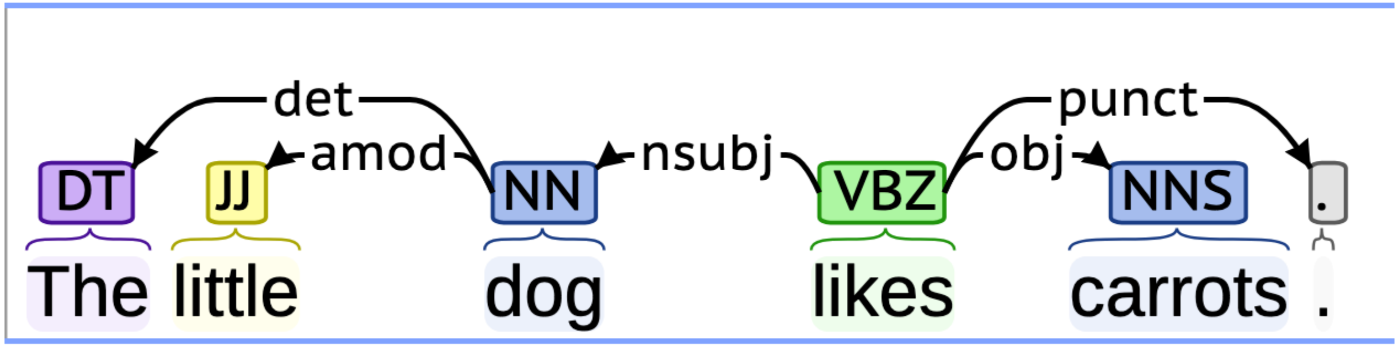
PARSING

Les difficultés des parsers

- Les ambiguïtés
- Les structures complexes
- Les unités phraséologiques (MWE : multi-word expressions, MWU : multi-word units). Le projet PARSEME

Les PARSERS (analyseurs syntaxiques)

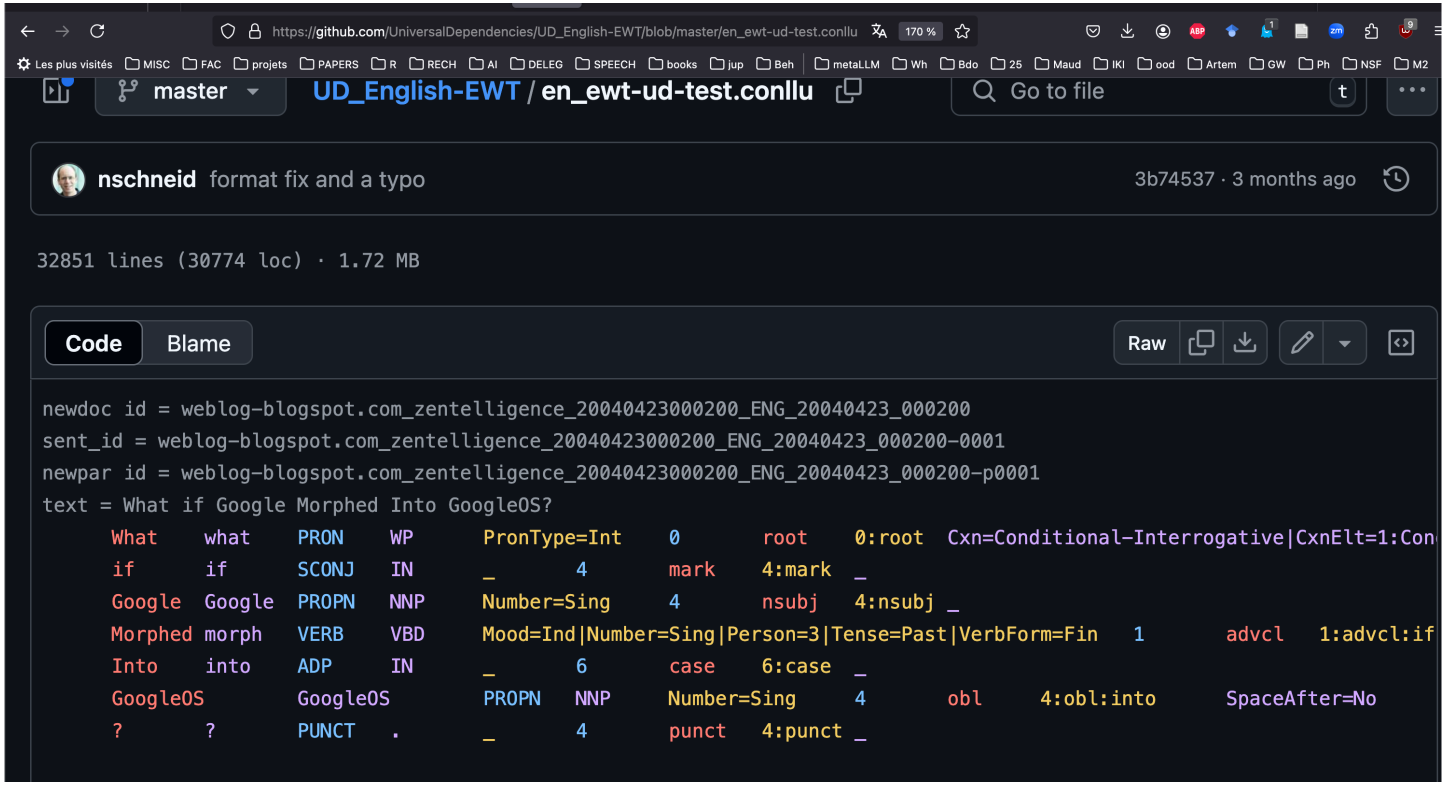
- Deux générations d’analyse : analyse en arbres, puis analyse en relations de dépendances
- Des relations de dépendances (deprel) entre une tête (HEAD, GOV) et son dépendant



