Extraction des interrogatives de corpus francophones annotés en dépendances universelles

Valentin D. Richard

LORIA, Université de Lorraine
ILLC, University of Amsterdam

27 mars 2023

Mon contexte

Mon projet de doctorat :

Aspects compositionnel et dynamique de la sémantique inquisitrice

.

Mon contexte

Mon projet de doctorat :

Aspects compositionnel et dynamique de la sémantique inquisitrice

Requiert :

Analyse des structures des interrogatives

(1)
a. Où va Marie ?
b. Où Marie va-t-elle ?
c. Où est-ce que Marie va ?
d. !Marie va où ?
e. %Où que Marie va ?
f. !C'est où que Marie va ?
g. !Où c'est que Marie va ?

.

Mon contexte

Mon projet de doctorat :

Aspects compositionnel et dynamique de la sémantique inquisitrice

Requiert :

Analyse des structures des interrogatives

(1)
a. Où va Marie ?
b. Où Marie va-t-elle ?
c. Où est-ce que Marie va ?
d. !Marie va où ?
e. %Où que Marie va ?
f. !C'est où que Marie va ?
g. !Où c'est que Marie va ?

Motivation :

- Grosse littérature sur les interrogatives en français*
- Mais désir de travailler sur des **données attestées**

. * ex. (LARRIVÉE et GURYEV 2021), GGF (DELAVEAU, CAPPEAU et DAGNAC 2021)

Choix des corpus

D'une part :

- Mais peu sur certaines structures
  - ex. subordonnées interrogatives non-standards :

(2) avant ça je me posais jamais la question de est-ce que j'aime faire ça (Omar Sy)

- Beaucoup d'études empiriques avec annotation manuelles **à partir de texte brut**

.

Choix des corpus

D'une part :

- Mais peu sur certaines structures
  - ex. subordonnées interrogatives non-standards :

(2) avant ça je me posais jamais la question de est-ce que j'aime faire ça (Omar Sy)

- Beaucoup d'études empiriques avec annotation manuelles **à partir de texte brut**

D'autre part :

- Corpus structurés en Dépendances Universelles (UD)
  - maintenu par Bruno Guillaume et Guy Perrier
- French Question Bank *

→ **Pas de trait pour le type de proposition** (interrogative / déclarative / ...)

. * (SEDDAH et CANDITO 2017)

Choix des corpus

D'une part :

- Mais peu sur certaines structures
  - ex. subordonnées interrogatives non-standards :

(2) avant ça je me posais jamais la question de est-ce que j'aime faire ça (Omar Sy)

- Beaucoup d'études empiriques avec annotation manuelles **à partir de texte brut**

D'autre part :

- Corpus structurés en Dépendances Universelles (UD)
  - maintenu par Bruno Guillaume et Guy Perrier
- French Question Bank *

→ **Pas de trait pour le type de proposition** (interrogative / déclarative / ...)

⇒ **Programme d'identification automatique des interrogatives**
à partir de corpus UD : **FUDIA**

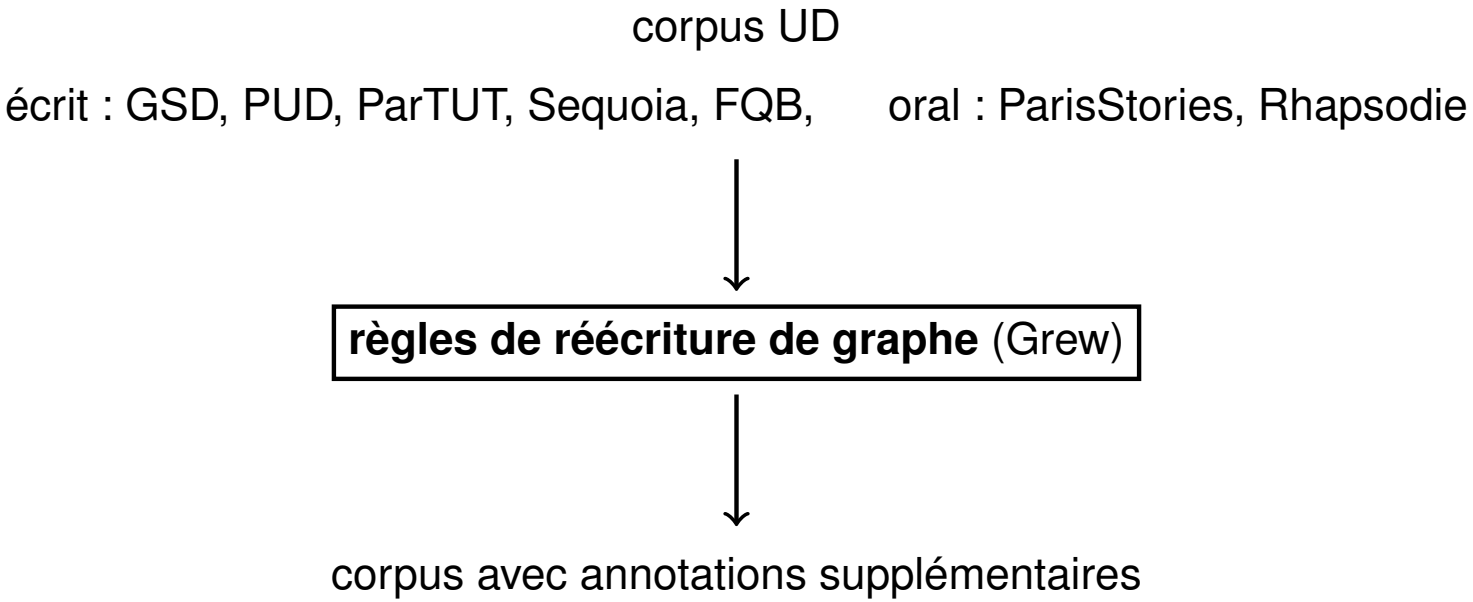
. * (Seddah et Candito 2017)

