



Universal Dependencies et GREW-match

Karën Fort

karen.fort@univ-lorraine.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

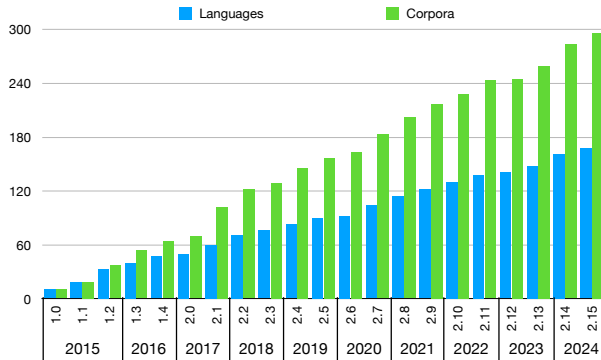
en taille, en qualité, en genre

Skolt Sami :

- ▶ 245 phrases
- ▶ 350 locuteurs en 2016

UD German-HDT :

- ▶ 189 928 phrases



Corpus disponibles pour le français

▼		French	7	635K	     	IE, Romance
French treebanks						
▶	GSD	400K	(L)(F)	    		★★★★☆
▶	Sequoia	70K	(L)(F)	    		★★★★☆
▶	ParTUT	28K	(L)(F)	    		★★★★☆
▶	ParisStories	42K	(L)(F)			★★★★☆
▶	Rhapsodie	44K	(L)(F)			★★★★☆
▶	PUD	24K	(L)(F)	 		★★★★☆
▶	FQB	23K	(L)(F)	 		★★★★☆

See [here](#) for comparative statistics of French treebanks.

<https://universaldependencies.org>

Sources

Universal Dependencies

Présentation

Annotations

GREW-match

Pour finir

Un schéma d'annotation unique

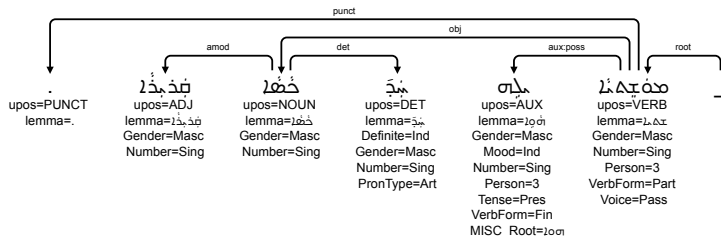
obl POS : 17 tags

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

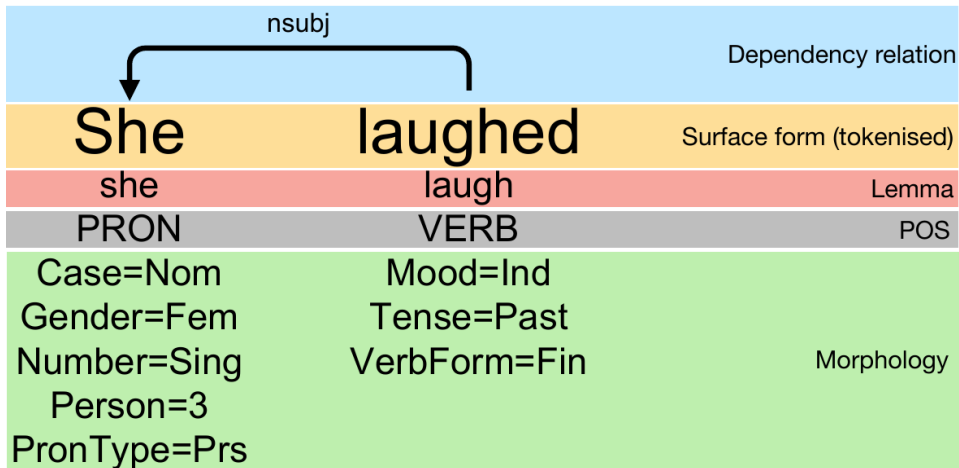
fac Lemmatisation

fac Morphologie : 23 traits disponibles

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

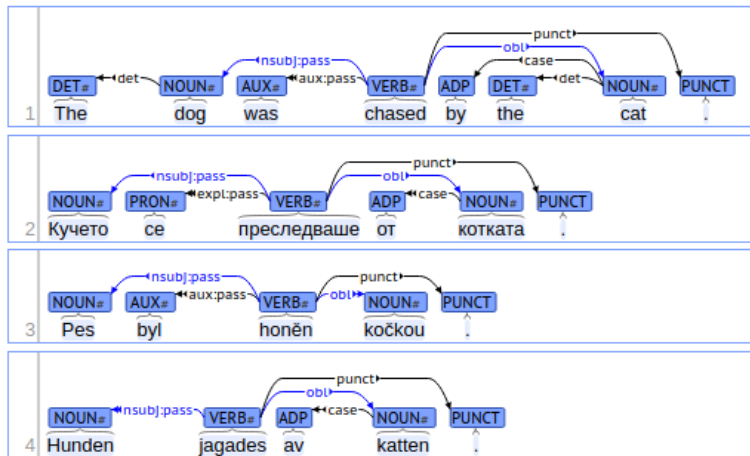


Les couches UD



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

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

Interface

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

The screenshot displays the Universal Grew web interface. At the top, a navigation bar includes links for 'Tutorial', 'UD 2.11' (the active tab), 'SUD 2.11', 'UD Latest', 'SUD Latest', 'UD Auto', and 'SUD Auto'. On the left, a sidebar titled '241 corpora' lists various datasets such as 'UD_Abaza-ATB@2.11', 'UD_Afrikaans-AfriBooms@2.11', and 'UD_Akkadian-PISANDUB@2.11'. The main area shows the selected corpus 'UD_English-GUM@2.11' with a search input field containing the number '1'. Below this, there are checkboxes for 'Clustering 1' with options 'No', 'Key', and 'Whether'. Further down, there are checkboxes for 'lemma', 'upos', 'xpos', 'features', and 'textform/wordform', along with a 'sentences order' dropdown set to 'initial' and a 'context' checkbox. On the right, there are tabs for 'Basic' and 'n-grams', and sub-tabs for 'Clustering' and 'Misc'. A section titled 'Request on graphs:' provides search filters for 'form', 'lemma', 'upos', and 'tags'. Another section titled 'Request on metadata:' provides search filters for 'sent_id' and 'full text (with regexp)'.

