



Universal Dependencies et GREW-match

Karën Fort

karen.fort@sorbonne-universite.fr / <http://karenfort.org>

28 février 2020



Quelques sources d'inspiration

- ▶ Bruno Guillaume et Guy Perrier, LORIA
- ▶ Tutoriel : <http://match.grew.fr/?tutorial=yes>
- ▶ *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



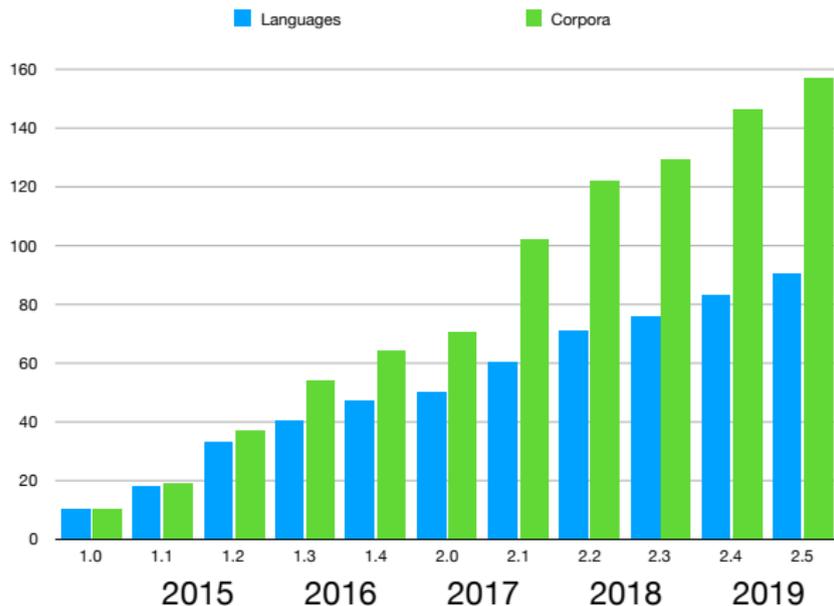
Le projet Universal Dependencies (UD) :

- ▶ 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 :

- ▶ 36 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	  W	 ★★★★★
▶	GSD	400K	(L)F	   W	 ★★★★★
▶	Sequoia	70K	(L)F	   W	 ★★★★★
▶	FQB	24K	(L)F	 	 ★★★★★
▶	Spoken	35K	(L)F		 ★★★★★
▶	PUD	24K	F	 W	 ★★★★★
▶	FTB	573K	(L)F		 ★★★★★
▶	CrapBank	-		 	 ★★★★★