Parsers en dépendance (suite)

- Des catégories posées comme universelles (upos)
- Des catégories issues de jeux d’étiquettes (tagsets : xpos)
- Un format tabulaire vertical :
- Des banques d’exemples (Treebanks) pour le train, dev, test
- Des outils d’apprentissage : UDPipe

Dependency parsing : format CONNL-U



UD_English-EWT / en_ewt-ud-test.conllu

nschneid format fix and a typo 3b74537 · 3 months ago

32851 lines (30774 loc) · 1.72 MB

```
newdoc id = weblog-blogspot.com_zentelligence_20040423000200_ENG_20040423_000200
sent_id = weblog-blogspot.com_zentelligence_20040423000200_ENG_20040423_000200-0001
newpar id = weblog-blogspot.com_zentelligence_20040423000200_ENG_20040423_000200-p0001
text = What if Google Morphed Into GoogleOS?
      What    what    PRON    WP      PronType=Int    0       root    0:root  Cxn=Conditional-Interrogative|CxnElt=1:Con
      if      if      SCONJ   IN      _       4       mark    4:mark  _
      Google  Google  PROPN   NNP     Number=Sing     4       nsubj   4:nsubj _
      Morphed morph   VERB    VBD     Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   1       advcl   1:advcl:if
      Into    into    ADP     IN      _       6       case    6:case  _
      GoogleOS        GoogleOS        PROPN   NNP     Number=Sing     4       obl     4:obl:into      SpaceAfter=No
      ?       ?       PUNCT   .       _       4       punct   4:punct _
```

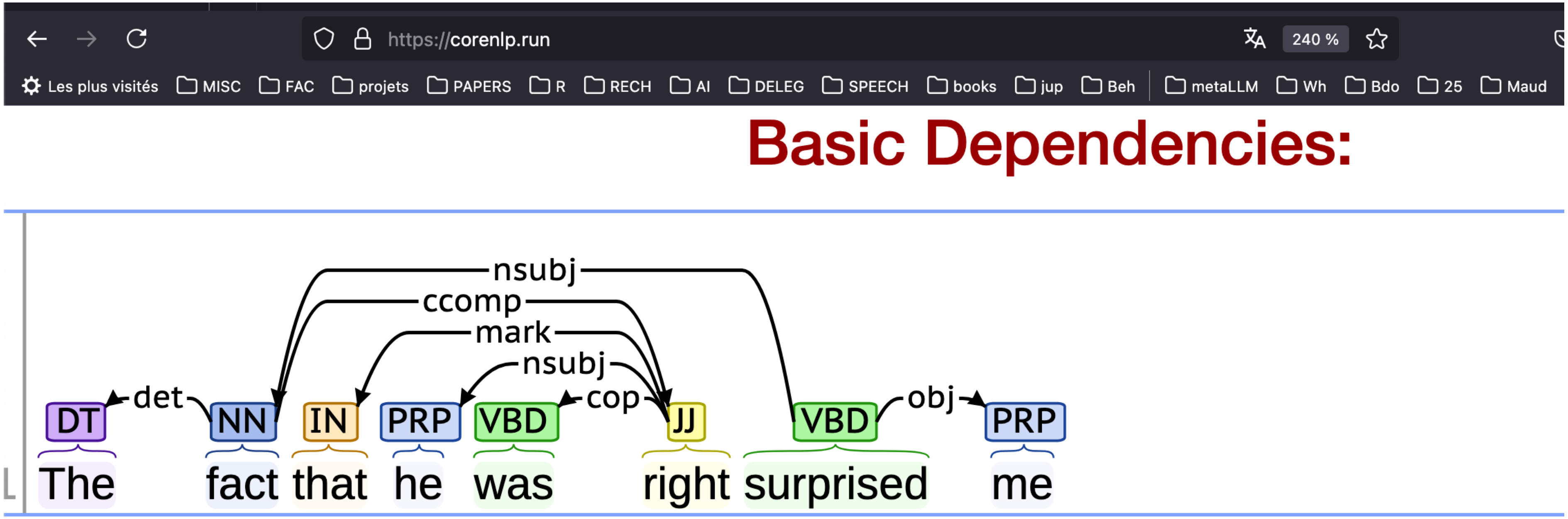
CoNLL-U "Conference on Natural Language Learning - Universal Dependencies"

