Universität Stuttgart

Mathilde Regnault,
Thomas Rainsford

Create and manage your own treebank. Workshop in Konstanz.

Corpus Annotation

What is a parser?

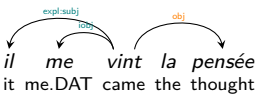
Program performing automatic syntactic analysis

- Symbolic parsers
- Statistical parsers
- Nowadays mostly with RNN

  comprehensive reading: Jurafsky D. and Martin J. H., *Speech and Language Processing*, 3rd ed. draft.

Evaluation:

- UPOS: accuracy of the Part-of-speech tags (UD tagset)
- UAS: Unlabeled Attachment Score
- **LAS: Labeled Attachment Score**
- Performed over:
  - dev: development set (c. 10% of the data)
  - test: test set (c. 10% of the data)



the thought came to me (*L'École des femmes*)

Online parsing with UDPipe

Some parsers are available as **webservices**.

ex. UDPipe [Straka and Straková, 2020]

- **Functions**: tagging, lemmatization, parsing
- **Models**: trained on UD treebanks
- **Input options**: plain text, tokenized text and CoNLL-U
- **Output options**: text, table, trees

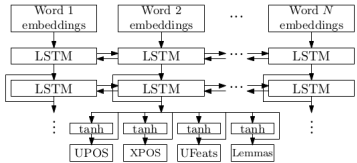


Figure 2: Tagger and lemmatizer model.

Abbildung: UDPipe's architecture [Straka, 2018]

Can ChatGPT annotate my data?

Yes...

represented in CoNLL-U format:

plaintext — Copier le code

```
# text = To proceed then to my Propositions, I shall begin with this.
1   To   to   PART   _   _   2   mark   _   _
2   proceed proceed VERB   _   _   10  advcl  _   _
3   then   then   ADV _   _   2   advmod  _   _
4   to   to   ADP _   _   6   case   _   _
5   my   my   DET _   _   6   det _   _
6   Propositions   Propositions   NOUN   _   _   2   obl _   _
7   ,   ,   PUNCT   _   _   10  punct  _   _
8   I   I   PRON   _   _   10  nsubj  _   _
9   shall   shall   AUX _   _   10  aux _   _
10  begin   begin   VERB   _   _   0   root   _   _
11  with   with   ADP _   _   12  case   _   _
12  this   this   PRON   _   _   10  obl _   _
```

Abbildung: Parsing of one sentence with *ChatGPT* after 4 attempts

But...

- it is **time-consuming**
- and **prone to bugs**.

Local parsing with HOPS

HOPS: *Honest Parser of Sentences* [Grobol and Crabbé, 2021]

- easy to install and use
- proved useful on historical data [Grobol et al., 2022]

Model	UPOS	UAS	LAS
[Straka et al., 2019]	96.26	91.83	86.75
mBERT	96.19	92.03	87.52
BERTrade-petit	96.60	92.20	87.95
BERTrade-mBERT	97.11	93.86	90.37
BERTrade-FlauBERT	97.15	93.96	90.57
BERTrade-CamemBERT	97.29	94.36	90.90

Tabelle: Results on SRCMF test

Installing a parser

+ Train your own models

– Requires resources (machine + time)

NLP teams are often eager to learn about your resources...

Manual Annotation

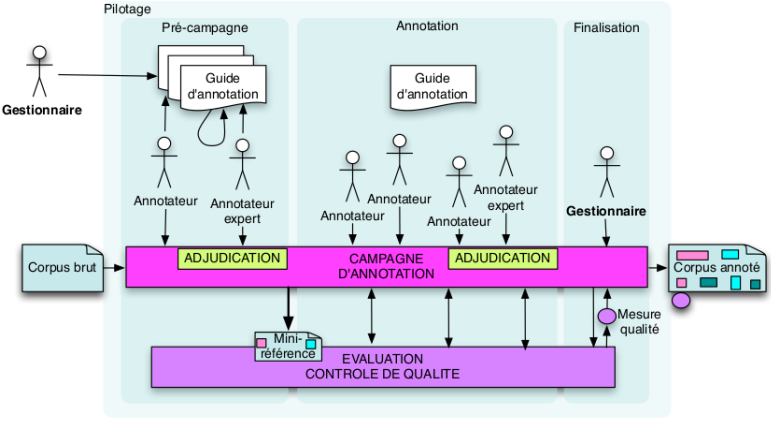


Abbildung: Annotation campaign, K. Fort, *Master's course in Sorbonne Université*