<https://universaldependencies.org>

Sources

Universal Dependencies

Présentation

Annotations

GREW-match

Pour finir

Un schéma d'annotation unique

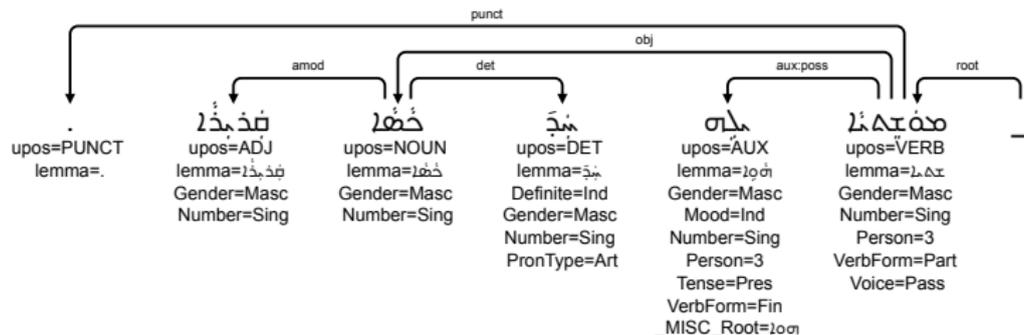
obl POS : 17 tags

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

fac Lemmatisation

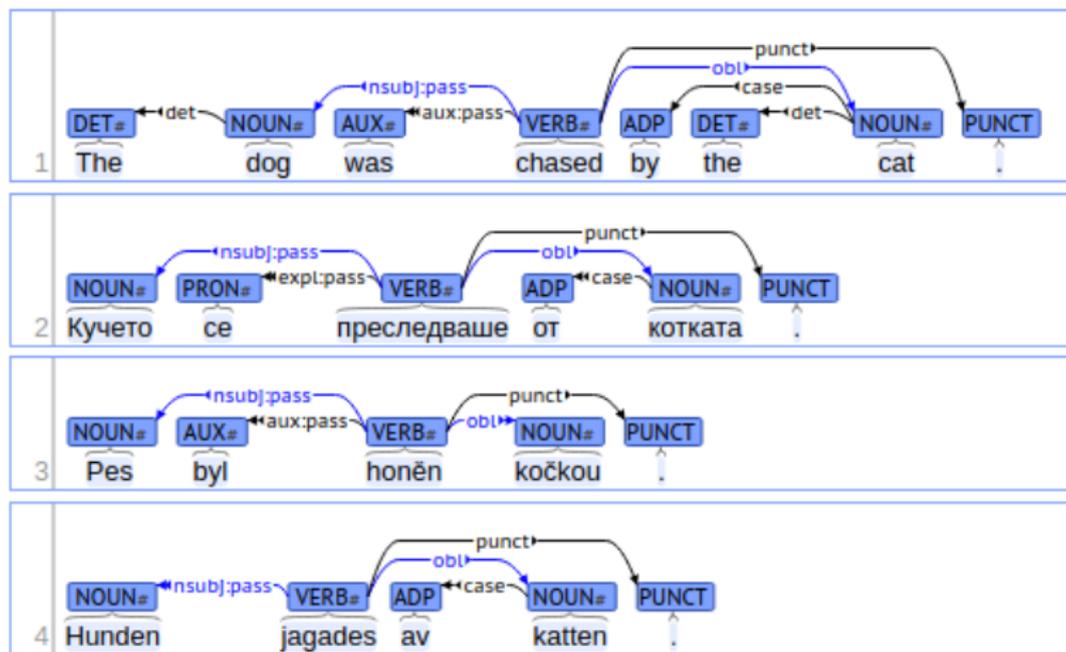
fac Morphologie : 23 traits disponibles

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



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]

<http://grew.fr/>

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
- ▶ interface disponible en ligne (accès à plusieurs corpus)

Interface

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

The screenshot shows the Grew-match web interface. At the top is a dark blue navigation bar with links 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 is a sidebar with a tree view showing 'UD_Faroese-OfI@2.5', 'Finnish', and 'French'. Under 'French', several corpora are listed, including 'UD_French-FTB@2.5' which has a red warning icon. The main area displays 'UD_English-LinES@2.5 [5243 trees, 94217 tokens]' with a 'Hide corpora list' button. Below this is a large empty search area with a '1' in the top left corner. On the right is a search filter panel with tabs for 'Basic' and 'n-grams'. Under 'Basic', there are sections for 'Clusters' and 'Misc'. The 'Misc' section contains several search criteria with input fields: '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)'.

Documentation :

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

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 d'un **même** verbe en français.

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	PROP	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@2.5

/!\ << (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@2.5

/!\ tokénisation préalable

Chercher

▶ *Sahara occidental*

Exclusions

En français, dans UD_French-GSD@2.5

/!\ 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@2.5

/!\ 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

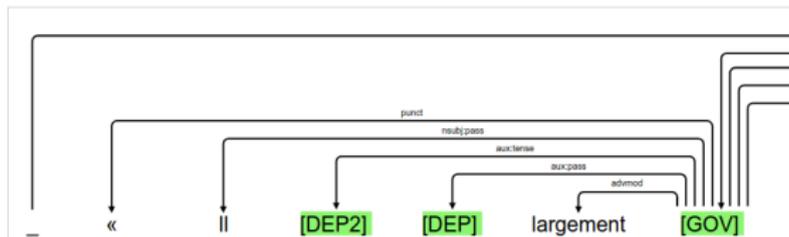
GREW-match trouve toutes les permutations

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.

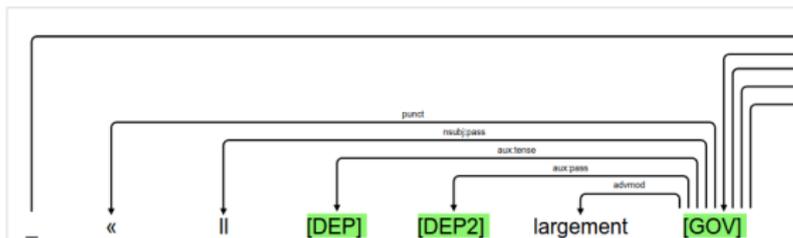


Get more results

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

/!\ <> (différent)

Accords

- ▶ chercher l'accord 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

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.