NTNU | Norwegian University of Science and Technology

Aligning the Norwegian UD Treebank with Entity and Coreference Information

LREC-COLING 2024, Tollef Jørgensen

Aligning the Norwegian UD Treebank with Entity and Coreference Information

A system for alignment of Universal Dependencies treebanks

Specialized usage:
Annotated (noisy) data, based on (subsets of) a treebank.
Handle formats/conversion.

General usage:
Existing treebank annotations, but maps to an older version of UD

Aligning the Norwegian UD Treebank with Entity and Coreference Information

Specialized usage for data-specific formatting and more

Specialized usage:
Annotated (noisy) data, based on (subsets of) a treebank. Handle formats/conversion.

General usage:
Existing treebank annotations, but maps to an older version of UD

Aligning the Norwegian UD Treebank with Entity and Coreference Information

General case, where any pair of misaligned CoNLL-data is supported

Specialized usage:
Annotated (noisy) data, based on (subsets of) a treebank. Handle formats/conversion.

General usage:
Existing treebank annotations, but maps to an older version of UD

Overview

- The problem
- The data
- Conversion
- Alignment
- Examples of the specialized and general cases

Annotated treebank,
outdated UD version

Subsets of UD, noisy data

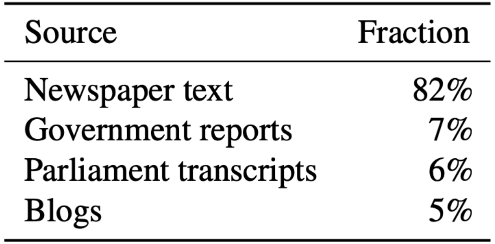
Why write about this?

1. Coreference data is tricky.
    - Formatting issues and edge-cases often arise. Lack of existing tools.
2. UD Treebanks are great sources of lingusitic information, and we wish to integrate more data into them, and hopefully share useful resources.
3. Combining the above, we added Norwegian as the 12th supported language in CorefUD (version 1.1)
    - An initiative to unify coreference corpora.
    - Results at the CRAC workshop, EMNLP 2023.

https://ufal.mff.cuni.cz/corefud

The Norwegian dep treebank (NDT)

- Presented at LREC 2014 (Solberg et al., 2014)
- Mostly newspaper text
- Norwegian *Bokmål* and Norwegian *Nynorsk* (two written forms)
  - 20,000 and 17,600 sentences
  - 310k and 301k tokens
- Converted and included in Universal Dependencies (Øvrelid and Hohle, 2016)



Source	Fraction
Newspaper text	82%
Government reports	7%
Parliament transcripts	6%
Blogs	5%

Adding entity and coreference data

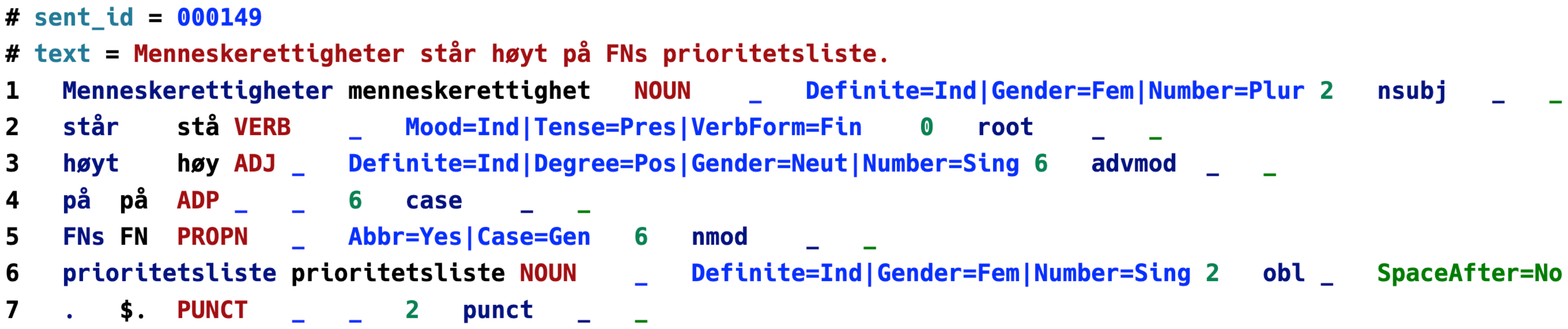
- Two largely varied annotation processes
- Entities
  - NorNE (Jørgensen et al., 2019)
- Coreference
  - NARC (Mæhlum et al., 2022)

