



Universal Dependencies et GREW-match

Karën Fort

karen.fort@sorbonne-universite.fr / <https://members.loria.fr/KFort/>



Quelques sources d'inspiration

- ▶ Bruno Guillaume et Guy Perrier, LORIA
- ▶ Tutoriel : <http://match.grew.fr/?tutorial=yes>
- ▶ Tutoriel TALN de Grew-match, avec l'accord de B. Guillaume
- ▶ *Conversion et améliorations de corpus du français annotés en Universal Dependencies*, B. Guillaume (séminaire au Lattice, 14 janvier 2020)
- ▶ *Application de la réécriture de graphes au traitement automatique des langues* [Bonfante et al., 2018]

Sources

Universal Dependencies

Présentation

Annotations

GREW-match

Pour finir



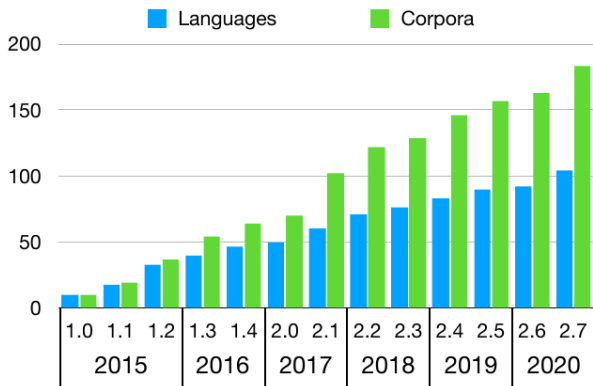
Le projet Universal Dependencies (UD) :

- ▶ Depuis 2014, 2 versions par an
- ▶ Projet collaboratif de production de corpus annotés en dépendances
- ▶ <http://universaldependencies.org>



Des corpus très variables

en taille, en qualité, en genre








Skolt Sami :

- ▶ 200 phrases
- ▶ 350 locuteurs en 2016

UD German-HDT :

- ▶ 189 928 phrases

Corpus disponibles pour le français

French		8	1,157K	     	IE, Romance
French treebanks					
▶	ParTUT	28K	(L)F	   	    
▶	GSD	400K	(L)F	   	    
▶	Sequoia	70K	(L)F	   	 
▶	FQB	24K	(L)F	 	 
▶	Spoken	35K	(L)F		    
▶	PUD	24K	F	 	    
▶	FTB	573K	(L)F		 
▶	CrapBank	-		 	    

<https://universaldependencies.org>

Sources

Universal Dependencies

Présentation

Annotations

GREW-match

Pour finir

Un schéma d'annotation unique

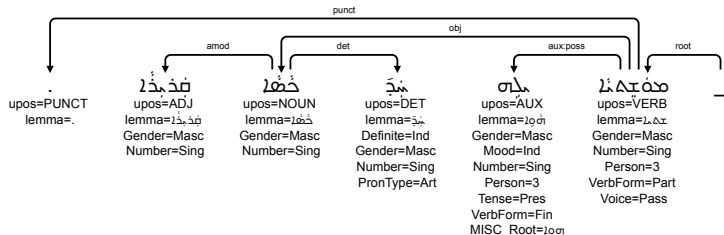
obl POS : 17 tags

obl Dépendances syntaxiques : 37 étiquettes universelles
(sous-typage possible)

fac Lemmatisation

fac Morphologie : 23 traits disponibles

Têtes sur les mots lexicaux (noms, verbes, adj et certains adv)

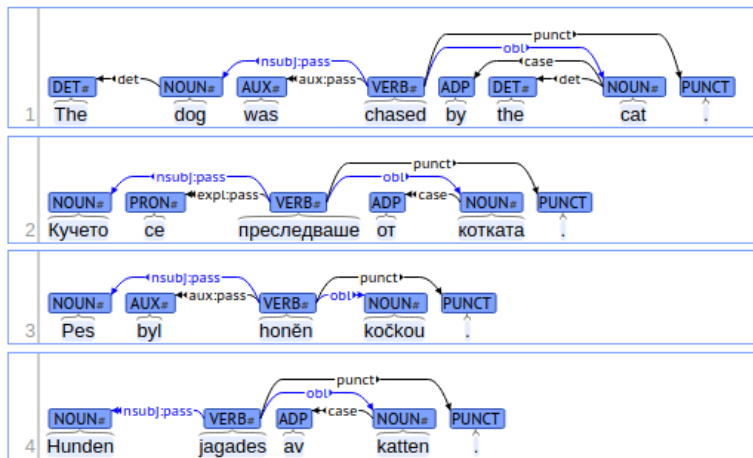


Les couches UD

nsubj		Dependency relation
She	laughed	Surface form (tokenised)
she	laugh	Lemma
PRON	VERB	POS
Case=Nom Gender=Fem Number=Sing Person=3 PronType=Prs	Mood=Ind Tense=Past VerbForm=Fin	Morphology

