** (Corblin 1985)
- **Anaphora** = interpretation of the referent via an antecedent (Kleiber 2001)
- **Coreference chain** = all the coreferent expressions (**mentions**) that designate the **same referent**

  → Tracking of the discursive becoming of the referent throughout the text (Schnedecker 2019)

Example [1] : « **[Paul]**$_p$ est venu me voir : **[il]**$_p$ avait quelque chose à me demander. **[Cet étourdi]**$_p$ avait oublié **[son]**$_p$ manteau chez moi et **[il]**$_p$ voulait que je le **[lui]**$_p$ ramène. »

Traduction [1] : « **[Paul]**$_p$ came to see me : **[he]**$_p$ had something to ask me. **[This forgetful man]**$_p$ had left **[his]**$_p$ coat at my house and **[he]**$_p$ wanted me to bring it back to **[him]**$_p$. »

Problem statement

- **Coreference = strict / exact coreference** : the referent is exactly the same
- **Theorical framework**. What phenomena should be taken into account ?
  - The case of **fuzzy coreference** :
  - Example [2] : « **[Toutes ces femmes]$_f$** se tenaient d'un côté du salon comme un régiment en déroute, et de l'autre côté, entourée de Pauline, de sa mère et de quelques hommes de bon sens qui ne craignaient pas de causer respectueusement avec elle, Laurence siégeait comme une reine affable qui sourit à son peuple et le tient à distance. Les rôles étaient bien changés, et le malaise croissait d'un côté, tandis que la véritable dignité triomphait de l'autre. **[On]$_f$** n'osait plus chuchoter, **[on]$_f$** n'osait même plus regarder, si ce n'est à la dérobée. Enfin, quand le départ des plus désappointées eut éclairci les rangs, **[on]$_d$** osa s'approcher, mendier une parole, un regard, toucher, demander l'adresse de la lingère, le prix des bijoux, le nom des pièces de théâtre le plus à la mode à Paris, et des billets de spectacle pour le premier voyage qu'**[on]$_d$** ferait à la capitale. » George SAND, Pauline, 1881. DEMOCRAT.
- **Interpretation** (Ferreira et al. 2002) vs **annotation** : a corpus linguistics issue
- Fuzziness is part of language, but how to deal with it in a corpus annotation ?

Problem statement

- **Coreference = strict / exact coreference** : the referent is exactly the same
- **Theorical framework**. What phenomena should be taken into account ? What to do when a doubt persists ?
  - The case of **fuzzy coreference** :
  - Traduction [2] : « **[All these women]$_f$** stood on one side of the room like a routed regiment, and on the other side, surrounded by Pauline, her mother and a few men of good sense who where not afraid to chat respectfully with her, Laurence sat like an affable queen who smiled at her people and kept them at a distance. The roles were well changed, and discomfort grew on one side, while true dignity triumphed on the other. **[One]$_f$** no longer dared to whisper, **[one]$_f$** no longer even dared to look up, except in secret. Finally, when the departure of the most disappointed had cleared the ranks, **[one]$_d$** dared to approach, beg for a word, a look, a touch, ask for the address of the dressmaker, the price of jewellery, the names of the most fashionable plays in Paris, and tickets for the first trip **[one]$_d$** would make to the capital. » George SAND, Pauline, 1881. DEMOCRAT.
- **Interpretation** (Ferreira et al. 2002) vs **annotation** : a corpus linguistics issue
- Fuzziness is part of language, but how to deal with it in a corpus annotation ?

Contents

1. **Annotation methodology** for coreference chains
2. **Linguistic analysis** of non-strict and fuzzy coreference
3. **Considerations** on annotation of fuzzy coreference
4. **Discussion** : what status for the coreference annotator ?

Annotating the coreference chains : framework

- **ANR Democrat project (2016-2020)** : https://www.lattice.cnrs.fr/democrat/
  - Description and modeling of coreference chains : tools for corpus annotation (in diachrony and comparative languages) and automatic processing (Landragin 2016)
  - Collaboration between 4 french laboratories :
    - Lattice (Montrouge), LiLPa (Strasbourg), ICAR (Lyon) & IHRIM (Lyon)
  - Goal : developing research on language and textual structuring thanks to different outputs :
    - A **Corpus** published in 2019 : 58 blocks of text of about 10,000 words (about 200,000 mentions)
      - Helping to create an integrated and discursive model of the reference and more precisely of the construction of coreference chains
    - An **annotation and exploration tool** adapted to coreference (TXM extension URS)
    - An automatic **coreference resolution system**

➢ Bridging the fields of linguistics and NLP via corpus linguistics

Annotating the coreference chains : framework

- **Annotation procedure adopted in TXM :**

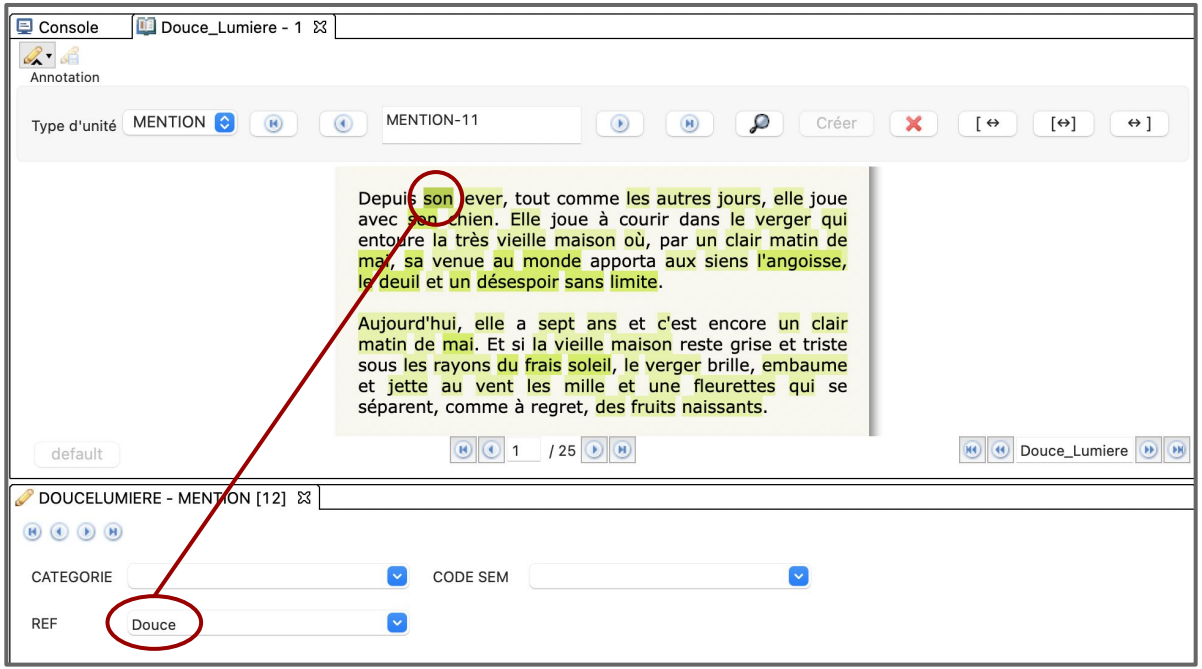
  1. **Manual** annotation of referents (labelling referring expressions)
  2. **Automatic** construction of chains according to the label of the referent