Follows the UD data and identifiers

Problematic!
We'll get to this...

What do we have?

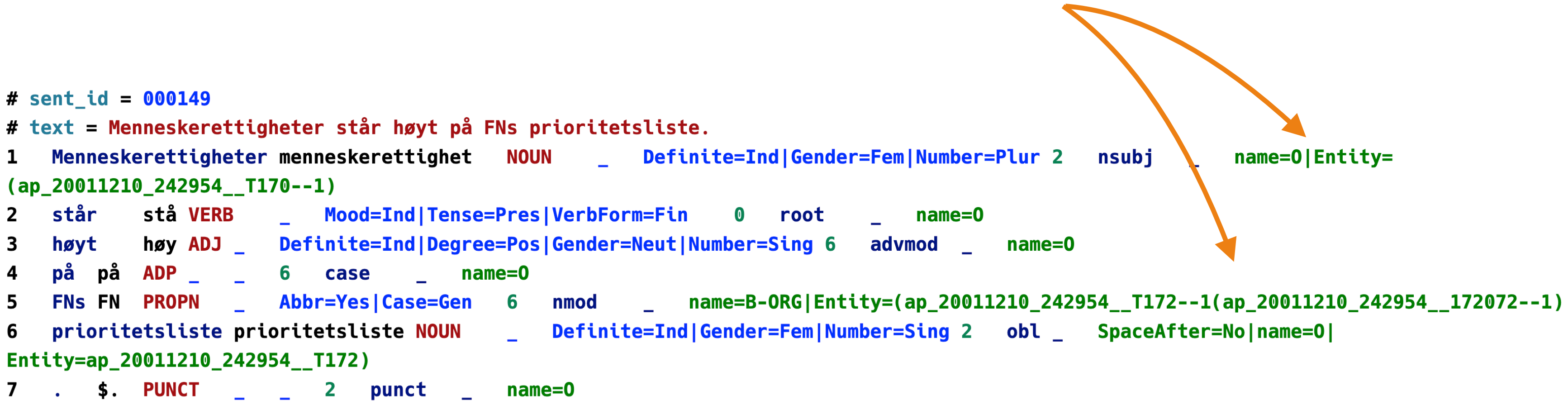
1. A UD treebank
2. Annotated data in an external format (coreference)
3. Annotated, but outdated, treebank (entities)



```
# sent_id = 000149
# text = Menneskerettigheter står høyt på FNs prioritetsliste.
1   Menneskerettigheter menneskerettighet   NOUN    _   Definite=Ind|Gender=Fem|Number=Plur 2   nsubj   _   _
2   står    stå VERB    _   Mood=Ind|Tense=Pres|VerbForm=Fin    0   root    _   _
3   høyt    høy ADJ _   Definite=Ind|Degree=Pos|Gender=Neut|Number=Sing 6   advmod  _   _
4   på  på  ADP _   _   6   case    _   _
5   FNs FN  PROPN   _   Abbr=Yes|Case=Gen   6   nmod    _   _
6   prioritetsliste prioritetsliste NOUN    _   Definite=Ind|Gender=Fem|Number=Sing 2   obl _   SpaceAfter=No
7   .   $.  PUNCT   _   _   2   punct   _   _
```

What do we want?