Un meilleur parallélisme entre les différentes langues

exemple en anglais, bulgare, tchèque, suédois



<https://universaldependencies.org/introduction.html>

Sources

Universal Dependencies

GREW-match

- Présentation

- Rechercher des nœuds

- Rechercher des relations

- Précédence, filtrage, tri

- Pour les curieux

Pour finir

GREW [Perrier and Guillaume, 2019]

Outil de TAL

- ▶ fondé sur la réécriture de graphes
- ▶ permet de :
 - ▶ rechercher un motif (dans une annotation)
 - ▶ transformer une annotation
- ▶ code librement disponible sous licence CeCILL V2.1

GREW vs GREW-match

grew

- ▶ <https://grew.fr/>
- ▶ ré-écriture de graphe : permet de transformer une annotation
- ▶ règle : (*motif*, *commande*)

grew-match

- ▶ <https://match.grew.fr/>
- ▶ permet de rechercher un motif dans une annotation (d'un corpus)
- ▶ règle : (*motif*)

Interface

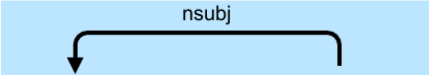
<http://match.grew.fr/>

The screenshot shows the Grew-match web interface. At the top, there is a navigation bar with tabs for 'Grew-match', 'Tutorial', 'UD 2.5', 'SUD 2.4', 'Misc UD/SUD', 'Sequoia', 'AMR', 'PARSEME', 'Orfeo', 'Help', 'Report', and 'Grew website'. On the left side, there is a sidebar menu with a tree view showing 'UD_Faroese-OfI@2.5', 'Finnish', and 'French'. Under 'French', several corpora are listed: 'UD_French-FTB@2.5' (with a warning icon), 'UD_French-FQB@2.5', 'UD_French-GSD@2.5', 'UD_French-PUD@2.5', 'UD_French-ParTUT@2.5', 'UD_French-Sequoia@2.5', and 'UD_French-Spoken@2.5'. The main content area is titled 'UD_English-LinES@2.5 [5243 trees, 94217 tokens]' and contains a 'Hide corpora list' button and a 'Relation tables' button. On the right side, there are search options: 'Basic' (selected), 'n-grams', 'Clusters', and 'Misc'. Below these are search criteria: 'Search for a form', 'Search for a lemma (does not exist in all languages)', 'Search for a POS (upos)', 'Search for a dependency relation', 'Search for both relations and tags', and 'Filter with NAP (Negative Application Patterns)'. Each search criterion has a corresponding input field.

Documentation :

- ▶ exemples fournis (cliquer à droite)
- ▶ <http://grew.fr/pattern/> (plus complet)

Grew-match sur UD

UD annotation	Grew-match syntax
	<code>pattern { M -[nsubj]-> N }</code>
She	<code>pattern { N [form="laughed"] }</code>
she	<code>pattern { N [lemma="laugh"] }</code>
PRON	<code>pattern { N [upos=VERB] }</code>
Case=Nom	
Gender=Fem	
Number=Sing	<code>pattern { N [Tense=Past] }</code>
Person=3	
PronType=Prs	
laughed	
laugh	
VERB	
Mood=Ind	
Tense=Past	
VerbForm=Fin	

Sources

Universal Dependencies

GREW-match

Présentation

Rechercher des nœuds

Rechercher des relations

Précédence, filtrage, tri

Pour les curieux

Pour finir

Recherches simples

En français (pour commencer), dans UD_French-GSD@2.5

/!\ forme (*mangeait*) vs lemme (*manger*)

/!\ POS (AUX) vs dépendance (aux)

/!\ POS (ADJ) vs trait (Plur, Fem)

Chercher

- ▶ la forme *avaient*
- ▶ le lemme *avoir*
- ▶ le POS PROPN (nom propre)
- ▶ les adjectifs au féminin pluriel

Écrire un *pattern* GREW

```
pattern {  
  N [upos="NUM"]  
}
```

Distinguer les éléments de :

- ▶ la syntaxe GREW :

```
pattern {  
  []  
}
```

- ▶ le schéma d'annotation (ici, UD) :

- ▶ upos
- ▶ NUM

- ▶ **votre** nommage de nœuds :

- ▶ N

Se documenter

En français, dans UD_French-GSD@2.5

Truc

- ▶ pour visualiser toutes les entrées qui ont un trait (ou voir comment il s'écrit) :

```
pattern {  
  N [upos="ADJ", Number, Gender]  
}
```