  ➢ All the referring expressions with the same referent label are in the same coreference chain

  - Annotators = 31 collaborators of the Democrat project (master students, PhD students, researchers, engineers, etc.)
    - 1 or 2 block / person on average

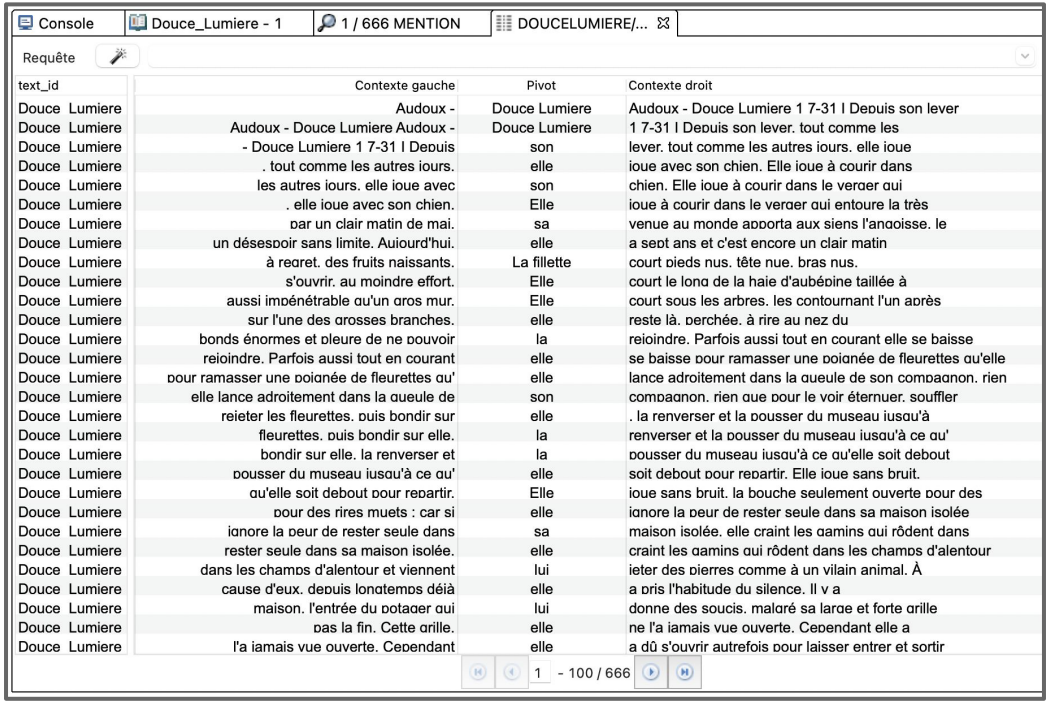
Annotating the coreference chains : framework

1. **Manual** annotation of referents :



Annotating the coreference chains : framework

2. **Automatic** construction of chains according to the label of the referent



text_id	Contexte gauche	Pivot	Contexte droit
Douce_Lumiere	Audoux -	Douce Lumiere	Audoux - Douce Lumiere 1 7-31 I Depuis son lever
Douce_Lumiere	Audoux - Douce Lumiere Audoux -	Douce Lumiere	1 7-31 I Depuis son lever. tout comme les
Douce_Lumiere	- Douce Lumiere 1 7-31 I Depuis	son	lever. tout comme les autres jours. elle joue
Douce_Lumiere	. tout comme les autres jours.	elle	joue avec son chien. Elle joue à courir dans
Douce_Lumiere	les autres jours. elle joue avec	son	chien. Elle joue à courir dans le verger qui
Douce_Lumiere	. elle joue avec son chien.	Elle	joue à courir dans le verger qui entoure la très
Douce_Lumiere	par un clair matin de mai.	sa	venue au monde apporta aux siens l'angoisse. le
Douce_Lumiere	un désespoir sans limite. Aujourd'hui.	elle	a sept ans et c'est encore un clair matin
Douce_Lumiere	à regret. des fruits naissants.	La fillette	court pieds nus. tête nue. bras nus.
Douce_Lumiere	s'ouvrir. au moindre effort.	Elle	court le long de la haie d'aubépine taillée à
Douce_Lumiere	aussi impénétrable qu'un gros mur.	Elle	court sous les arbres. les contournant l'un après
Douce_Lumiere	sur l'une des grosses branches.	elle	reste là. perchée. à rire au nez du
Douce_Lumiere	bonds énormes et pleure de ne pouvoir	la	rejoindre. Parfois aussi tout en courant elle se baisse
Douce_Lumiere	rejoindre. Parfois aussi tout en courant	elle	se baisse pour ramasser une poignée de fleurettes qu'elle
Douce_Lumiere	pour ramasser une poignée de fleurettes qu'	elle	lance adroitement dans la gueule de son compagnon. rien
Douce_Lumiere	elle lance adroitement dans la gueule de	son	compagnon. rien que pour le voir éternuer. souffler
Douce_Lumiere	rejeter les fleurettes. puis bondir sur	elle	. la renverser et la pousser du museau jusqu'à
Douce_Lumiere	fleurettes. puis bondir sur elle.	la	renverser et la pousser du museau jusqu'à ce qu'
Douce_Lumiere	bondir sur elle. la renverser et	la	pousser du museau jusqu'à ce qu'elle soit debout
Douce_Lumiere	pousser du museau jusqu'à ce qu'	elle	soit debout pour repartir. Elle joue sans bruit.
Douce_Lumiere	qu'elle soit debout pour repartir.	Elle	joue sans bruit. la bouche seulement ouverte pour des
Douce_Lumiere	pour des rires muets : car si	elle	ignore la peur de rester seule dans sa maison isolée
Douce_Lumiere	ignore la peur de rester seule dans	sa	maison isolée. elle craint les gamins qui rôdent dans
Douce_Lumiere	rester seule dans sa maison isolée.	elle	craint les gamins qui rôdent dans les champs d'alentour
Douce_Lumiere	dans les champs d'alentour et viennent	lui	jeter des pierres comme à un vilain animal. À
Douce_Lumiere	cause d'eux. depuis longtemps déjà	elle	a pris l'habitude du silence. Il y a
Douce_Lumiere	maison. l'entrée du potager qui	lui	donne des soucis. malgré sa large et forte grille
Douce_Lumiere	pas la fin. Cette grille.	elle	ne l'a jamais vue ouverte. Cependant elle a
Douce_Lumiere	l'a jamais vue ouverte. Cependant	elle	a dû s'ouvrir autrefois pour laisser entrer et sortir