- a standardized format for annotating linguistic data, particularly for tasks in computational linguistics and natural language processing (NLP)
- . The CoNLL-U format is an extension of the original CoNLL format, specifically designed to support the Universal Dependencies project
- .**Key Features of CoNLL-U**

1. **Structure**: CoNLL-U files are plain text files (UTF-8 encoded) that contain three types of lines[1]
2. : Word lines with 10 tab-separated fields for token annotation
3. Blank lines marking sentence boundaries
4. Comment lines starting with '#' for sentence-level metadata
5. **Token Annotation**: Each word or token is described using 10 fields, including[1]
6. : ID: Token identifier
7. FORM: The word form as it appears in the text
8. LEMMA: The base or dictionary form of the word
9. UPOS: Universal part-of-speech tag
10. HEAD: Syntactic head in dependency parsing
11. DEPREL: Dependency relation to the HEAD
12. **Extensibility**: The CoNLL-U format can be extended to CoNLL-U Plus for additional annotation layers while maintaining compatibilit
13. **Standardization**: CoNLL-U has become a standard format in NLP due to its simplicity and effectiveness in handling annotated linguistic data

- The CoNLL-U format facilitates various NLP tasks, including part-of-speech tagging, syntactic parsing, and named entity recognition, by providing a structured and standardized way to represent linguistic annotation

CoreNLP (Stanford)