- ▶ pour visualiser toutes les entrées qui n'ont pas un trait :

```
pattern {  
  N [upos="ADJ", !Number]  
}
```

Recherches simples 2

En russe, dans UD_Russian-GSD@2.5

Chercher

- ▶ tous les pronoms personnels я (je)

Sources

Universal Dependencies

GREW-match

Présentation

Rechercher des nœuds

Rechercher des relations

Précédence, filtrage, tri

Pour les curieux

Pour finir

Recherches de dépendances

En français (pour commencer), dans UD_French-GSD@2.5

/!\ la flèche part de la tête de la dépendance

/!\ catégorisation UD

Chercher

- ▶ toutes les relations sujet

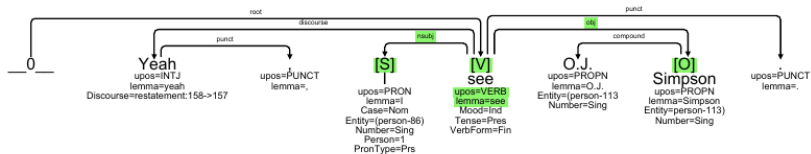
Suites

Chercher

Des exemples d'utilisation de deux auxiliaires en français.

Attention !

pattern { V [upos=VERB, lemma="see"]; V -[nsubj]-> S; V -[obj]-> O }



Solution

The screenshot shows the Grew web interface with a search query for a dependency relation pattern. The interface includes a top navigation bar with tabs for Grew-match, Tutorial, UD, SUD, Sequoia, AMR, PARSEME, Orfeo, Help, Report, and Grew website. A left sidebar lists various corpora, with UD_French-GSD@2.7 highlighted in green. The main content area displays the search results for the pattern `{ GOV - [aux|aux:pass|aux:tense]-> DEP ; GOV - [aux|aux:pass|aux:tense]-> DEP2 }` in the UD_French-GSD@2.7 corpus, which contains 16341 trees and 416740 tokens. The search results are filtered by lemma, upos, xpos, features, textform/wordform, and sentences order. The search results show that more than 1000 results were found in 69.64% of the corpus.

UD_French-GSD@2.7 [16341 trees, 416740 tokens]

```
1 % Search for a dependency relation
2 % Available relations are:
3 % acl, acl:relcl, advcl, advmod, amod, appos, aux, aux:pass,
4 % compound, conj, cop, csubj, dep, det, discourse, obj, expl,
5 % fixed, flat, neg, nmod, nmod:poss, nsubj, nsubj:pass, numm
6
7 pattern { GOV - [aux|aux:pass|aux:tense]-> DEP ;
8 GOV - [aux|aux:pass|aux:tense]-> DEP2
9 }
```

Relation tables

Basic n-grams

Clusters Misc

Search for a form
Search for a lemma (does not exist in all languages)
Search for a POS (upos)
Search for a dependency relation
Search for both relations and tags
Filter with NAP (Negative Application Patterns)

lemma upos xpos features textform/wordform sentences order:
initial context

Search Save Export Clustering

More than 1000 results found in 69.64% of the corpus [0.26s]

Questions de langues

En japonais, on peut utiliser plusieurs auxiliaires

Chercher

Combien d'auxiliaires peuvent être utilisés conjointement en japonais ?

Aller plus loin : Relation tables

UD_French-GSD@2.5

mark

nmod

nsubj

nsubj:caus

nsubj:pass

nummod

obj

obj:agent

DEP→ § GOV	TOTAL	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
TOTAL	20188	21	3	36	1		88		8021	33		8638	3251			36	5	55
ADJ	1367		1	1			4		938	1		354	64			4		
ADP	21						1		9			10	1					
ADV	34								15	1		15	3					
AUX	2											1	1					
CCONJ																		
DET	2											1	1					

Sources

Universal Dependencies

GREW-match

Présentation

Rechercher des nœuds

Rechercher des relations

Précédence, filtrage, tri

Pour les curieux

Pour finir

Notions d'ordre

En français, dans UD_French-GSD@2.7

/!\ << (avant) vs < (juste avant)

Chercher

- ▶ tous les cas où le sujet apparaît après le verbe :
 1. chercher toutes les relations sujet
 2. filtrer selon l'ordre des nœuds

Traduire ce qu'on cherche

En français (pour commencer), dans UD_French-GSD@2.7

/!\ tokénisation préalable

Chercher

▶ *Sahara occidentale*

Exclusions

En français, dans UD_French-GSD@2.7

/!\ without

Chercher

- ▶ tous les cas où le sujet apparaît avant le verbe
- ▶ mais pas juste avant

Trier les sorties

En français, dans UD_French-GSD@2.7

/!\ clustering key (en bas)