- CorefUD uses the CoNLL-U format
  - Treebanks with added information fields



```
# sent_id = 000149
# text = Menneskerettigheter står høyt på FNs prioritetsliste.
1   Menneskerettigheter menneskerettighet   NOUN    _   Definite=Ind|Gender=Fem|Number=Plur 2   nsubj   _   name=O|Entity=
(ap_20011210_242954__T170--1)
2   står    stå VERB    _   Mood=Ind|Tense=Pres|VerbForm=Fin    0   root    _   name=O
3   høyt    høy ADJ _   Definite=Ind|Degree=Pos|Gender=Neut|Number=Sing 6   advmod  _   name=O
4   på  på  ADP _   _   6   case    _   name=O
5   FNs FN  PROPN   _   Abbr=Yes|Case=Gen   6   nmod    _   name=B-ORG|Entity=(ap_20011210_242954__T172--1(ap_20011210_242954__172072--1)
6   prioritetsliste prioritetsliste NOUN    _   Definite=Ind|Gender=Fem|Number=Sing 2   obl _   SpaceAfter=No|name=O|
Entity=ap_20011210_242954__T172)
7   .   $.  PUNCT   _   _   2   punct   _   name=O
```

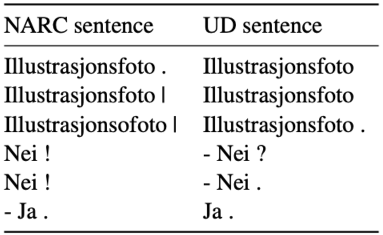
NTNU | Norwegian University of Science and Technology

What's the problem?

- The NARC dataset (Mæhlum et al., 2022)
  - Norwegian Anaphora Resolution Corpus
- Based on the source material in the treebank, but has several ommitted documents/sentences.
- Annotations are sometimes made on subsets, interleaved in the original.
  - [w1, w2, w3, ..., wN-2, wN-1, wN] (original UD)
    - [w1, **w2, w3**, ..., wN-2, **wN-1, wN**] (annotated subsets)

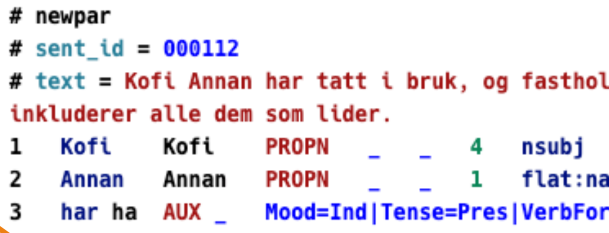
What's the problem?

- No clear mapping.
  - UD: sentence-level id
  - NARC: document-level id
    - cannot match just on raw text, as there are duplicated sentences, among other issues...
- Annotated on a character-level basis
  - Includes "noisy" tokens that must be removed and span-corrected
  - CoNLL-U requires word-level



NARC sentence	UD sentence
Illustrasjonsfoto .	Illustrasjonsfoto
Illustrasjonsfoto |	Illustrasjonsfoto
Illustrasjonsofoto |	Illustrasjonsfoto .
Nei !	- Nei ?
Nei !	- Nei .
- Ja .	Ja .