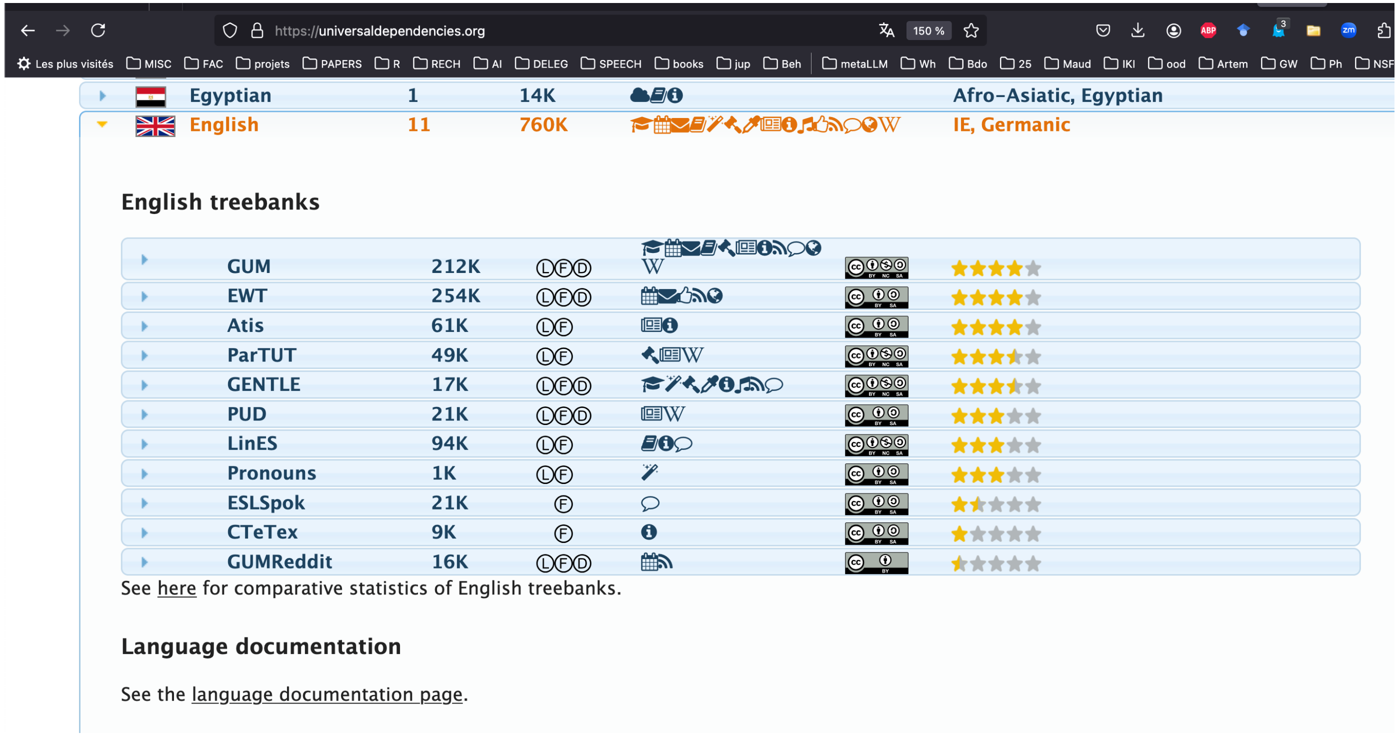


Des Treebanks disponibles sur Github

- https://github.com/UniversalDependencies

- Documentation :
- https://universaldependencies.org/

En anglais



Egyptian	1	14K	Afro-Asiatic, Egyptian
English	11	760K	IE, Germanic

English treebanks

GUM	212K	ⓁⒻⒹ	CC BY NC SA	★★★★☆
EWT	254K	ⓁⒻⒹ	CC BY SA	★★★★☆
Atis	61K	ⓁⒻ	CC BY SA	★★★★☆
ParTUT	49K	ⓁⒻ	CC BY NC SA	★★★½☆
GENTLE	17K	ⓁⒻⒹ	CC BY NC SA	★★★½☆
PUD	21K	ⓁⒻⒹ	CC BY SA	★★★☆☆
LinES	94K	ⓁⒻ	CC BY NC SA	★★★☆☆
Pronouns	1K	ⓁⒻ	CC BY SA	★★★☆☆
ESLSpok	21K	Ⓕ	CC BY SA	★½☆☆☆
CTeTex	9K	Ⓕ	CC BY SA	★☆☆☆☆
GUMReddit	16K	ⓁⒻⒹ	CC BY	½☆☆☆☆

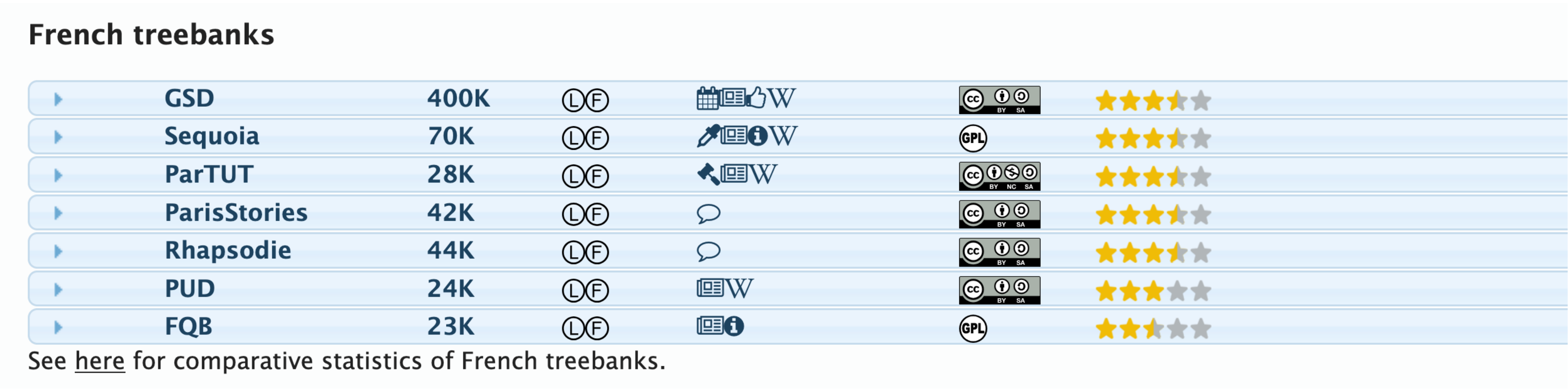
See here for comparative statistics of English treebanks.

