



# À LA CROISÉE DE LA CORÉFÉRENCE ET DES EXPRESSIONS POLYLEXICALES

---

Anaëlle PIERREDON

Jianying LIU

Encadrants du stage :

Agata Savary

Jean-Yves Antoine

Anaïs Halftermeyer

# Définitions

## Coréférence

Procédé linguistique dans lequel plusieurs expressions réfèrent à un même élément du discours.

« Il a ouvert **la porte** et l'a refermée. »

## Expressions polylexicales

- Utilisation fréquente
- Termes complexes composés de plusieurs mots
- non-compositionnalité sémantique

**Composants** : tokens annotés comme appartenant à une expression polylexicale.

# Hypothèse de départ



Les composants individuels d'une expression polylexicale sont rarement susceptibles d'appartenir à des chaînes de coréférences.

Exemple:

« Il a retourné sa veste et l'a suspendue dans l'armoire. »

# Types d'expressions polylexicales

- **LVC** (Light Verb Construction)
- **IRV** (Inherently reflexive verbs)
- **MVC** (Multi-verb constructions)
- **VID** (Verbal Idioms)

LVC.full : Faire une présentation

LVC.cause : donner un conseil

s'apercevoir, s'évanouir

laisser tomber, vouloir dire

l'emporter, se faire des idées

Caractéristique spécifique : non-compositionnalité sémantique




# PLAN

## I - Traitements

1. Méthode
2. Choix des corpus
3. Chaîne de traitements

## II - Résultats

1. Les différents cas
  2. Format de sortie
  3. Analyse des résultats
- 



# Traitements

1. Méthode et Outils utilisés
2. Choix des corpus
3. Chaîne de traitements

# Méthode et Outils utilisés

## 1. Repérage des candidats:

Seen2Seen

pour expression polylexicale

OFCORS

pour chaîne de coréférence

## 2. Analyse manuelle des résultats

# Seen2Seen

- Caroline Pasquer
- Annotation en expressions polylexicales
- Méthode symbolique interprétable