Here are other recommendations: [Grouin et al., 2011, Fort et al., 2012a, Fort et al., 2012b, Fort, 2016]

HOPS demo

Train your model:

- Select your data (train/dev/test)
  in the right **data format**
- Select your word embeddings model and choose configuration
  List from the *transformers* library (*HuggingFace*)
  If no such model available for your language:
  - training without (noBERT config)
  - use a multilingual model
  - create your own :-)
- Experiments:
  - Do several trainings (3-5) to start at random in the data (random seed)
  - Play with the parameters (ex. the number of epochs)
- Let's take a quick look into a server

Corpus Management

2

Get to know the data

Uploading your treebanks on Arborator-GREW:

- Google or GitHub account
- **Correct CoNLL-U files**

Let's try! Select one of the options below:

- Middle French repository

  Grandes Chroniques de France, 1375-1380.
  Gold annotation from UD_Middle_French-PROFITEROLE v. 2.14
  parsed with UDPipe + Old French model and HOPS + SRCMF-Flaubert model
  (no specific model for Middle French... yet)

- Early Modern English repository

  Robert Boyle, *The Sceptical Chymist*, 1661. Source: Wikisource
  Annotated with an English HOPS model (not state-of-the-art)

- Upload your own treebank

What do you think of the quality of the annotation? How should we measure it?

Exploit data

With GREW online/local softwares, you can...

- **visualize** and **share** your data (Arborator-GREW),
- **edit** your treebank with GREW-web
  and **retrain the parser** directly on the platform

Here are some projects we did:

- Enrich a lexicon with verb valency frames using GREW: GitHub repository
- Enrich a lexicon with semantic types for noun
  temporal nouns used as modifiers without a preposition, ex. *that day, last week...*
- **Error mining** for PROFITEROLE

References I

[Fort, 2016] Fort, K. (2016).
Collaborative annotation for reliable natural language processing: Technical and sociological aspects.
John Wiley & Sons.

[Fort et al., 2012a] Fort, K., François, C., Galibert, O., and Ghribi, M. (2012a).
Analyzing the impact of prevalence on the evaluation of a manual annotation campaign.
In *International Conference on Language Resources and Evaluation (LREC)*.

[Fort et al., 2012b] Fort, K., Nazarenko, A., and Rosset, S. (2012b).
Modeling the complexity of manual annotation tasks: a grid of analysis.
In *International Conference on Computational Linguistics*, pages 895–910.

[Grobol and Crabbé, 2021] Grobol, L. and Crabbé, B. (2021).
Analyse en dépendances du français avec des plongements contextualisés.
In *Actes de la 28ème Conférence sur le Traitement Automatique des Langues Naturelles*.

[Grobol et al., 2022] Grobol, L., Regnault, M., Suarez, P. O., Sagot, B., Romary, L., and Crabbé, B. (2022).
Bertrade: Using contextual embeddings to parse old french.
In *13th Language Resources and Evaluation Conference*.

References II

[Grouin et al., 2011] Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011).
Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview.
In *Proceedings of the 5th linguistic annotation workshop*, pages 92–100.

[Straka, 2018] Straka, M. (2018).
UDPipe 2.0 prototype at CoNLL 2018 UD shared task.
In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

[Straka and Straková, 2020] Straka, M. and Straková, J. (2020).
UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings.
In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 124–129, Marseille, France. European Language Resources Association (ELRA).

[Straka et al., 2019] Straka, M., Straková, J., and Hajič, J. (2019).
Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing.
arXiv:1908.07448 [cs].