Méthode

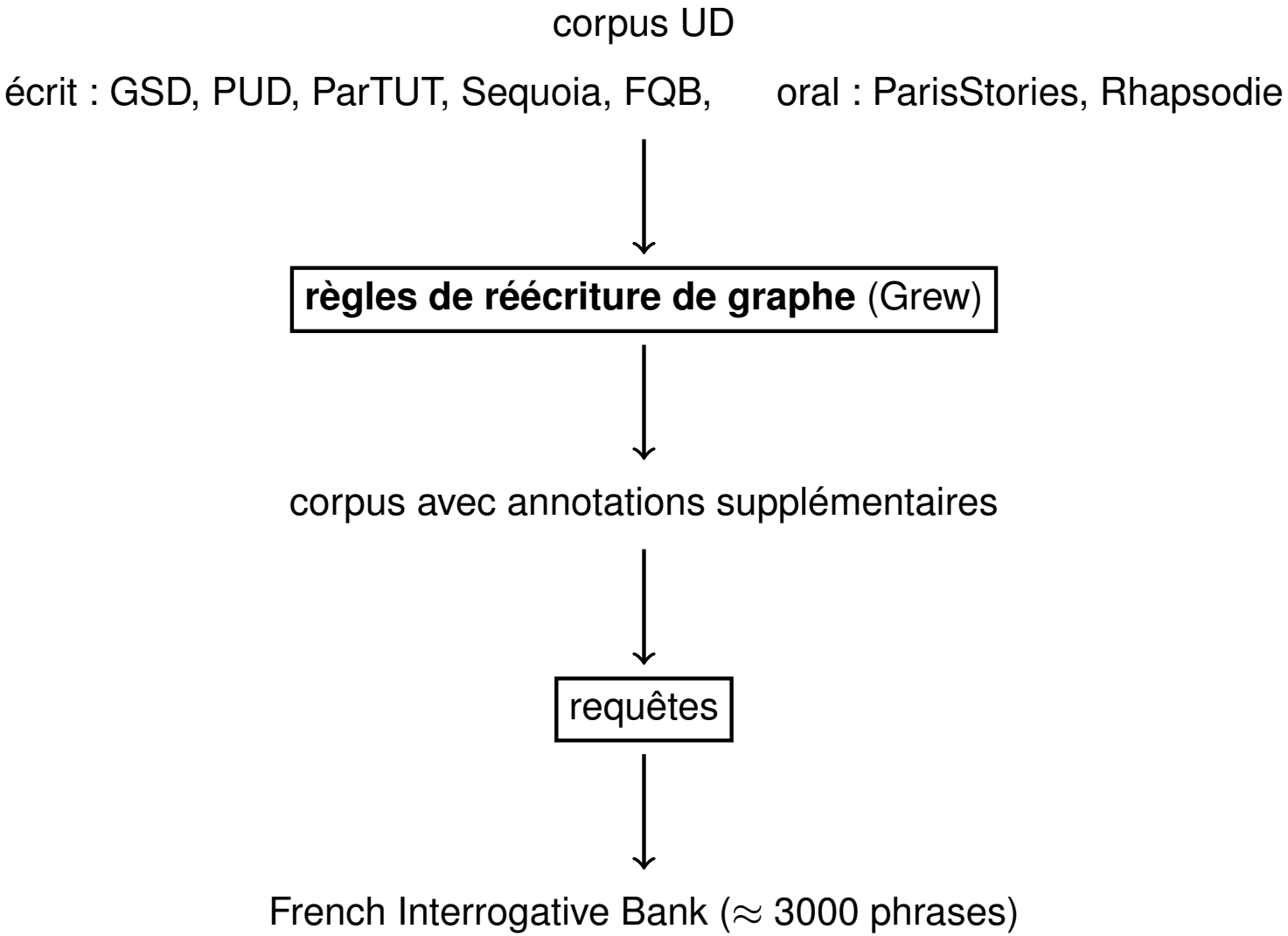
corpus UD

écrit : GSD, PUD, ParTUT, Sequoia, FQB, oral : ParisStories, Rhapsodie

Méthode


corpus UD
écrit : GSD, PUD, ParTUT, Sequoia, FQB,
oral : ParisStories, Rhapsodie
règles de réécriture de graphe (Grew)
corpus avec annotations supplémentaires