# OFCORS

- Théo Azzouza
- Annotation en chaînes de coréférence





# Corpus

1ère expérience: sous-corpus

Sequoia de PARSEME

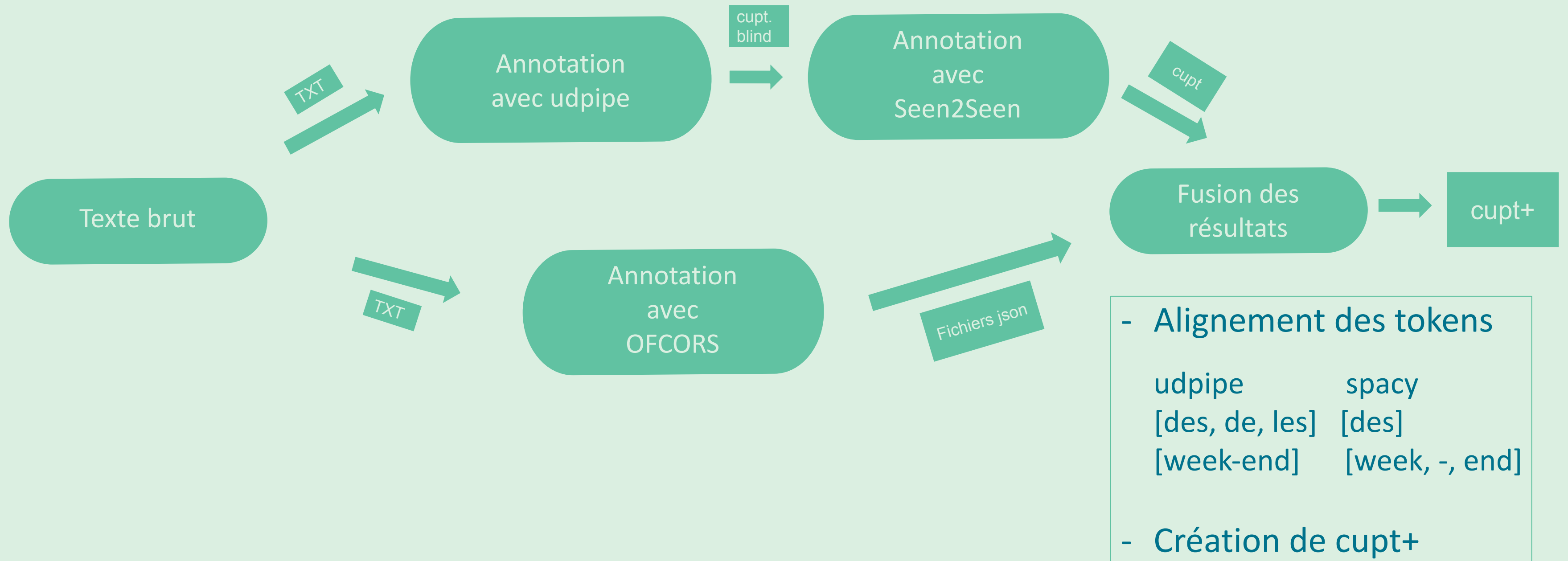
2 emea + 19 frwiki

2e expérience: Est Républicain 2003

total: 11 770 articles

notre essai : 100 articles plus de 300  
mots

# Chaîne de traitements





# Résultats

1. Les différents cas
2. Format de sortie
3. Analyse des résultats

# CAS 1 - Expression comprise dans la mention

TOKENS	MWE	MENTION
chez	*	*
7	*	1
patients	*	1
atteints	1	1
de	*	1
la	*	1
maladie	1	1
de	*	1
Paget	*	1
.	*	*

« L'histologie osseuse a été évaluée 6 mois après le traitement par 5 mg d'acide zolédronique chez 7 patients atteints de la maladie de Paget. »

# CAS 2 - Identiques

TOKENS	MWE	MENTION
[...]	*	*
mise	1	1
en	1	1
évidence	1	1
[...]	*	*

# CAS 3 - Mention comprise dans l'expression

TOKENS	MWE	MENTION
De	*	*
nombreux	*	*
protagonistes	*	*
ont	*	*
trouvé	1	*
la	1	1
mort	1	1
depuis	*	*
...	...	...

« De nombreux protagonistes ont trouvé la mort depuis la signature du contrat. »

# CAS 4 - Chevauchement

TOKENS	MWE	MENTION
est	*	*
pris	1	*
en	1	*
flagrant	1	1
délit	1	1
d'	*	1
extorsion	*	1
de	*	1
fonds	*	1

« Ce dernier est  
pris en flagrant  
délit d'extorsion  
de fonds. »

# CAS 4 - Chevauchement

TOKENS	MWE	MENTION
...	...	...
semblent	*	*
avoir	*	*
subi	1	*
des	*	1
retouches	1	1
...	...	...



# Format de sortie

```
"LVC.full": {  
  "TYPE": "LVC.full",  
  "COREF": "17/79",  
  "MWES": [  
    ...  
    {  
      "FICHIER": "frwiki_1_mwe_coref.cupt",  
      "PHRASE": "combat contre les institutions, mené sans relâche,  
qui dura près de deux décennies et qui, en dehors de son activité  
d'institutrice, l'occupa à plein temps.",  
      "TOKENS": "['combat', 'mené']",  
      "COREF": "['47:224', '*']",  
      "CAS": "{ '224': 3 }",  
      "CHAÎNE(S)": {  
        "47": "{ '219': ['un', 'combat'], '224': ['combat'] }"  
      }  
    }  
    ...  
  ]  
}
```

# Analyse des résultats

- Les composants des MVC (*vouloir dire*) et des IRV (*s'apercevoir*) ne se trouvent jamais dans des chaînes de corréférence
- Les seules chaînes de corréférence correctes observées contiennent des composants de LVC

**Chaîne correcte : chaîne de corréférence vérifiée manuellement et pour laquelle les composants semblent effectivement être corréférents.**

# Analyse des résultats : les LVC

Corpus	Chaînes correctes intersectées avec des LVC	Total de LVC
EMEA	3 + 0	184 + 7
FRWIKI	1 + 0	79 + 1
EST RÉPUBLICAIN	3 + 0	122 + 3 selon Seen2seen