1 - 100 / 666

Annotating the coreference chains : framework

- Selection of the phenomena to be annotated :
  - **Annotation manual** : https://www.lattice.cnrs.fr/democrat/files/ANR-15-CE38-0008-DEMOCRAT_livrable_L1_corpus.pdf
  - **Strict / Exact coreference** but also :
    - ➢ Possessive determiners : « **son** manteau » / « **her** coat »
    - ➢ Zero pronouns : « Il entra et ø commanda un café » / « He came in and ø ordered a coffee »

      → “maillons faibles”
    - ➢ Evolutive referent = one chain (discourse referent)

Annotating the coreference chains

- **Annotation produced for Democrat** :
  - **For the corpus** : 4 blocks of text of about 10,000 words annotated in references according to the recommendations of the Democrat annotation manual :

Tile	Author	Source	Date	Text type	Text Genre	Tool
Pauline	G. Sand	Wikisource	1881	Narrative	Novel	TXM
Le Diable au corps	R. Radiquet	Wikisource	1923	Narrative	Novel	Analec
Douce Lumière	M. Audoux	Wikisource	1937	Narrative	Novel	Analec
Est Républicain	-	Ortolang	2003	Non narrative	Press articles	Analec

  - **For the double annotation** (annotation as an aid to the calculation of the inter-rater reliability) : 2 block of text of about 2,000 words

Tile	Author	Source	Date	Text type	Text Genre	Tool
Elisabeth Seton	L. Conan	Wikisource	1881	Narrative	Biography	TXM
Aden Arabie	P. Nizan	ebooksgratuits	1931	Non narrative	Pamphlet	TXM

Annotating the coreference chains

- Example [1] :

  « **[Paul]**$_{p}$ est venu **[me]**$_{j}$ voir : **[il]**$_{p}$ avait [**quelque chose**]$_{q}$ à **[me]**$_{j}$ demander. **[Cet étourdi]**$_{p}$ avait oublié **[[son]**$_{p}$ **manteau]**$_{m}$ [**chez [moi]**$_{j}$]$_{h}$ et **[il]**$_{p}$ voulait que **[je]**$_{j}$ **[le]**$_{m}$ **[lui]**$_{p}$ ramène. »

  - **2 singletons** (1 mention) : q (quelque chose) & h (chez moi)
  - **3 chains** :
    - p (6 mentions) : Paul, il, Cet étourdi, son, il, lui
    - j (4 mentions) : me, me, moi, je
    - m (2 mentions) : son manteau, le

Annotating the coreference chains

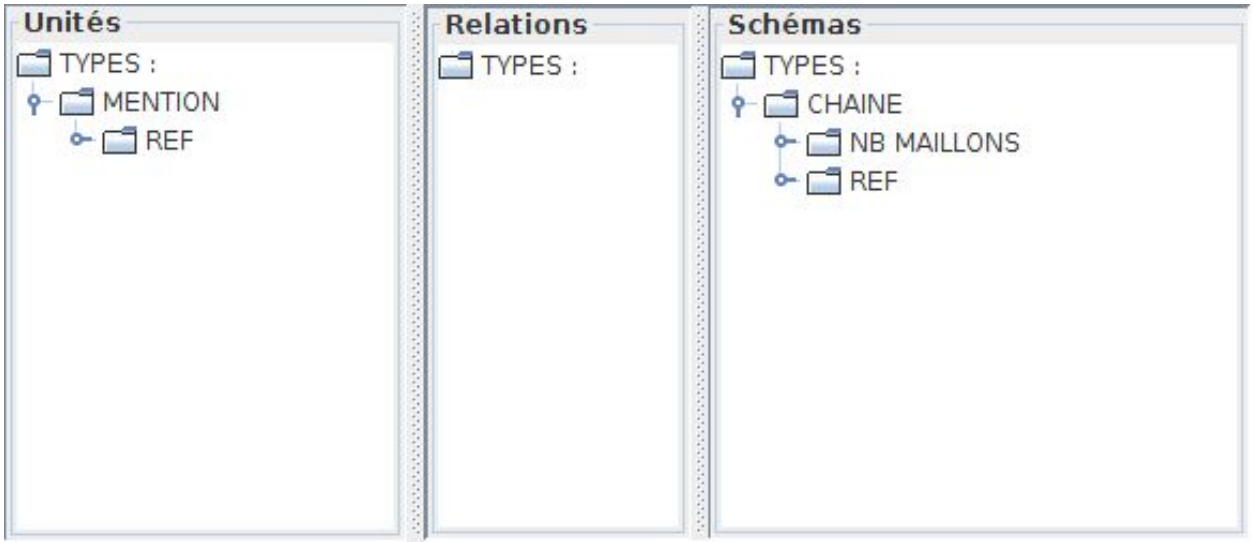
- Traduction [1] :

  « **[Paul]**$_p$ came to see **[me]**$_j$: **[he]**$_p$ had [**something**]$_q$ to ask **[me]**$_j$. **[This forgetful man]**$_p$ had left **[[his]**$_p$ **coat]**$_m$ at [**[my]**$_j$ **house**]$_h$ and **[he]**$_p$ wanted **[me]**$_j$ to bring **[it]**$_m$ back to **[him]**$_p$. »

  - **2 singletons** (1 mention) : q (something) & h (my house)
  - **3 chains** :
    - p (6 mentions) : Paul, he, This forgetful man, his, he, him
    - j (4 mentions) : me, me, my, me
    - m (2 mentions) : his coat, it

Annotating the coreference chains

- **URS in TXM**



Non-strict coreference

- **The choice of the referent**
  - **Salience** = selection criterion (Landragin 2005)
    - Clefts include several expressions that carry reference = source of salience

Exemple [3] « Absolument certain, répliqua **[Roger]**$_{r}$, puisque **[c']**$_{l}$est **[elle]**$_{l}$ **[qui]**$_{l}$ **[me]**$_{r}$ l'a dit » Adèle BOURGEOIS, Nemoville, 1917. DEMOCRAT.