Méthode


corpus UD
écrit : GSD, PUD, ParTUT, Sequoia, FQB,
oral : ParisStories, Rhapsodie
règles de réécriture de graphe (Grew)
corpus avec annotations supplémentaires
requêtes
French Interrogative Bank (≈ 3000 phrases)

Conception des règles

Variables utilisées :

- Relations de dépendances
- traits morphosyntaxiques, e.g. `PronType=Int`
- ordre des mots, ex. sujet inversé

Conception des règles

Variables utilisées :

- Relations de dépendances
- traits morphosyntaxiques, e.g. `PronType=Int`
- ordre des mots, ex. sujet inversé
- pas basé sur les points d'interrogations
- lexique des classes **fermées** seulement : mots interrogatif
  - but : "découvrir" les prédicats acceptant les interrogatives

Conception des règles

Variables utilisées :

- Relations de dépendances
- traits morphosyntaxiques, e.g. `PronType=Int`
- ordre des mots, ex. sujet inversé
- pas basé sur les points d'interrogations
- lexique des classes **fermées** seulement : mots interrogatif
  - but : "découvrir" les prédicats acceptant les interrogatives

Élaboration des règles :

- grandes lignes à partir de la littérature
- en regardant le corpus
  - annotations UD utilisées en pratique
  - cas limites
- établissement d'heuristiques pour identifier certaines motifs
  ex."si" interrogatif vs. conditionnel

Annotations ajoutées

À	quelle	heure	vient	-elle	?

.

Annotations ajoutées

À	quelle	heure	vient	-elle	?
	`PronType=Int`				

- Mot interrogatif

.

Annotations ajoutées

À	quelle	heure	vient	-elle	?
	`PronType=Int`		`ClauseType=Int`		

- Mot interrogatif
- Tête de la proposition interrogative

.

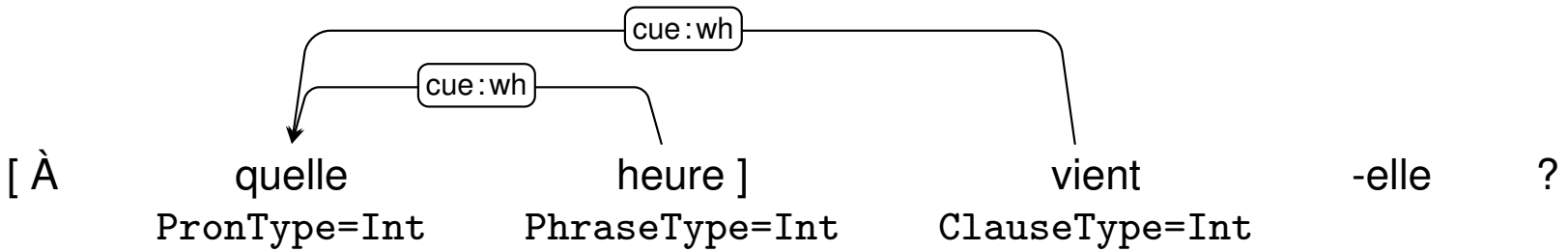
Annotations ajoutées

[À	quelle	heure]	vient	-elle	?
	`PronType=Int`	`PhraseType=Int`	`ClauseType=Int`		

- Mot interrogatif
- Tête de la proposition interrogative
- Tête du syntagme interrogatif

.

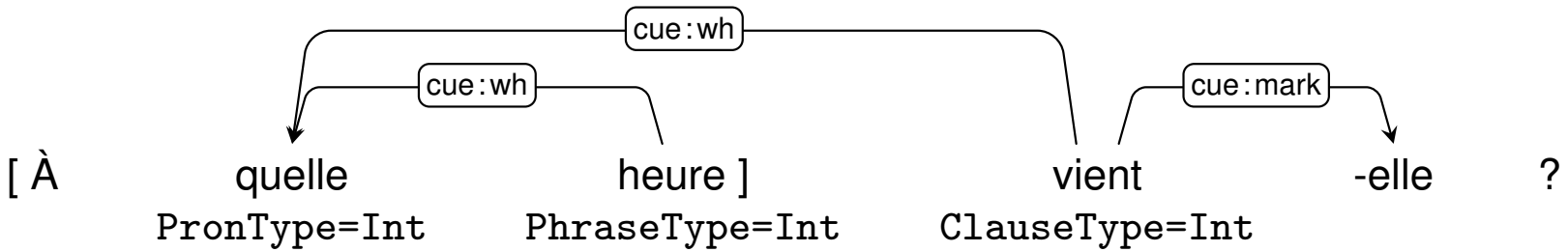
Annotations ajoutées



- Mot interrogatif
- Tête de la proposition interrogative
- Tête du syntagme interrogatif
- Arc(s) vers le mot interrogatif*