no_bokmaal-ud-train.conllu



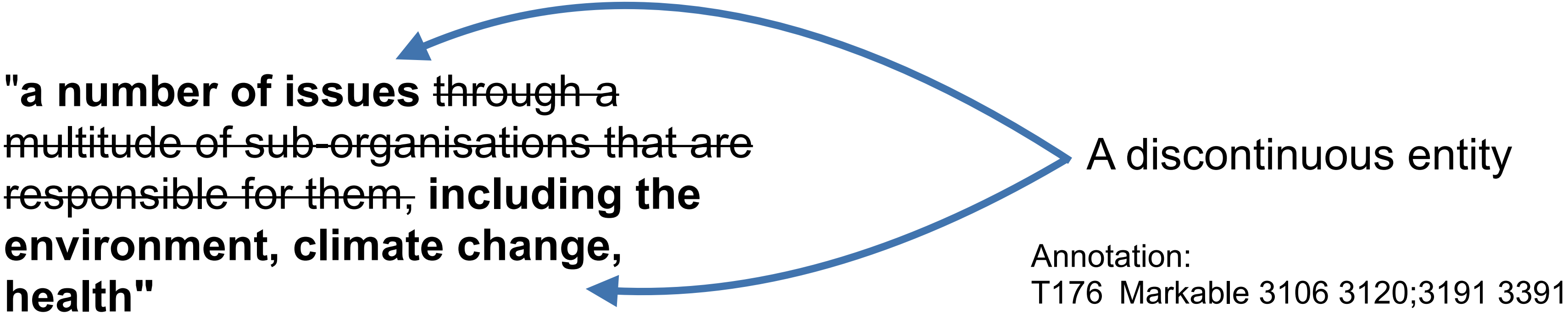
```
# newpar
# sent_id = 000112
# text = Kofi Annan har tatt i bruk, og fasthol
inkluderer alle dem som lider.
1   Kofi    Kofi    PROPN   _   _   4   nsubj
2   Annan   Annan   PROPN   _   _   1   flat:na
3   har ha  AUX _   Mood=Ind|Tense=Pres|VerbFor
```

ap_20011210-242954.ann

```
T41    Markable 803  806 han
T42    Markable 823  833 Kofi Annan
T43    Markable 839  854 de fattiges
T44    Markable 862  864 vi
T45    Markable 869  872 FNs

...

R16    Coref Arg1:T42 Arg2:T41
R17    Coref Arg1:T45 Arg2:T218
```

Coref: A custom BRAT parser

Discontinuous mentions



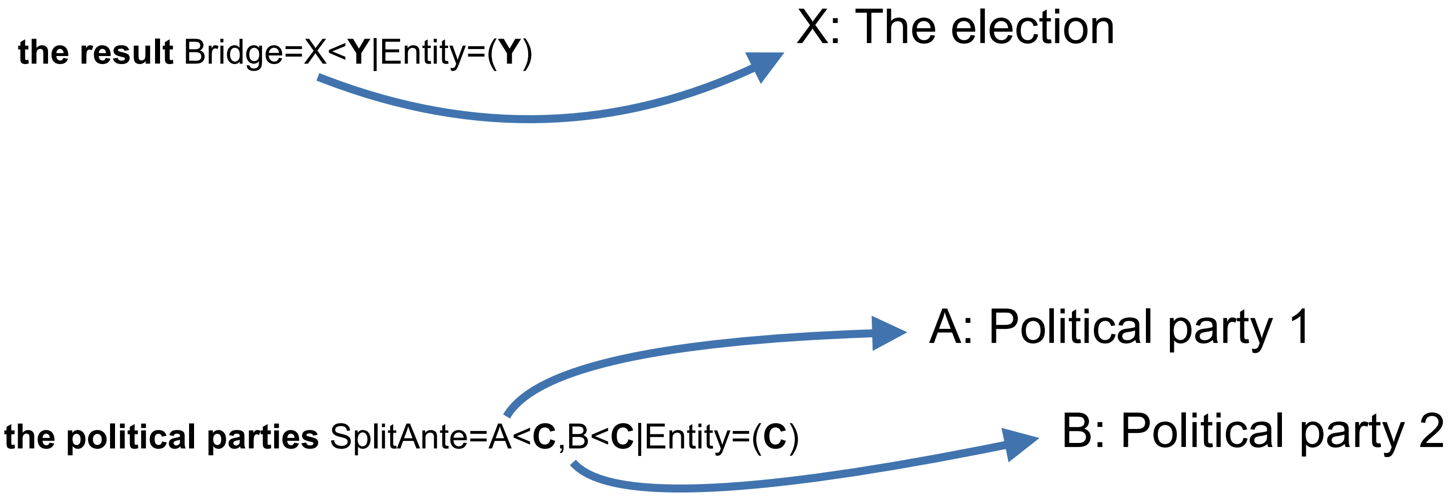
"a number of issues ~~through a multitude of sub-organisations that are responsible for them,~~ **including the environment, climate change, health"**