Traduction [3] « Absolutely certain, replied **[Roger]**$_{r}$, since **[it]**$_{l}$ was **[she]**$_{l}$ **[who]**$_{l}$ told **[me]**$_{r}$ » Adèle BOURGEOIS, Nemoville, 1917. DEMOCRAT.

  - **Ambiguity** = choice between different alternatives (Fuchs 1996)
    - Candidate referents = mutually exclusive and not necessarily semantically close

Exemple [4] « **[Bouvard]**$_{b}$ **[l']**$_{l}$engagea à mettre bas sa redingote. **[Lui]**$_{b|l}$, **[il]**$_{b|l}$ se moquait du qu'en-dira-t-on ! » Gustave FLAUBERT, Bouvard et Pécuchet, 1881. DEMOCRAT.

Traduction [4] « **[Bouvard]**$_{b}$ urged **[him]**$_{l}$ to shed his topcoat. **[He]**$_{b|l}$ didn't mind what people would say ! » Gustave FLAUBERT, Bouvard et Pécuchet, 1881. DEMOCRAT.

Non-strict coreference

- **The choice of the referent**
  - **Abstract anaphora** (Asher, 1993) = ‘reference to abstract object in discourse’ (ex : events)
    - May even have a resumptive value :

Exemple [5] « Aux courses désordonnées s’étaient tout de suite ajoutés les jeux hardis et violents. Douce, légère et souple, suivait avec intérêt tous les mouvements de son camarade. Et derrière lui, elle faisait des culbutes savantes, franchissait des obstacles, grimpait jusqu’au faîte des arbres pour se nicher entre les feuilles ou se balancer entre les branches. Puis Noël se lassa de **[tout cela]**. » Marguerite AUDOUX, Douce Lumière, 1937. DEMOCRAT.

Traduction [5] « The discorderly races were immediately followed by bold and violent games. Douce, soft and flexible, followed with interest all the movements of his friend. And behind him, she was doing skilful somersaults, overcoming obstacles, climbing to the tops of trees to nestle between the leaves or swinging between the branches. Then Noël grew tired of **[all this]**. » Marguerite AUDOUX, Douce Lumière, 1937. DEMOCRAT.

Non-strict coreference

- **The choice of the referent**
  - **Near identity** = semantically close referents (Recasens 2010)
    - A detailed classification taking into account, for example, meronymy :

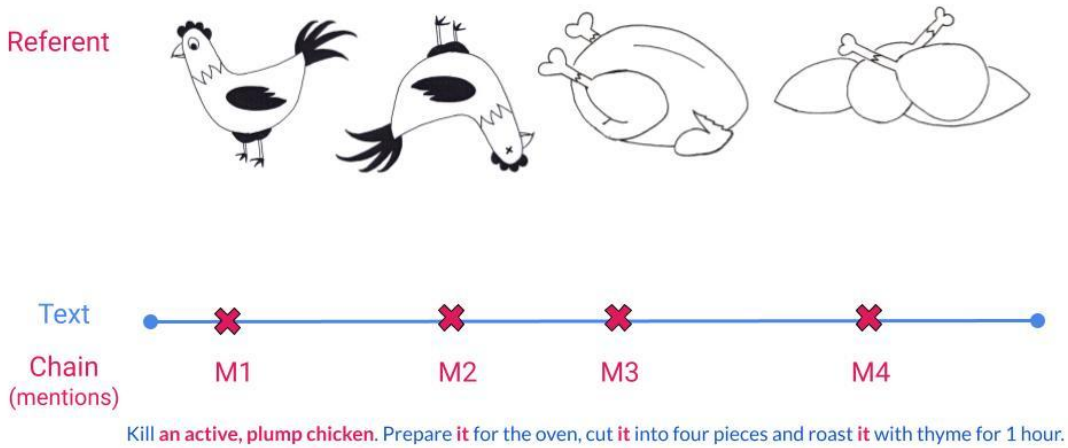
Exemple [6] « Ils ne sortaient pas sans **[leur louchet]**$_{I}$, et coupaient en deux les vers blancs, d'une telle force que **[le fer de [l'outil]**$_{I}$**]**$_{f}$ s'en enfonçait de trois pouces. » Gustave FLAUBERT, Bouvard et Pécuchet, 1881. DEMOCRAT

Traduction [6] « They would not come out without **[their draining spade]**$_{I}$, and cut the white worms in half, so forcefully that **[the iron of [the tool]**$_{I}$**]**$_{f}$ would sink three inches into them. » Gustave FLAUBERT, Bouvard et Pécuchet, 1881. DEMOCRAT

.

Non-strict coreference

- **The choice of the referent**
  - **Evolutive referent** = transformation of the referent : breaking point ? (Charolles et Schnedecker 1993)



(Brown & Yule 1983)

Non-strict coreference

- **The choice of the referent**
  - **Evolutive referent** = transformation of the referent : breaking point ? (Charolles et Schnedecker 1993)

Exemple [7] « Je connais un peu **[Debbie Harry]$_d$**. Une nuit, à **[ses]$_d$** débuts, je **[l']$_d$**avais emmenée voir la tour Eiffel. Et **[la blonde enfant]$_d$** s'était étonnée : "On ne peut pas monter la nuit ?" Mais il faisait un froid glacial, c'était à Paris et deux ans avant. Depuis, **[elle]$_d$** était devenue **[Madame Blondie]$_b$**, épousant **[son]$_b$** guitariste. Et **[elle]$_b$** avait vendu un million d'albums. Je pouvais toujours tenter de **[la]$_b$** joindre. » Philippe MANŒUVRE, L'Enfant du rock, 1985. FRANTEXT.

Traduction [7] « I know a little bit about **[Debbie Harry]$_d$**. One night, at **[her]$_d$** beginnings, I took **[her]$_d$** to see the Eiffel Towel. And **[the blonde child]$_d$** was astonished : "We can't go up at night ?" But it was freezing cold, it was in Paris and two years before. Since then, **[she]$_d$** had become **[Madame Blondie]$_b$**, marrying **[her]$_b$** guitarist. And **[she]$_b$** had sold a million albums. I could always try to reach **[her]$_b$**. » Philippe MANŒUVRE, L'Enfant du rock, 1985. FRANTEXT.