. * Seulement dans le FIB enrichi

Annotations ajoutées

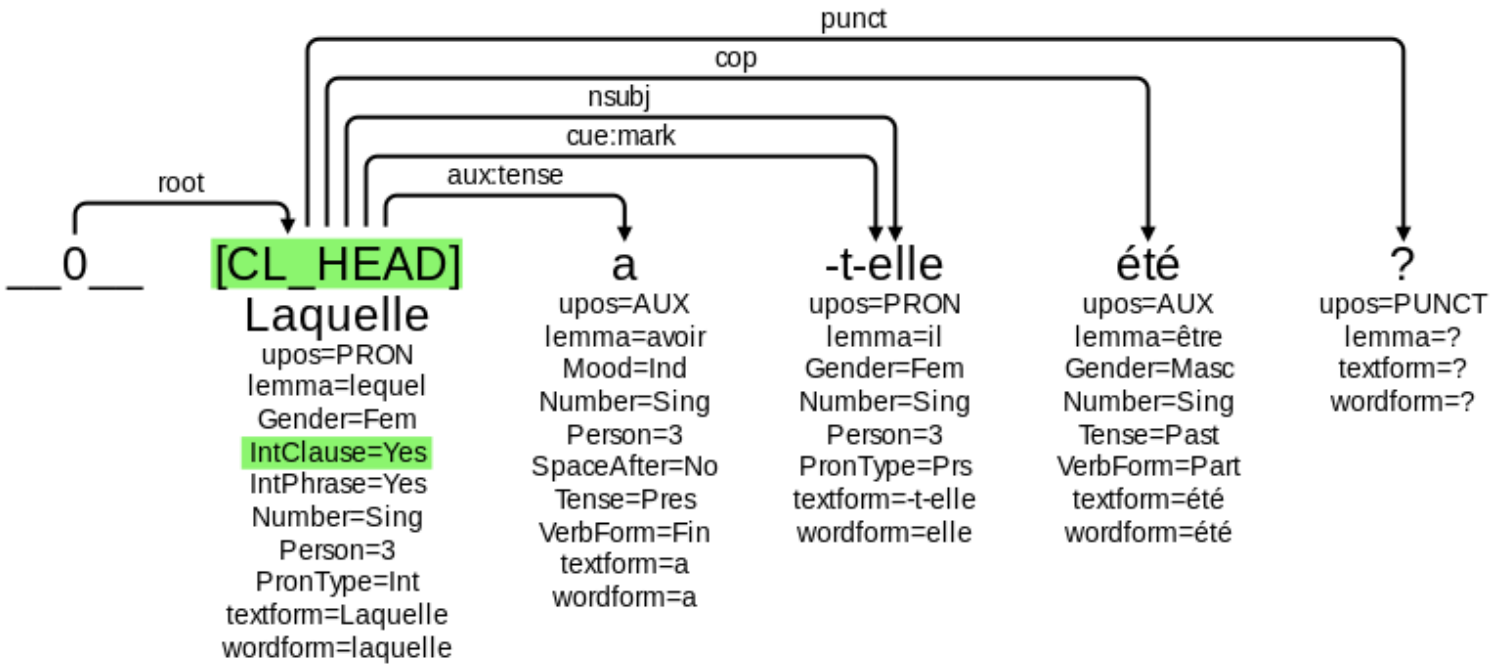


- Mot interrogatif
- Tête de la proposition interrogative
- Tête du syntagme interrogatif
- Arc(s) vers le mot interrogatif*
- Arc vers le marquage morphosyntaxique de l'interrogative*