## Exemple :

« Aclasta ne doit être utilisé, chez les patients souffrant de la maladie osseuse de Paget, que par un médecin expérimenté dans le traitement de cette maladie. »

# Analyse des résultats : les VID

## Exemple d'erreur :

« Après plusieurs mois, **la Chine** met **une sourdine** à ses objections et le ministère français des Affaires étrangères lève son **veto**. »

## Exemple limite :

« Créé par la Fédération nationale qui perpétue le souvenir de l'homme d'Etat meusien qui fut ministre de la Guerre et l'initiateur d'un système de défense qui porte **son nom**, le prix **André-Maginot** récompense des travaux liés au civisme et au devoir de mémoire. »

Corpus	Nombre total de VID
EMEA	32
FRWIKI	98
ER	302 selon Seen2seen



# Conclusion

Le niveau de compositionnalité est différent selon les types des expressions

**Pour invalider notre hypothèse il faudrait donc observer une chaîne de coréférence correcte contenant un composant d'un VID et la validation devra se faire sur le plus grand corpus possible.**

Ensuite :

- Corpus ANCOR
- Distribution des différents cas selon le type
- Procédé de validation de l'hypothèse

# Références

- Guide d'annotation de PARSEME : <https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/>
  - Corpus PARSEME FR: [https://gitlab.com/parseme/parseme\\_corpus\\_fr](https://gitlab.com/parseme/parseme_corpus_fr)
  - Corpus Est Républicain : <http://redac.univ-tlse2.fr/corpus/estRepublicain.html>
  - Outil Seen2Seen : [https://gitlab.com/cpasquer/st\\_2020](https://gitlab.com/cpasquer/st_2020)
  - Outil OFCORS : <https://gitlab.com/Stanoy/ofcors/-/tree/master>
  - Outil DECOFRE : <https://github.com/LoicGrobol/decofre>
- 
- Pasquer C., Savary A., Ramisch C., Antoine J.-Y. (2020) Verbal Multiword Expression Identification: Do We Need a Sledgehammer to Crack a Nut?. *The 28th International Conference on Computational Linguistics (COLING-20)*, Barcelona, Spain. [⟨hal-03013636⟩](#)
  - Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., ... & Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (pp. 107-118). [⟨hal-03014927⟩](#)
  - Desoyer A., Landragin F., Tellier I., Lefeuvre A., Antoine J.-Y. (2014) Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR, *Traitement Automatique des Langues, TAL*, vol. 55 [⟨halshs-01153297⟩](#)

# Merci



[Git du projet](#)