Non-strict coreference

- **The choice of the referent**
  - **Evolutive referent** = transformation of the referent : breaking point ? (Charolles et Schnedecker 1993)

Exemple [7] « Je connais un peu **[Debbie Harry]$_d$**. Une nuit, à **[ses]$_d$** débuts, je **[l']$_d$**avais emmenée voir la tour Eiffel. Et **[la blonde enfant]$_d$** s'était étonnée : "On ne peut pas monter la nuit ?" Mais il faisait un froid glacial, c'était à Paris et deux ans avant. Depuis, **[elle]$_d$** était devenue **[Madame Blondie]$_d$**, épousant **[son]$_d$** guitariste. Et **[elle]$_d$** avait vendu un million d'albums. Je pouvais toujours tenter de **[la]$_d$** joindre. » Philippe MANŒUVRE, L'Enfant du rock, 1985. FRANTEXT.

Traduction [7] « I know a little bit about **[Debbie Harry]$_d$**. One night, at **[her]$_d$** beginnings, I took **[her]$_d$** to see the Eiffel Towel. And **[the blonde child]$_d$** was astonished : "We can't go up at night ?" But it was freezing cold, it was in Paris and two years before. Since then, **[she]$_d$** had become **[Madame Blondie]$_d$**, marrying **[her]$_d$** guitarist. And **[she]$_d$** had sold a million albums. I could always try to reach **[her]$_d$**. » Philippe MANŒUVRE, L'Enfant du rock, 1985. FRANTEXT.

Non-strict coreference

- **Fuzzy (co)reference**
  - **Fuzzy reference** : fuzzy identification of the referent
    - *Referential opacity* (Quine 1977), *indirect denotation* (Frege 1892)
    - A matter of viewpoint regarding the categorisation of the referent
  - **Fuzzy coreference** : fuzziness about the coreference **relation**
    - *Anaphora with fuzzy antecedent* (Landragin 2007)
    - **Typical cases** :
      - Plural groups : fuzzy groups, fuzzy reference
      - In french, some pronouns :
        - « on » : undefined / generic / specific value
        - « ce » : underdetermination

Good enough interpretation of “on”

- **“on”** : *homo* (latin) → human being
  - Difficult to categorize : “personal”, “impersonal”, “indefinite” pronoun…
  - **Identification of the referent** : masc, fem, sing, plur
    - Precise : usually equal to “nous” or “je” + “tu” → « Elle **nous** a appelé et **on** est venus. » // “She called **us** and **we** came”
    - Generic : general truth value (as in proverbs) → « Quand **on** veut **on** peut. » // “When **you** want **you** can”
    - Indefinite : to avoid specifying the subject → « **On** m’a dit que tu étais là. » // “**I was told** you were there”
    - Impersonal : no referent → « **On** est mercredi. » // “**It**’s wednesday”
    - Fuzzy : vague referent, often a mix between precise/indefinite or precise/generic
      - ➢ An often denigrated use (also indefinite)
      - ➢ But really useful to report on the vagueness (in literature for example)

Good enough interpretation of “on”

- **“on”** : *homo* (latin) → human being
  - **A pronoun with no equivalent** (Fløttum et al. 2007) :
    - in English : “one”, “they”, “you”, “we”, or passive / impersonal structure (indefinite uses)
    - in Spanish : “uno” (indefinite), se (reflexive pronoun), first or third person
    - in German and Swedish : “**man**” (indefinite uses in particular)
- **Good-enough representations of language comprehension** (Ferreira et al. 2000, 2002) :
  - Experiments in psycholinguistics :
    - The structure of the passive sentences is fragile (atypical order : patient before agent)
    - Subjects are able to process certain expressions superficially

  ➢ Language comprehension - and thus the resolution of a (co)reference - is sometimes simply “good-enough”
    - A reader can handle fuzziness, but what about annotation ?
      - What depth of treatment for the annotation of referring expressions ? (Charolles 2014)

Fuzzy coreference

Exemple [8]

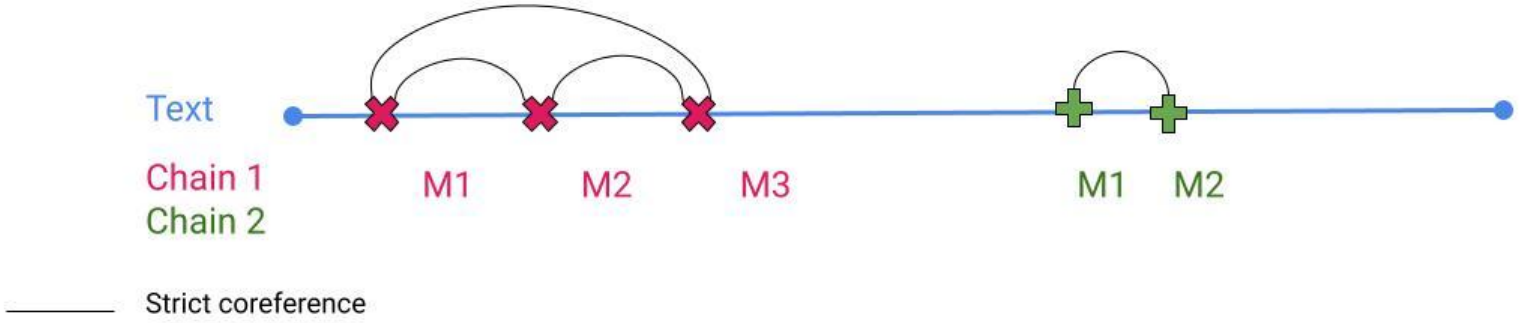
« La classe finie, d'autres tourments **[l']**$_{d}$attendaient sur la route qu'**[elle]**$_{d}$ suivait en compagnie de **[filles et garçons regagnant leur demeure]**$_{f}$. Toutes les malices étaient bonnes à faire à **[cette gnangnan]**$_{d}$ **[qui]**$_{d}$ ne se défendait pas et ne se **[méfiait]**$_{d}$ jamais. **[On]**$_{o1}$ **[la]**$_{d}$ poussait brusquement dans un fossé vaseux, ou dans un buisson plein d'épines d'où **[elle]**$_{d}$ sortait salie et déchirée. Quand vint la neige, **[elle]**$_{d}$ fut toute désignée pour recevoir les boules, qu'**[on]**$_{o2}$ **[lui]**$_{d}$ jetait de préférence au visage. **[Elle]**$_{d}$ pensait à Noël. S'il était là, il saurait bien **[la]**$_{d}$ défendre. Mais la ferme des Barray était peu éloignée du village, et Noël n'avait rien à faire sur la route qui conduisait au Verger, distant de plus d'un kilomètre. Il y avait bien Marguerite Dupré, une grande qui prenait parfois **[sa]**$_{d}$ défense, mais alors c'était elle qu'**[on]**$_{o3}$ attaquait, Marguerite Dupré, dont la maison n'était pas très éloignée de celle de **[la petite]**$_{d}$, prenait, en même temps qu'**[elle]**$_{d}$, le même sentier. Mais, arrivée là, **[Douce]**$_{d}$ ne craignait plus rien, **[elle]**$_{d}$ courait plus vite qu'une oie et **[devançait]**$_{d}$ facilement **[les méchants]**$_{m}$. » Marguerite AUDOUX, Douce Lumière, 1937.