. * Seulement dans le FIB enrichi

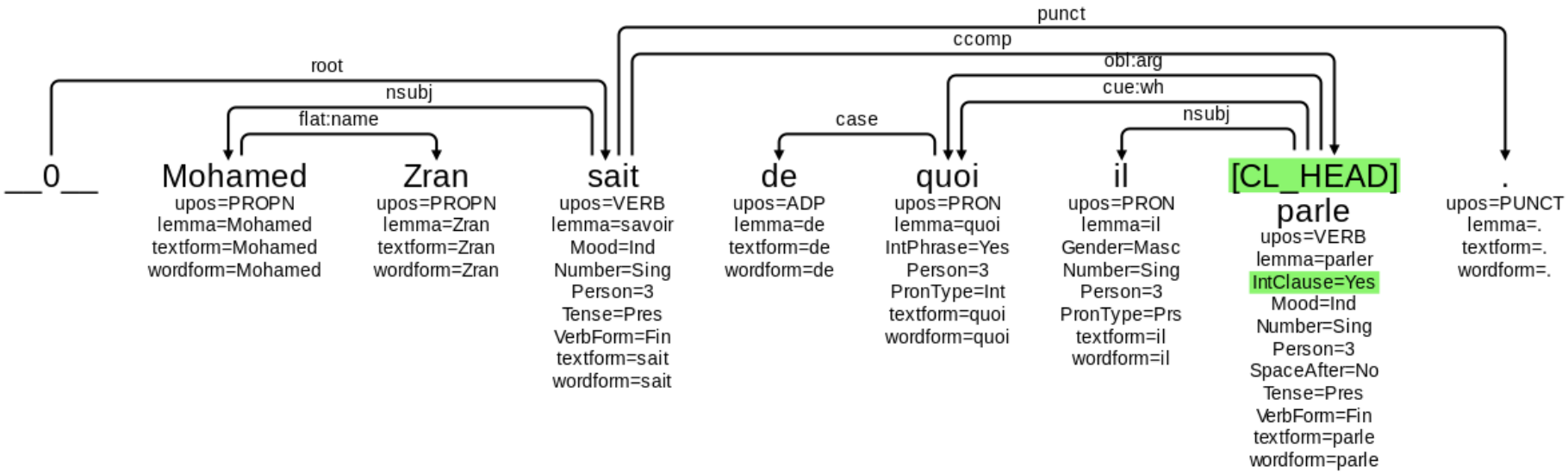
Exemple 1

- `WH = PH_HEAD = CL_HEAD` (GSD)



Exemple 2

- Interrogative enchâssée (GSD)



Évaluation

Estimation des **erreurs d'annotation** (dont faux positifs) :

- extraction des phrases avec `ClauseType=Int` sur une partie du corpus
- étiquetage à la main du type d'erreur

Évaluation

Estimation des **erreurs d'annotation** (dont faux positifs) :

- extraction des phrases avec `ClauseType=Int` sur une partie du corpus
- étiquetage à la main du type d'erreur

Estimation du nombre de **faux négatifs** :

1. extraction des phrase sans `ClauseType=Int` mais `PronType=Int`
2. extraction du reste des phrases du FQB

Résultat de l'évaluation

Catégorie \ Ensemble	`ClauseType=Int`	Sans `ClauseType=Int` mais `PronType=Int`	Total
0. Bien annoté	490	0	490
1. Faute d'origine	15	6	21
2. Faute de FUDIA			
- corrigée plus tard	26	55	81
- pas corrigée	2	2	4
3. Autre	0	7	7
Total	533	70	603

Table – Nombre de phrase étiquetées par catégorie et par ensemble.

Résultat de l'évaluation

Catégorie \ Ensemble	`ClauseType=Int`	Sans `ClauseType=Int` mais `PronType=Int`	Total
0. Bien annoté	490	0	490
1. Faute d'origine	15	6	21
2. Faute de FUDIA			
- corrigée plus tard	26	55	81
- pas corrigée	2	2	4
3. Autre	0	7	7
Total	533	70	603

Table – Nombre de phrase étiquetées par catégorie et par ensemble.

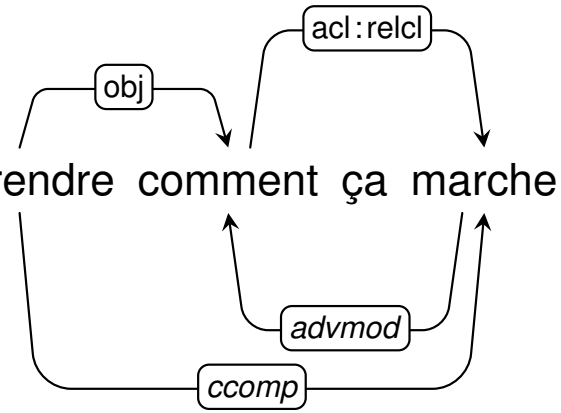
Sur le reste de FQB : 60 phrase

- 57 déclaratives
- 2 déclaratives questionnantes
- 1 faute de FUDIA
  - lemmatisation `t-il` du token `t-il`

Annotation issues

- Mauvaises annotation d'origine :

Rhapsodie :

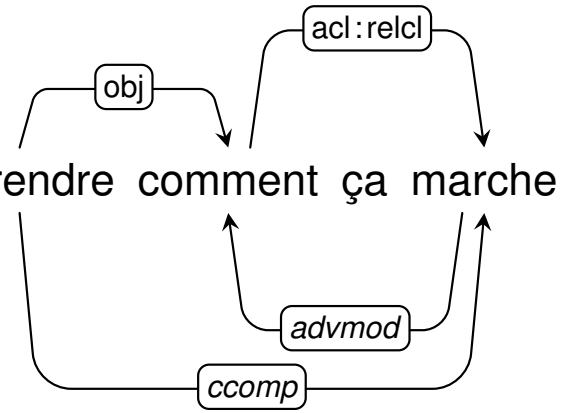


Attendu :