Language documentation

See the language documentation page.

https://universaldependencies.org/

En français



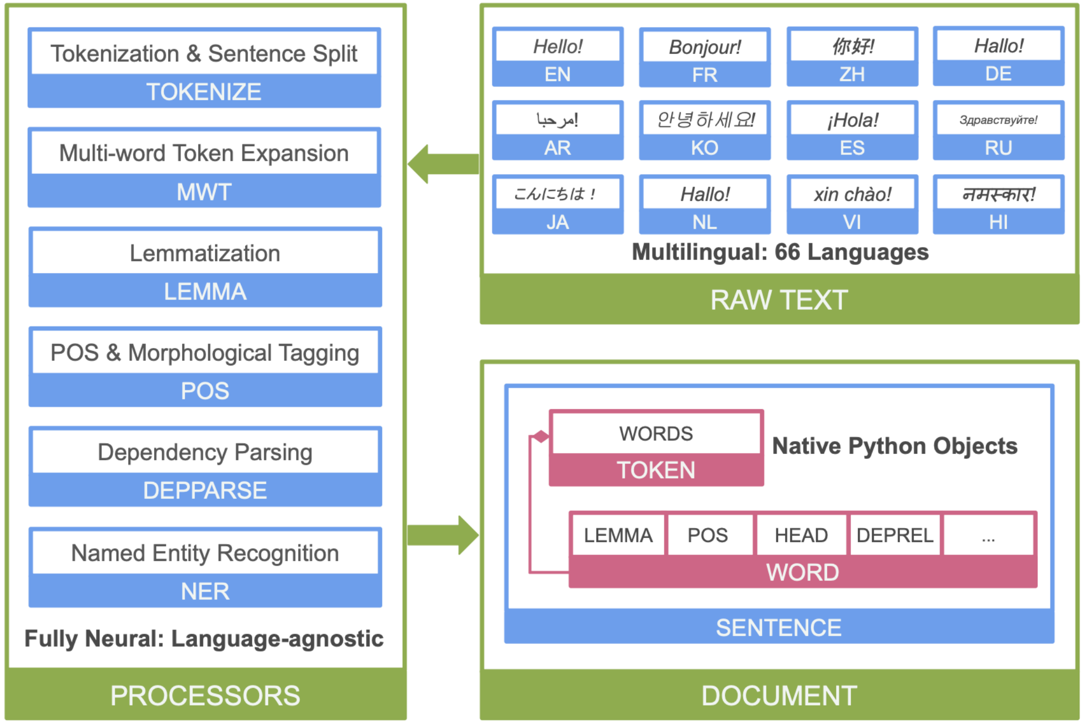
French treebanks

▸	**GSD**	**400K**	ⓁⒻ	📅📰👍W	CC BY SA	★★★⯪☆
▸	**Sequoia**	**70K**	ⓁⒻ	💉📰ℹW	GPL	★★★⯪☆
▸	**ParTUT**	**28K**	ⓁⒻ	🔨📰W	CC BY NC SA	★★★⯪☆
▸	**ParisStories**	**42K**	ⓁⒻ	💬	CC BY SA	★★★⯪☆
▸	**Rhapsodie**	**44K**	ⓁⒻ	💬	CC BY SA	★★★⯪☆
▸	**PUD**	**24K**	ⓁⒻ	📰W	CC BY SA	★★★☆☆
▸	**FQB**	**23K**	ⓁⒻ	📰ℹ	GPL	★★⯪☆☆

See here for comparative statistics of French treebanks.