A discontinuous entity

Annotation:
T176 Markable 3106 3120;3191 3391

Coref: A custom BRAT parser

Bridging and **Split Antecedents**



the result Bridge=X<**Y**|Entity=(**Y**)

X: The election

the political parties SplitAnte=A<**C**,B<**C**|Entity=(**C**)

A: Political party 1

B: Political party 2

Coref: A custom BRAT parser

BRAT format (.txt+.ann)

Doc: ap_20011210-242954.txt

Line 1: FN for fred og utvikling , i de fattiges tjeneste

(UN for peace and development, at the service of the poor)

Annotation: ap_20011210-242954.ann

```
T1	Markable 0 51		FN for fred og utvikling , i de fattiges tjeneste
T2	Markable 7 24		fred og utvikling
T3	Markable 7 11		fred
T4	Markable 15 24		utvikling
T5	Markable 29 49		de fattiges tjeneste
```

Any referable mention (markable)

Character indexes in doc.

jsonline

Doc: ap_20011210-242954.jsonl

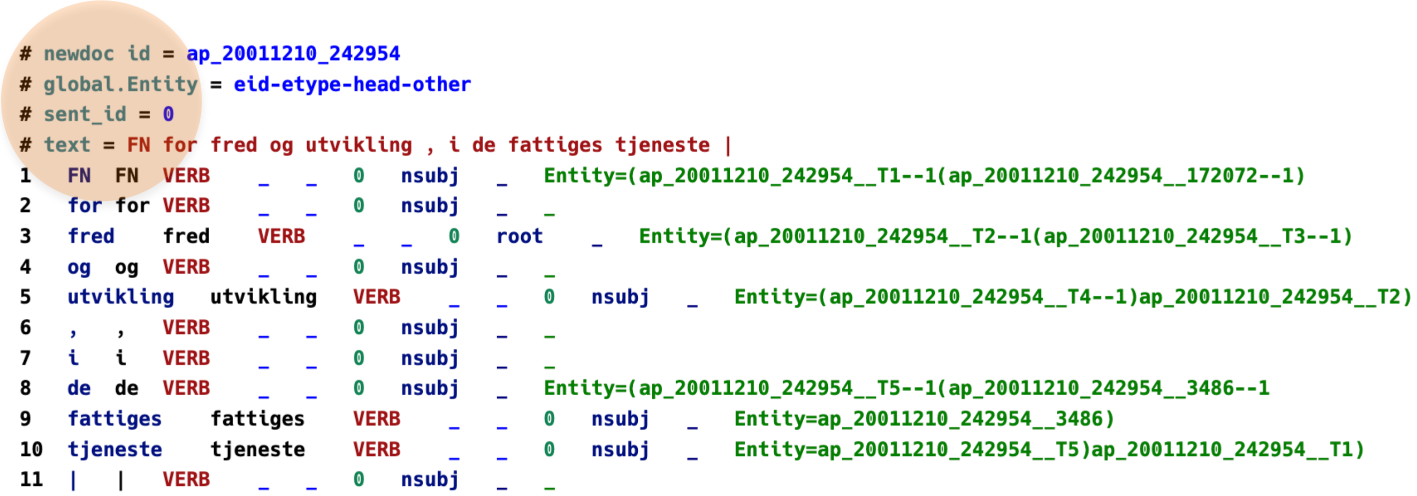
```
{"doc_key": "ap_20011210-242954",
"sentences": [["FN", "for", "fred", "og",
"utvikling", ",", "i", "de", "fattiges",
"tjeneste", ... ]
"clusters": [[[0, 2], [252, 254], [375, 377],
[438, 459], [463, 467], [869, 872], [1032,
1034],  [ ... ]]
```