Annotation issues

- Mauvaises annotation d'origine :

Rhapsodie :



Attendu :

- Traits manquants, e.g. `PronType=Int`

Trous dans les guides

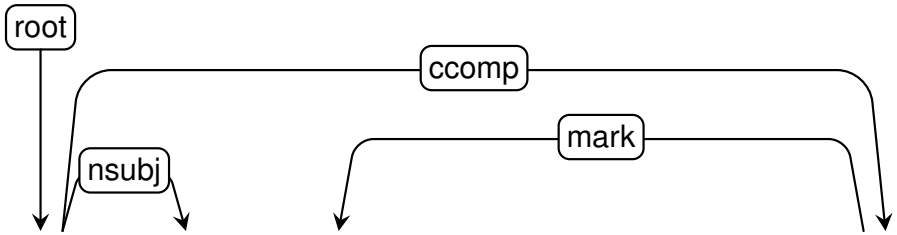
Absence de certains phénomènes dans les corpus UD

- e.g. particule québécoise *-tu*

(3) [...] bon, on va-tu prendre un café ? (CFPQ)

Reannotating fixed expressions

6 annotations différentes de "est-ce que" observées, e.g.

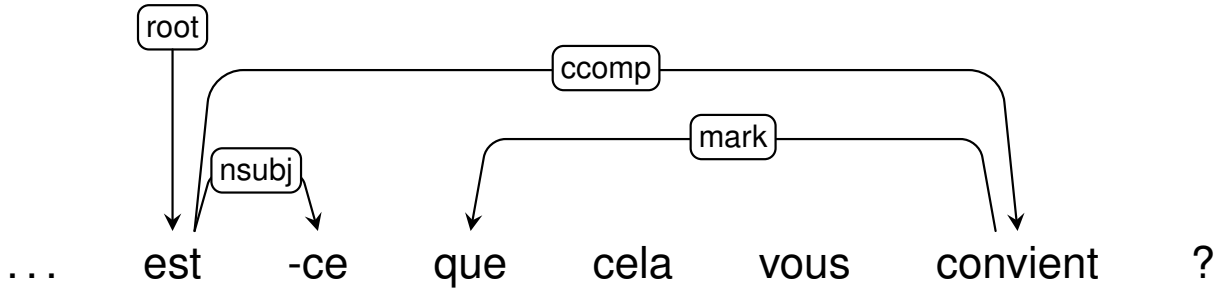


(ParTUT)

.

Reannotating fixed expressions

6 annotations différentes de "est-ce que" observées, e.g.



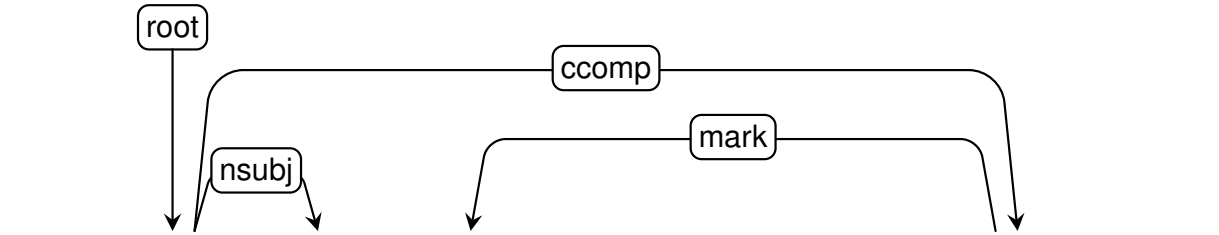
(ParTUT)