Documentation :

- ▶ exemples fournis (cliquer à droite)
- ▶ tutorial (en haut)

Grew-match sur UD

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

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@2.11

/!\ 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@2.11

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@2.11

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@2.11

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

/!\ catégorisation UD

Chercher

► toutes les relations sujet

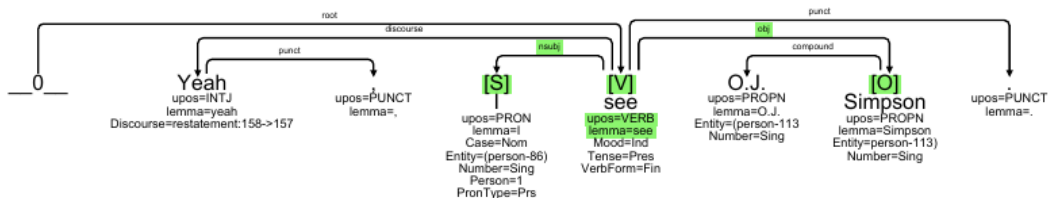
Suites

Chercher

Des exemples d'utilisation de deux auxiliaires en français.

Attention !

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



Solution

 Tutorial **UD 2.11** SUD 2.11 UD Latest SUD Latest UD Auto SUD Auto ?

241 corpora
Filter: ✕
UD_French-ParTUT@2.11
UD_French-GSD@2.11
UD_French-Sequoia@2.11
UD_French-FQB@2.11
UD_French-Rhapsodie@2.11
UD_French-ParisStories@2.11
UD_French-PUD@2.11
UD_French-FTB@2.11 
UD_Frisian_Dutch-Fame@2.11
UD_Galician-TreeGal@2.11
UD_Galician-CTG@2.11
UD_German-HDT@2.11

Hide corpora list UD_French-GSD@2.11 ⓘ ↺ 📄

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

Clustering 1: ☒ No ☐ Key ☐ Whether
☒ lemma ☒ upos ☐ xpos ☒ features ☐ textform/wordform ⓘ
sentences order: ☐ context

Search 🔍 Count ≡

Basic n-grams

Clustering Misc

Request on graphs:
Search for a
Search for a
Search for a
Search for a dependency relation
Search for both relations and tags
Filter with NAP (Negative Application Patterns)
Request on metadata:
Search for a
Search in full text (with regexp)

More than 1000 results found in 73.09% of the corpus [0.306s]

Save 📄 TSV 📄 CoNLL 📄

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@2.11

obj:agent [23]

mark [6725]

nmod [35439]

nsubj [18998]

nsubj:caus [132]

nsubj:pass [3664]

nummod [3582]

obj [12950]

obj:agent [112]

obj:lvc [554]

obl [19]

obl:agent [1554]

obl:arg [8325]

obl:mod [15606]

DEP→ ↓ GOV	TOTAL	PRON	NOUN	PROPN	ADJ	NUM	X	VERB	ADV	ADP	AUX	CCONJ	DET	INTJ	PUNCT	SCONJ	SYM	_
TOTAL	18998	8238	7463	3183	75	31	4	2	2									
VERB	14420	7036	5280	2020	59	19	3	1	2									
NOUN	2771	718	992	1044	9	7	1											
ADJ	1400	356	972	68	2	2												
PRON	192	69	83	37	3													
PROPN	99	16	76	4	2			1										
ADV	30	11	16	2		1												
SYM	27	2	18	5		2												
NUM	26	13	13															
ADP	21	10	10	1														
X	5	1	3	1														
DET	4	3		1														
AUX	3	3																
CCONJ																		

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@2.11

/!\ << (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@2.11

/!\ tokénisation préalable

Chercher

► *Sahara occidental*

Exclusions

En français, dans UD_French-GSD@2.11

/!\ 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@2.11

/!\ 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

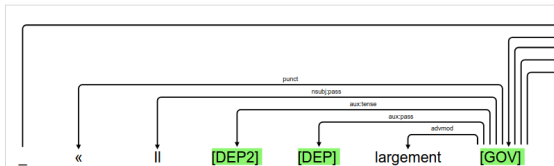
GREW-match trouve toutes les permutations

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.

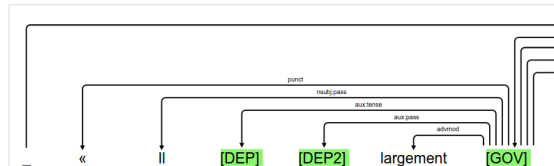


Get more results

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)

GREW match sur le français (GSD)

https://universal.grew.fr/?corpus=UD_French-GSD@2.13

/!\ <> (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)

https://universal.grew.fr/?corpus=UD_French-GSD@2.13

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 vietnamien ? 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 l'ukrainien



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

Application de la réécriture de graphes au traitement automatique des langues,
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.