Chercher et trier

- ▶ tous les cas où le sujet apparaît après le verbe
- ▶ classés par lemme

Sources

Universal Dependencies

GREW-match

Présentation

Rechercher des nœuds

Rechercher des relations

Précédence, filtrage, tri

Pour les curieux

Pour finir

Filtrer dans le pattern vs à l'extérieur

```
pattern {  
  N[];  
  N.Number <> Sing  
}
```

VS

```
pattern {  
  N[]  
}  
without {  
  N.Number="Sing"  
}
```

Multiplication des petits pains

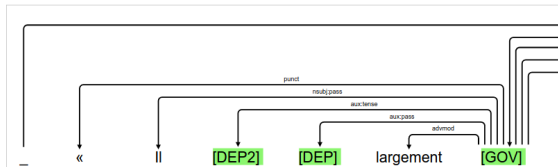
GREW-match trouve toutes les permutations

Get more results

1 / 10

fr-ud-test_00003 [1/2]
fr-ud-test_00003 [2/2]
fr-ud-test_00020 [1/2]
fr-ud-test_00020 [2/2]
fr-ud-test_00060 [1/2]
fr-ud-test_00060 [2/2]
fr-ud-test_00079 [1/2]
fr-ud-test_00079 [2/2]
fr-ud-test_00085 [1/2]
fr-ud-test_00085 [2/2]

« Il a été largement démontré que la population civile du territoire non autonome du Sahara occidental est l'objet de diverses atteintes aux droits humains, comme la détention arbitraire, les coups et les tortures », écrit l'ONG internationale, implantée dans 35 pays, citée par l'agence de presse sahraouie.

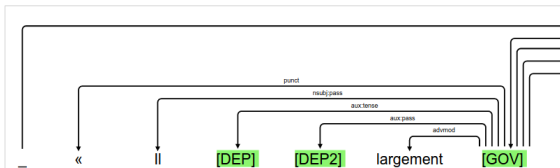


Get more results

2 / 10

fr-ud-test_00003 [1/2]
fr-ud-test_00003 [2/2]
fr-ud-test_00020 [1/2]
fr-ud-test_00020 [2/2]
fr-ud-test_00060 [1/2]
fr-ud-test_00060 [2/2]
fr-ud-test_00079 [1/2]
fr-ud-test_00079 [2/2]
fr-ud-test_00085 [1/2]
fr-ud-test_00085 [2/2]

« Il a été largement démontré que la population civile du territoire non autonome du Sahara occidental est l'objet de diverses atteintes aux droits humains, comme la détention arbitraire, les coups et les tortures », écrit l'ONG internationale, implantée dans 35 pays, citée par l'agence de presse sahraouie.



Sources

Universal Dependencies

GREW-match

Pour finir

CQFR : Ce Qu'il Faut Retenir

TD

Bibliographie



- ▶ Comment trouver de l'aide
- ▶ Chercher puis raffiner
- ▶ Accès à d'autres corpus qu'UD (Orféo, Sequoia, AMR)

TD à rendre, sera noté

- ▶ document pdf
- ▶ avec captures d'écran GREW
- ▶ envoyé par mail

GREW match sur le français (GSD)

http://match.grew.fr/?corpus=UD_French-GSD@2.7

/!\ <> (différent)

Accords

- ▶ chercher les désaccords en nombre du verbe avec le sujet
- ▶ enlever la copule des résultats obtenus ci-dessus
- ▶ chercher l'accord en nombre du verbe avec l'objet

GREW match 2 sur le français (GSD)

http://match.grew.fr/?corpus=UD_French-GSD@2.7

Dépendances longue distance

- ▶ chercher les relatives
- ▶ exclure les dépendances courtes : cas où le pronom relatif est directement dépendant de la tête de la relative (*le livre que j'ai lu*)

GREW match 3

Caractéristiques d'une langue inconnue

- ▶ est-ce qu'il existe des auxiliaires en hébreu ? des prépositions ? des déterminants ?
- ▶ est-ce une langue SOV ? SVO ? autre ?
- ▶ est-ce que les adjectifs sont plutôt avant les noms ou après ?
- ▶ Mêmes questions pour le norvégien



Bonfante, G., Guillaume, B., and Perrier, G. (2018).

Application de la réécriture de graphes au traitement automatique de
volume 1 of Série Logique, linguistique et informatique.

ISTE editions.



Perrier, G. and Guillaume, B. (2019).

GREW, a tool for annotating corpora and exploiting annotated corpora.

Journées scientifiques "Linguistique informatique, formelle de terrain".

Poster.