From markables to coreference clusters

JSONLine to CoNLL-U

From **clusters** [[0, 2], [8, 10], [...]] to **token-based MISC fields**

Still needs alignment! (document to sentence-level)

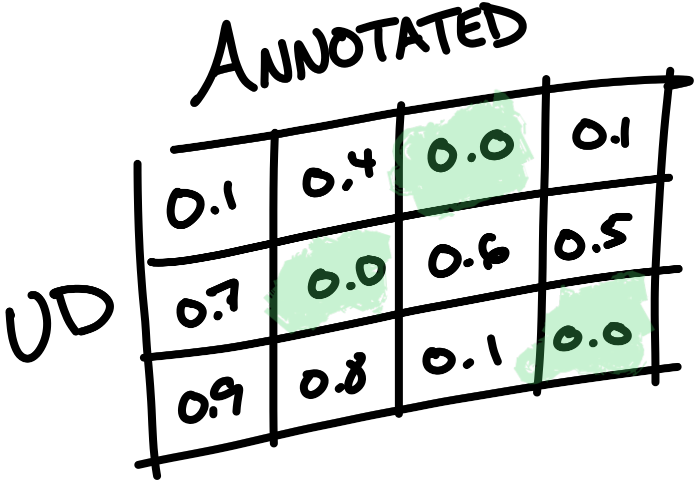


```
# newdoc id = ap_20011210_242954
# global.Entity = eid-etype-head-other
# sent_id = 0
# text = FN for fred og utvikling , i de fattiges tjeneste |
1	FN	FN	VERB	_	_	0	nsubj	_	Entity=(ap_20011210_242954__T1--1(ap_20011210_242954__172072--1)
2	for	for	VERB	_	_	0	nsubj	_	_
3	fred	fred	VERB	_	_	0	root	_	Entity=(ap_20011210_242954__T2--1(ap_20011210_242954__T3--1)
4	og	og	VERB	_	_	0	nsubj	_	_
5	utvikling	utvikling	VERB	_	_	0	nsubj	_	Entity=(ap_20011210_242954__T4--1)ap_20011210_242954__T2)
6	,	,	VERB	_	_	0	nsubj	_	_
7	i	i	VERB	_	_	0	nsubj	_	_
8	de	de	VERB	_	_	0	nsubj	_	Entity=(ap_20011210_242954__T5--1(ap_20011210_242954__3486--1
9	fattiges	fattiges	VERB	_	_	0	nsubj	_	Entity=ap_20011210_242954__3486)
10	tjeneste	tjeneste	VERB	_	_	0	nsubj	_	Entity=ap_20011210_242954__T5)ap_20011210_242954__T1)
11	|	|	VERB	_	_	0	nsubj	_	_
```

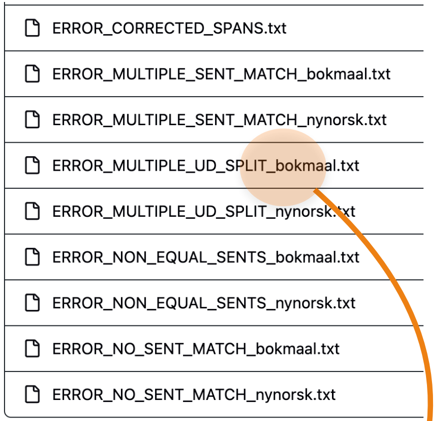
NTNU | Norwegian University of Science and Technology