Fuzzy coreference

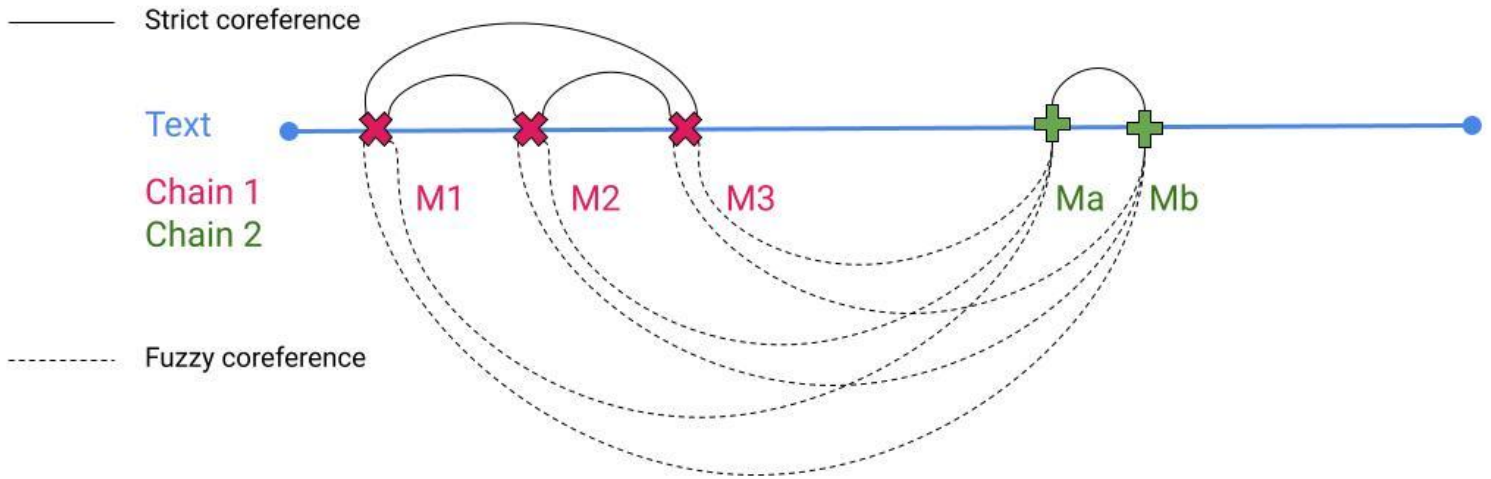
Traduction [8]

« Once the class was over, other torments awaited **[her]**$_{d}$ on the road **[she]**$_{d}$ was following in the company of **[girls and boys returning home]**$_{f}$. All kinds of mischief were good for **[namby-pampy]**$_{d}$ **[who]**$_{d}$ never defended **[herself]**$_{d}$ and never **[mistrusted]**$_{d}$ anyone. **[One]**$_{o1}$ would suddenly push **[her]**$_{d}$ into a muddy ditch, or into a bush full of torns from which **[she]**$_{d}$ would come out dirty and torn. When the snow came, **[she]**$_{d}$ was the perfect target for the balls, which **[they?]**$_{o2}$ preferably threw in **[her]**$_{d}$ face. **[She]**$_{d}$ was thinking about Noël. If he were there, he would know how to defend **[fer]**$_{d}$. But the Barray farm was not far from the village, and Noël had nothing to do on the road leading to the orchard, more than a kilometer away. There was indeed Marguerite Dupré, a tall girl who sometimes defended **[her]**$_{d}$ but then it was she that **[they?]**$_{o3}$ attacked. Marguerite Dupré, whose house was not very far from **[the little girl]**$_{d}$'s house, took at the same moment the same path as **[her]**$_{d}$. But once there, **[Douce]**$_{d}$ was safe, **[she]**$_{d}$ ran faster than a goose and easily **[beat]**$_{d}$ **[the villains]**$_{m}$. » Marguerite AUDOUX, Douce Lumière, 1937.

Strict coreference



Fuzzy coreference between 2 chains



Non-strict coreference annotation in corpus

- **Annotation → making choices** : What phenomena ? How to annotate them ?
- **Some cases of non-strict coreference considered in corpus**
  - **ARRAU** (Poesio et Artstein 2008) & **Phrase Detective** (Chamberlain, Poesio et Kruschwitz 2016) : ambiguity
  - **ACE** (Doddington et al. 2004) : 5 types of relation + coreference et metonymy
  - **OntoNotes** (Pradhan et al. 2011) : identical coreference / appositive coreference
  - **WikiCoref** (Ghaddar et Langlais 2016) : identical coreference / attributive coreference / attributive coreference in copulative constructions
  - **NIDENT** (Recasens et al. 2010) : near identity typology
    - **Polish Coreference Corpus** (Ogrodniczuk, Glowinska et al. 2014) : identical coreference / near-identical coreference - inspiration NIDENT
  - **ANCOR** (Muzerelle, Lefeuvre, Antoine et al. 2013) : direct / indirect / pronominal / associative / pronominal associative anaphora
  - (Dipper et Zinsmeister 2011, 2012) : abstract / concrete anaphora

Non-strict coreference annotation in corpus

- **In Democrat** : strict framework but different annotation procedures
  - Different approaches of the annotation of fuzzy coreference (sub-corpus for the analysis = 30 texts):

    1. Grouping generic referents
    2. Grouping undefined referents
    3. Grouping all the “on” pronouns
    4. Grouping generic referents according to the text structure
    5. **Grouping generic or fuzzy expressions only if they are coreferent**
    6. Not grouping (or annotating) expressions whose referent is generic or undefined
    7. Alway identify a precise referent

  - Possible drifts :

    1. **An over-grouping of mentions** : the annotation of all generic and/or fuzzy referents in a single coreference chain even though not all these mentions are coreferent (approaches 1, 2, 3, 4 and sometimes 7). This also often means to annotate together mentions whose coreference relation is fuzzy.

    2. **An under-grouping of mentions** : the annotation, in different chains, of semantically linked referring expressions because they have a fuzzy coreference relation (approaches d’annotation 6 and sometimes 7).