- Problème : **expression grammaticalisée** et lexicalisée*

. *(DRUETTA 2003) et GGF

Reannotating fixed expressions

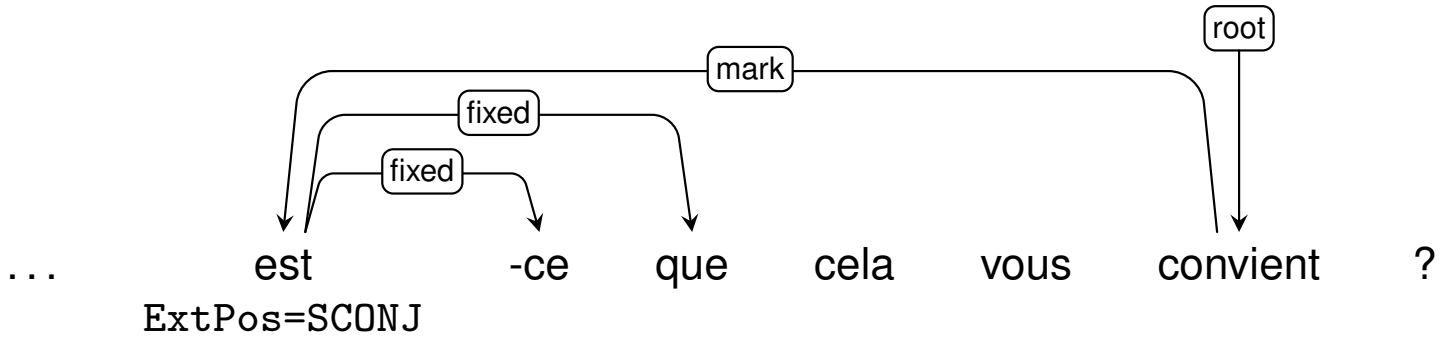
6 annotations différentes de "est-ce que" observées, e.g.



(ParTUT)

- Problème : **expression grammaticalisée** et lexicalisée*

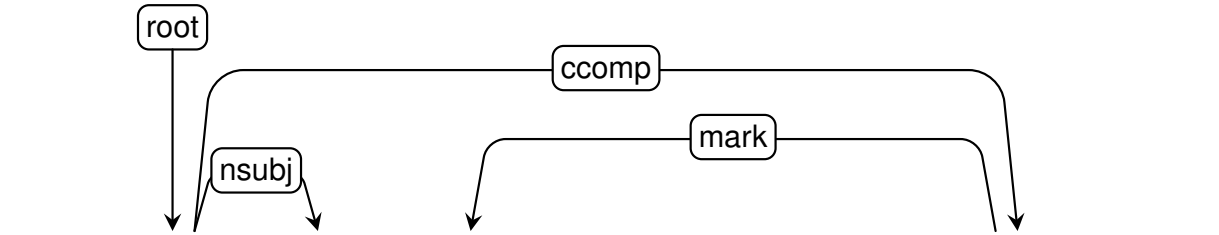
⇒ Réannotation comme marqueur figé



. *(DRUETTA 2003) et GGF

Reannotating fixed expressions

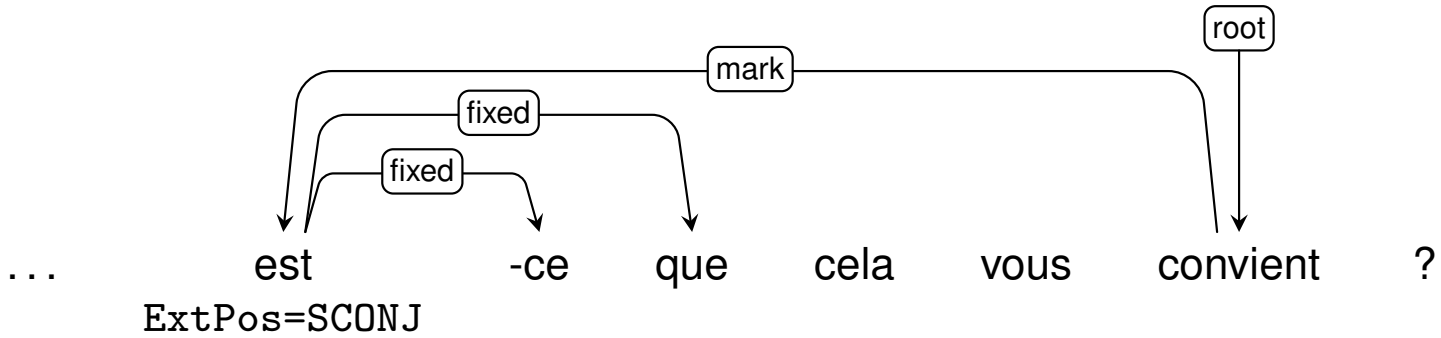
6 annotations différentes de "est-ce que" observées, e.g.



(ParTUT)