# cuprt+

```
# newdoc = 2003-01-02_69.txt
# source_sent_id = http://my/newcorpus/uri 2003-01-02_69.txt 1
# text = Décisions du Conseil A l'unanimité, nos édiles donnent leur accord sur la modification des statuts de la communauté de communes concernant : -
l'amélioration du traitement des déchets pour les points-tris, la déchetterie, voire la collecte sélective.
1  Décisions  décision  NOUN  _  Gender=Fem|Number=Plur  11  advcl  _  _  *  1;2  *
2-3  du  _  _  _  _  _  _  _  *  *  *
2  de  de  ADP  _  _  4  case  _  _  *  1;2;3  *
3  le  le  DET  _  Definite=Def|Gender=Masc|Number=Sing|PronType=Art  4  det  _  _  *  1;2;3  *
4  Conseil conseil  NOUN  _  Gender=Masc|Number=Sing  1  nmod  _  SpacesAfter='\n  *  1;2;3  *
5  A  à  ADP  _  _  7  case  _  _  *  2  *
6  l'  le  DET  _  Definite=Def|Gender=Fem|Number=Sing|PronType=Art  7  det  _  SpaceAfter=No  *  2;4;5  *
7  unanimité  unanimité  NOUN  _  Gender=Fem|Number=Sing  1  nmod  _  SpaceAfter=No  *  2;4;5  *
8  ,  ,  PUNCT  _  _  1  punct  _  _  *  2;5  *
9  nos  son  DET  _  Gender=Masc|Number=Plur|Poss=Yes|PronType=Prs  10  det  _  _  *  2;5;6  *
10  édiles  édile  NOUN  _  Gender=Masc|Number=Plur  11  nsubj  _  _  *  2;5;6  *
11  donnent  donner  VERB  _  Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin  0  root  _  _  1:LVC.full
*  *
12  leur  son  DET  _  Gender=Masc|Number=Sing|Poss=Yes|PronType=Prs  13  det  _  _  *  7  *
13  accord  accord  NOUN  _  Gender=Masc|Number=Sing  11  obj  _  _  1  7  *
14  sur  sur  ADP  _  _  16  case  _  _  *  *  *
15  la  le  DET  _  Definite=Def|Gender=Fem|Number=Sing|PronType=Art  16  det  _  _  *  8  *
16  modification  modification  NOUN  _  Gender=Fem|Number=Sing  11  obl  _  _  *  8  *
17-18  des  _  _  _  _  _  _  _  *  *  *
17  de  de  ADP  _  _  19  case  _  _  *  8;9  29:9
18  les  le  DET  _  Definite=Def|Gender=Masc|Number=Plur|PronType=Art  19  det  _  _  *  8;9  29:9
19  statuts  statut  NOUN  _  Gender=Masc|Number=Plur  16  nmod  _  _  *  8;9  29:9
20  de  de  ADP  _  _  22  case  _  _  *  8;9  29:9
21  la  le  DET  _  Definite=Def|Gender=Fem|Number=Sing|PronType=Art  22  det  _  _  *  8;9;10  29:9;7:10
22  communauté  communauté  NOUN  _  Gender=Fem|Number=Sing  19  nmod  _  _  *  8;9;10  29:9;7:10
23  de  de  ADP  _  _  24  case  _  _  *  8;9;10;11  29:9;7:10
24  communes  commune  NOUN  _  Gender=Fem|Number=Plur  22  nmod  _  _  *  8;9;10;11;12  29:9;7:10
25  concernant  concerner  VERB  _  Tense=Pres|VerbForm=Part  22  acl  _  _  *  *  *
```



# Sortie Seen2Seen

cupt.blind

```
# source_sent_id = http://my/newcorpus/uri 2003-01-02_69.txt 1
# text = Décisions du Conseil A l'unanimité, nos édiles donnent leur accord sur la modification des statuts de la communauté de communes concernant :
l'amélioration du traitement des déchets pour les points-tris, la déchetterie, voire la collecte sélective.
...
9   nos   son   DET   _   Gender=Masc|Number=Plur|Poss=Yes|PronType=Prs 10   det   _   _   _
10  édiles édile NOUN _   Gender=Masc|Number=Plur 11   nsubj _   _   _
11  donnent donner VERB _   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin 0   root   _   _   _
12  leur   son   DET   _   Gender=Masc|Number=Sing|Poss=Yes|PronType=Prs 13   det   _   _   _
13  accord accord NOUN _   Gender=Masc|Number=Sing 11   obj   _   _   _
14  sur    sur    ADP   _   16   case   _   _   _
15  la     le     DET   _   Definite=Def|Gender=Fem|Number=Sing|PronType=Art 16   det   _   _   _
16  modification modification NOUN _   Gender=Fem|Number=Sing 11   obl   _   _   _
...
```

cupt

```
# source_sent_id = http://my/newcorpus/uri 2003-01-02_69.txt 1
# text = Décisions du Conseil A l'unanimité, nos édiles donnent leur accord sur la modification des statuts de la communauté de communes concernant : -
l'amélioration du traitement des déchets pour les points-tris, la déchetterie, voire la collecte sélective.
...
9   nos   son   DET   _   Gender=Masc|Number=Plur|Poss=Yes|PronType=Prs 10   det   _   _   *
10  édiles édile NOUN _   Gender=Masc|Number=Plur 11   nsubj _   _   *
11  donnent donner VERB _   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin 0   root   _   _   1:LVC.full
12  leur   son   DET   _   Gender=Masc|Number=Sing|Poss=Yes|PronType=Prs 13   det   _   _   *
13  accord accord NOUN _   Gender=Masc|Number=Sing 11   obj   _   _   1
14  sur    sur    ADP   _   16   case   _   _   *
15  la     le     DET   _   Definite=Def|Gender=Fem|Number=Sing|PronType=Art 16   det   _   _   *
16  modification modification NOUN _   Gender=Fem|Number=Sing 11   obl   _   _   *
...
```