Non-strict coreference annotation in corpus

Titre	1	2	3	4	5	6	7
Aden Arabie	✓	-	-	-	-	-	-
Articles Wiki	-	-	-	✓	-	-	-
Bouvard et Pécuchet	✓	✓	-	-	-	-	-
Code Civil 1	-	-	-	-	-	✓	-
Code Civil 2	-	-	-	-	-	✓	-
Code de procédure pénale	-	-	-	-	-	-	-
Convention univ	-	-	-	-	-	-	-
Convention marin	-	-	-	-	-	✓	-
Convention aéro	-	-	-	-	-	-	-
Convention thon	-	-	-	-	-	✓	-
Douce Lumière	-	-	-	-	✓	-	-
De la ville au moulin	-	✓	-	-	-	-	-
Élisabeth Seton	-	-	-	-	-	✓	-
Est Républicain 1	-	-	-	-	-	-	✓
Est Républicain 2	-	-	-	-	✓	-	-
Génie du christianisme	-	-	-	-	✓	-	-
Jean-Christophe 1	-	-	✓	-	-	-	-
Jean-Christophe 2	-	-	✓	-	-	-	-
Madame de Hautefort	-	-	✓	-	-	-	-
Mademoiselle Fifi 1	-	-	-	✓	-	-	-
Mademoiselle Fifi 2	-	-	-	✓	-	-	-
Mademoiselle Fifi 3	-	-	-	✓	-	-	-
La morte amoureuse	✓	-	-	-	-	-	-
Le capitaine Fracasse	✓	-	-	-	-	-	-
Le Diable au corps	✓	-	-	-	-	-	-
Le ventre de Paris	-	✓	-	-	-	-	-
Nemoville	-	✓	-	-	-	-	-
Pauline	-	-	-	-	✓	-	-
Rosalie de Constant	-	-	-	-	-	-	✓
Sarrasine	✓	-	-	-	-	-	-
TOTAL	**6**	**4**	**3**	**4**	**4**	**5**	**2**

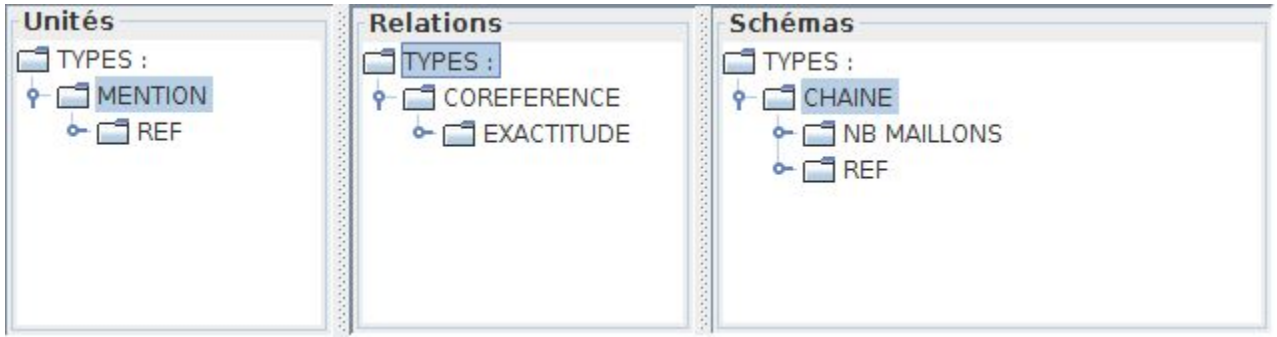
Tableau 5.1 – Sous-corpus de Democrat : les différentes conduites d'annotation.

Non-strict coreference annotation in corpus

- **Recommandations**
  - Some precisions in the annotation manual
    1. Distinction between the notions of generality, indefiniteness and fuzziness
    2. Distinction of each referent at the time of the annotation : not all the generic “on” are necessarily coreferent
    3. Distinction between labels and referents that are not coreferent : chains are built according to these labels
    4. Distinction between strict and fuzzy coreference for the “on” pronoun : not all the “on” are fuzzy
    5. Distinction between strict and fuzzy coreference and define how to annotate them : whether or not the fuzziness is taken into account
  - Caution with the structure of texts : the problem of text concatenation = coreference anyway ?

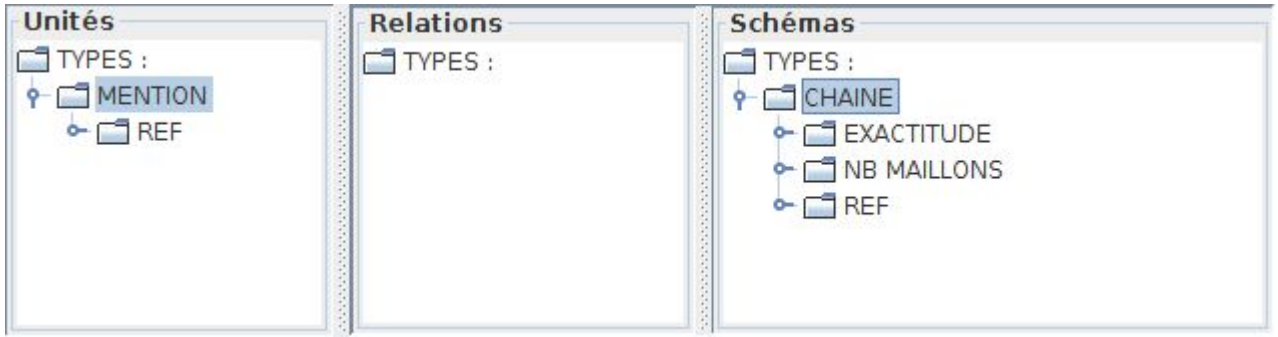
Non-strict coreference annotation in corpus

- **Recommandations**
  - Properties adapted to the fuzzy coreference in the annotation scheme
    - An “exactitude” feature to qualify the coreference **relation** between the units : strict or fuzzy



Non-strict coreference annotation in corpus

- **Recommandations**
  - Properties adapted to the fuzzy coreference in the annotation scheme
    - An “exactitude” feature to qualify the coreference in the **scheme** (chains) : strict or fuzzy



Expert or non-expert annotation ?