- Problème : **expression grammaticalisée** et lexicalisée*

⇒ Réannotation comme marqueur figé

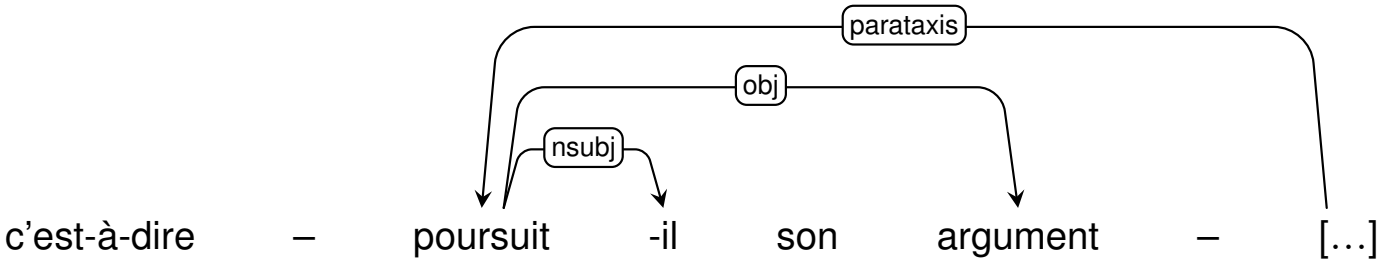


→ Pareil pour "qu'est-ce que" comme pronom

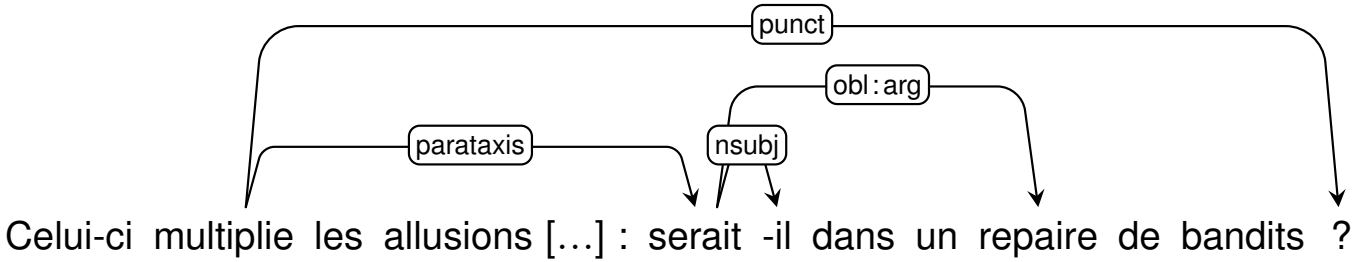
. *(DRUETTA 2003) et GGF

Inversion interrogative vs. stylistique

Parataxes avec inversion sujet-verbe :



(GSD)



(GSD)

- pas d'information suffisante en UD 2.11

Syntagme interrogative ?

Qu'est-ce qu'un **syntagme interrogative** quand WH est in situ ?

(4) Winnie est une imitation de quel animal ? (FQB)

Syntagme interrogative ?

Qu'est-ce qu'un **syntagme interrogative** quand WH est in situ ?

(4) Winnie est une imitation de quel animal ? (FQB)

- Test d'antéposition

(5) a. De quel animal (est-ce que) Winnie est une imitation ?
b. *Une imitation de quel animal (est-ce que) Winnie est ?

Syntagme interrogative ?

Qu'est-ce qu'un **syntagme interrogative** quand WH est in situ ?

(4) Winnie est une imitation de quel animal ? (FQB)

- Test d'antéposition

(5) a. De quel animal (est-ce que) Winnie est une imitation ?
b. *Une imitation de quel animal (est-ce que) Winnie est ?

- pas toujours concluant

(6) a. Elle est [en quête [de quel dragon]] ?
b. [De quel dragon] est-elle [en quête _] ?
c. [En quête [de quel dragon]] est-elle ?

Conclusion

FUDIA

- programme par règles
- identification de toutes les interrogatives

French Interrogative Bank

- corpus d'interrogatives annoté en UD
- version enrichie pour faire des statistiques

Projet :

- statistiques fines
- les comparer avec d'autres études


DELAVEAU, Annie, Paul CAPPEAU et Anne DAGNAC (2021). “Les phrases interrogatives”. In : **La Grande Grammaire du Français**. Sous la dir. d’Anne ABEILLÉ et Danièle GODARD. 1re éd. T. 2. Arles : Actes Sud/Imprimeries nationales Éditions, p. 1402-1437. ISBN : 978-2-330-14239-1.


DRUETTA, Ruggero (24 nov. 2003). “"Qu’est-Ce Tu Fais ?" État d’avancement de La Grammaticalisation de "Est-Ce Que". Première Partie”. In : **Linguæ & - Rivista di lingue e culture moderne** 1.2 (2), p. 67-88. ISSN : 1724-8698. URL : https://www.ledonline.it/index.php/linguae/article/view/154 (visité le 07/07/2022).


LARRIVÉE, Pierre et Alexander GURYEV (24 déc. 2021). “Variantes formelles de l’interrogation. Présentation”. In : **Langue francaise** 212.4, p. 9-24. ISSN : 0023-8368. URL : https://www.cairn.info/revue-langue-francaise-2021-4-page-9.htm?ref=doi (visité le 15/02/2022).


SEDDAH, Djamé et Marie CANDITO (2017). “Tour d’Horizon du French QuestionBank : Construire un Corpus Arboré de Questions pour le Français”. In : **ACor4French - Les corpus annotés du français**. Orléans, France : Association pour le Traitement Automatique des Langues.