# Sortie OFCORS

```
{
  "0": "D\u00e9cisions", "1": "du", "2": "Conseil", "3": "A", "4":
  "\u0020", "5": "unanimite\u00e9", "6": ",", "7": "nos", "8": "\u00e9diles",
  "9": "donnent", "10": "leur", "11": "accord", "12": "sur", "13": "la",
  "14": "modification", "15": "des", "16": "statuts", "17": "de", "18":
  "la", "19": "communaut\u00e9", "20": "de", "21": "communes", "22":
  "concernant", "23": ":", "24": "-", "25": "\u0020", "26":
  "am\u00e9lioration", "27": "du", "28": "traitement", "29": "des",
  "30": "d\u00e9chets", "31": "pour", "32": "les", "33": "points-tris",
  "34": ",", "35": "la", "36": "d\u00e9chetterie", "37": ",", "38":
  "voire", "39": "la", "40": "collecte", "41": "s\u00e9lective", "42": ".",
  "43": "-", "44": "Projet", "45": "de", "46": "r\u00e9habilitation",
  "47": "de", "48": "toutes", "49": "les", "50": "d\u00e9charges", "51":
  ";", "52": "\u00e9tablissement", "53": "d'", "54": "une", "55":
  "charte", "56": "foresti\u00e8re", "57": "avec", "58": "\u0020", "59":
  "ONF", "60": "et", "61": "les", "62": "propri\u00e9taires", "63":
  "priv\u00e9s", "64": ".", "65": "-", "66": "R\u00e9alisation", "67":
  "d'", "68": "une", "69": "aire", "70": "d'", "71": "accueil", "72":
  "des", "73": "gens", "74": "du", "75": "voyage", "76": ".", "77": "-",
  "78": "Prise", "79": "en", "80": "charge", "81": "des", "82":
  "travaux", "83": "d'", "84": "entretien", "85": "et", "86": "d'",

```

liste des tokens

```
{
  "1":
  {
    "CONTENT": ["D\u00e9cisions", "du", "Conseil"],
    "LEFT_CONTEXT": ["<start>"],
    "RIGHT_CONTEXT": ["A", "\u0020", "unanimite\u00e9", ",", "nos", "\u00e9diles", "donnent", "leur", "accord", "sur"],
    "START": "0",
    "END": "2",
    "SPAN_ID": "0-2"
  },
  "2":
  {
    "CONTENT": ["D\u00e9cisions", "du", "Conseil", "A", "\u0020", "unanimite\u00e9", ",", "nos", "\u00e9diles"],
    "LEFT_CONTEXT": ["<start>"],
    "RIGHT_CONTEXT": ["donnent", "leur", "accord", "sur", "la", "modification", "des", "statuts", "de", "la"],
    "START": "0",
    "END": "8",
    "SPAN_ID": "0-8"
  },

```

liste des mentions

```
{
  "type": "clusters",
  "clusters": {
    "0": ["193", "195", "194", "210", "20"],
    "1": ["97", "93", "96"],
    "2": ["227", "224"],
    "3": ["161", "160", "162", "163"],
    "4": ["159", "154"],
    "5": ["187", "192"],
    "6": ["104", "111", "105", "103", "10"],
    "7": ["171", "172", "169", "176", "17"],
    "8": ["50", "54"],
    "9": ["78", "76"],
    "10": ["220", "218", "222"],
    "11": ["237", "235", "232"],

```

liste des cha\u00eenes

# Chaîne de traitements

1

texte brut → cupt blind

- Annotation avec udpipe
- conllu → cupt.blind

3

texte brut → fichiers  
de coréférence

- Annotation avec OFCORS
- fichiers sortis:  
tokens.json,  
mentions\_output.json,  
resulting\_chains.json

2

cupt blind → cupt

- Annotation avec  
Seen2Seen

4

Fusion des résultats

- Alignement des tokens

udpipe	spacy
[des, de, les]	[des]
[week-end]	[week, -, end]

- Création de cupt+