- What status of the annotation in the production of this type of corpus ?
    - **« Non-expert »** :
        - Interpretation not influenced by any linguistic questioning
    - **« Expert »** :
        - Accurate linguistic modelling
        - Can nevertheless generate errors
- **Psycholinguistic experimentation** with Lucie Rousier-Vercruyssen :
    - Interpretation of the “on” pronoun in different contexts
- **Work with third year students in languages sciences** (corpus linguistics course) :
    - Categorisation of different examples of generic and undefined “on”
    - Basic calculation of the inter-rater agreement

Conclusion

- **Corpus analysis of coreference chains**
  - **Annotation** : raising awareness of the phenomenon of non-strict / fuzzy coreference
  - **Analysis** of the fuzzy coreference annotation in the Democrat corpus
    - **Categorisation** of annotation drifts
  - **Recommendations** for taking this phenomenon into account

Main contributions (in french)

- **Article** : Marine Delaborde, Frédéric Landragin. En quoi le pronom « on » a-t-il une valeur anaphorique ? Le cas des successions d'occurrences de « on ». *Les cahiers de praxématique*, Montpellier : Presses universitaires de la Méditerranée, 2006-, 2019, La gestion de l'anaphore en discours : complexités et enjeux, 72, pp.1-18. ⟨hal-02161902⟩ : **(co)referential complexity of the "on" pronoun**
- **Presentations** :
  - Marine Delaborde, Frédéric Landragin. De la coréférence exacte à la coréférence complexe : une typologie et sa mise en œuvre en corpus. *10èmes Journées internationales de Linguistique de Corpus*, Université Grenoble Alpes, Nov 2019, Grenoble, France. ⟨hal-02286100⟩ : **proposition of typology of fuzzy coreference for the annotation**
  - Marine Delaborde, Frédéric Landragin. En quoi le pronom "on" a-t-il une valeur anaphorique? Le cas des successions d'occurrences de "on". *Gérer L'Anaphore en Discours (GLAD 2018) : vers une approche interdisciplinaire / Managing Anaphora in Discourse : towards an interdisciplinary approach* , Apr 2018, Grenoble, France. ⟨halshs-01795213⟩ : **(co)referential complexity of the "on" pronoun**
  - Marine Delaborde, Frédéric Landragin. Traitement " good-enough " du pronom " on " : vers une modélisation de la coréférence floue. *Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS) 2018*, Jan 2018, Paris, France. ⟨halshs-01795228⟩ : **(co)referential complexity of the "on" pronoun : understanding vs annotation**
  - Frédéric Landragin, Marine Delaborde. Faut-il compter ou ignorer les occurrences de « ce » dans les chaînes de coréférences ?. *Ce disant, que fait-on ? Aspects grammaticaux et discursifs de ce en français*, Université de Strasbourg, 2018, Strasbourg, France. ⟨halshs-01836380⟩ : **(co)referential complexity of the "on" pronoun**
- **Poster :** Frédéric Landragin, Marine Delaborde, Yoann Dupont, Loïc Grobol. Description et modélisation des chaînes de référence. Le projet ANR Democrat (2016-2020) et ses avancées à mi-parcours. *Cinquième édition du Salon de l'Innovation en TAL (Traitement Automatique des Langues) et RI (Recherche d'Informations)*, May 2018, Rennes, France. 2018. ⟨hal-01797982⟩ **: Democrat project presentation**

Perspectives

- **Langages - thematique issue :** « La coréférence floue dans le corpus Democrat »
  traduction : Fuzzy coreference in Democrat
  - Typology and recommandations
- **Annotation of the « exactitude »** feature in the Democrat corpus with TXM
  - Validation of the annotation scheme proposed by inter-annotation agreements
  - Consideration of fuzzy coreference
    - Gain of **semantic informations** in a corpus (compared to a strict coreference annotation)
    - **Difficulties in calculating inter-rater agreement**
    - Addition of a feature in the annotation structure, **possible** in Democrat :
      - For the annotation and the coreference detection (NLP)

References

- Asher, N. (1993). *Reference to Abstract Objects in Discourse* (Springer). https://www.springer.com/gp/book/9780792322429
- Chamberlain, J., Poesio, M., & Kruschwitz, U. (2016). Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2039-2046.
- Charolles, M., & Schnedecker, C. (1993). Coréférence et identité : Le problème des référents évolutifs. *Langages*, *27*(112), 106-126. https://doi.org/10.3406/lgge.1993.1664
- Corblin, F. (1985). Remarques sur la notion d'anaphore. *Revue québécoise de linguistique*, *15*(1), 173-195.
- Dipper, S., & Zinsmeister, H. (2011). *Towards a standard for annotating abstract anaphora*. 6.
- Dipper, S., & Zinsmeister, H. (2012). Annotating abstract anaphora. *Language Resources and Evaluation*, *46*(1), 37-52. https://doi.org/10.1007/s10579-011-9160-1
- Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 837-840.
- Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, *11*(1), 11-15.
- Fløttum, K., Jonasson, K., & Norén, C. (2007). On : Pronom à facettes (De Boeck/Duculot).
- Frege, G. (1892). *Über Sinn und Bedeutung* (1. Auflage). Pfeffer.
- Fuchs, C. (1996). *Les ambiguïtés du français*. Ophrys.
- Ghaddar, A., & Langlais, P. (2016). *WikiCoref : An English Coreference-annotated Corpus of Wikipedia Articles*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.

References

- Kleiber, G. (2001). *L'anaphore associative*. Puf.
- Landragin, F. (2005). *Traitement automatique de la saillance*. 263-272.
- Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, 11-15.
- Landragin, F. (2007). L'anaphore à antécédent flou : Une caractérisation et ses conséquences sur l'annotation des relations anaphoriques. *Journée d'étude de l'Association pour le Traitement Automatique des LAngues (ATALA) sur la résolution des anaphores*, 3.
- Muzerelle, J., Lefeuvre, A., Antoine, J.-Y., Schang, E., Maurel, D., Villaneau, J., & Eshkol, I. (2011). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA (Éd.), *20e conférence sur le Traitement Automatique des Langues Naturelles* (p. 555-563). ATALA.
- Poesio, M., & Artstein, R. (2008, janvier 1). *Anaphoric Annotation in the ARRAU Corpus.*
- Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., & Xue, N. (2011). CoNLL-2011 Shared Task : Modeling Unrestricted Coreference in OntoNotes. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning : Shared Task*, 1-27.
- Quine, W. V. O., Dopp, J., & Gochet, P. (1977). *Le mot et la chose*. Flammarion.
- Recasens, Marta, Hovy, E., & Martí, A. (2010, juin 19). *A Typology of Near-Identity Relations for Coreference (NIDENT)*.
- Recasens, Martha. (2010). *Coreference : Theory, Resolution, Annotation and Evaluation*. University of Barcelona.
- Schnedecker, C. (2019). De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence. *Les cahiers de praxématique*. https://hal.archives-ouvertes.fr/hal-02317889