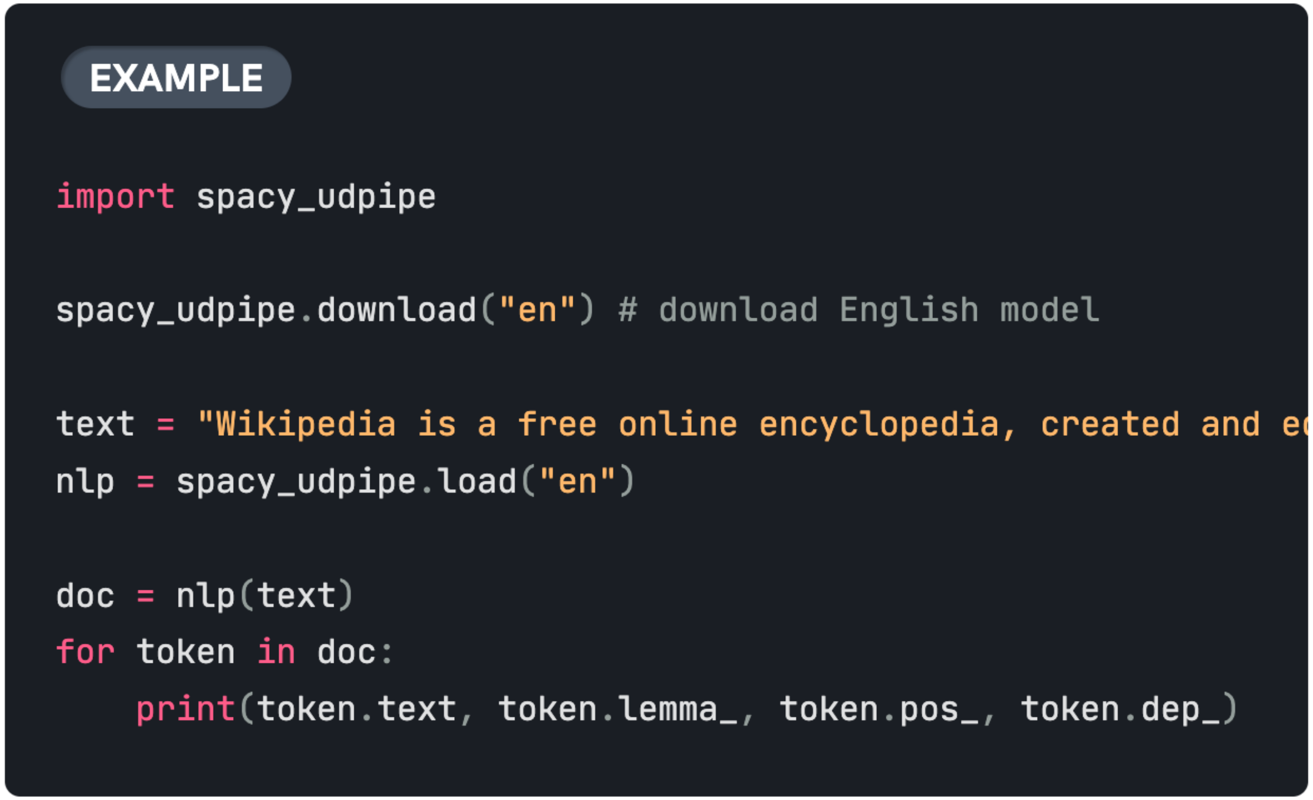
https://universaldependencies.org/

Stanza



https://stanfordnlp.github.io/stanza/.
https://aclanthology.org/2020.acl-demos.14.pdf

Spacy



EXAMPLE

```
import spacy_udpipe

spacy_udpipe.download("en") # download English model

text = "Wikipedia is a free online encyclopedia, created and e
nlp = spacy_udpipe.load("en")

doc = nlp(text)
for token in doc:
    print(token.text, token.lemma_, token.pos_, token.dep_)
```

- https://spacy.io/universe/project/spacy-udpipe

Le réentraînement de Treebanks spécialisées

- Stanza uses multiple English treebanks for different purposes:

- For general English, Stanza uses the Universal Dependencies (UD) treebanks, specifically the English Web Treebank (EWT)
- . For biomedical text, Stanza provides two separate syntactic analysis pipelines: a. One trained on the CRAFT (Colorado Richly Annotated Full Text) treebank
- .
- b. Another trained on the GENIA treebank
- . For clinical text, Stanza uses a combination of: a. The English Web Treebank (EWT).
- b. A silver-standard MIMIC treebank created from clinical notes in the MIMIC-III database
- .

- These treebanks are combined in various ways to improve robustness and performance across different domains. For instance, the CRAFT pipeline combines the English Web Treebank with the CRAFT treebank, while the clinical pipeline combines the EWT with the silver-standard MIMIC treebank.