Alignment

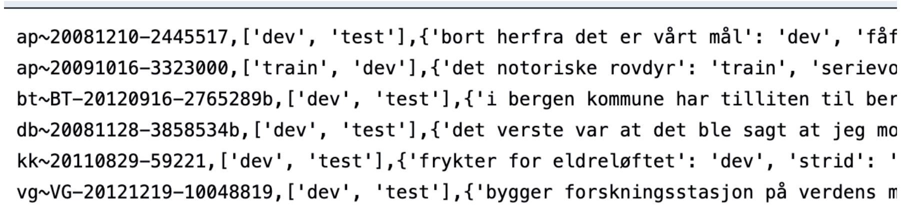
- Map raw texts to their UD identifiers
- Direct mapping of matching annotated sentences to UD IDs, where the joint index matches
- Sentence disambiguation
  - lazy-match: multiple candidates?
    - minimize alignment scores to UD
      - cost matrix ➡ linear sum assignment problem



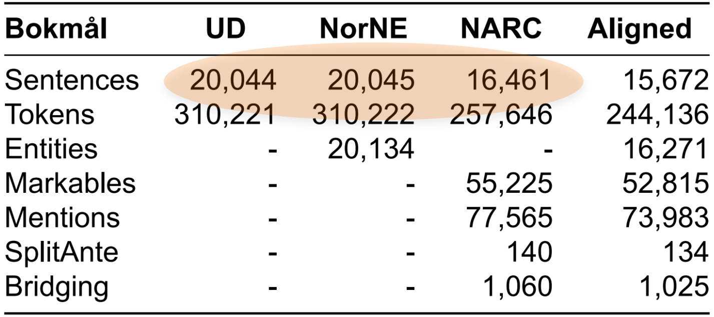
0.1	0.4	0.0	0.1
0.7	0.0	0.6	0.5
0.9	0.8	0.1	0.0



- ERROR_CORRECTED_SPANS.txt
- ERROR_MULTIPLE_SENT_MATCH_bokmaal.txt
- ERROR_MULTIPLE_SENT_MATCH_nynorsk.txt
- ERROR_MULTIPLE_UD_SPLIT_bokmaal.txt
- ERROR_MULTIPLE_UD_SPLIT_nynorsk.txt
- ERROR_NON_EQUAL_SENTS_bokmaal.txt
- ERROR_NON_EQUAL_SENTS_nynorsk.txt
- ERROR_NO_SENT_MATCH_bokmaal.txt
- ERROR_NO_SENT_MATCH_nynorsk.txt

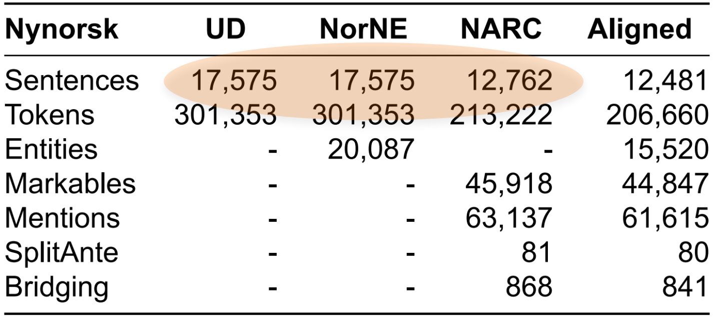


```
ap~20081210-2445517,['dev', 'test'],{'bort herfra det er vårt mål': 'dev', 'fåf
ap~20091016-3323000,['train', 'dev'],{'det notoriske rovdyr': 'train', 'serievo
bt~BT-20120916-2765289b,['dev', 'test'],{'i bergen kommune har tilliten til ber
db~20081128-3858534b,['dev', 'test'],{'det verste var at det ble sagt at jeg mo
kk~20110829-59221,['dev', 'test'],{'frykter for eldreløftet': 'dev', 'strid': '
vg~VG-20121219-10048819,['dev', 'test'],{'bygger forskningsstasjon på verdens m
```



Bokmål	**UD**	**NorNE**	**NARC**	**Aligned**
Sentences	20,044	20,045	16,461	15,672
Tokens	310,221	310,222	257,646	244,136
Entities	-	20,134	-	16,271
Markables	-	-	55,225	52,815
Mentions	-	-	77,565	73,983
SplitAnte	-	-	140	134
Bridging	-	-	1,060	1,025

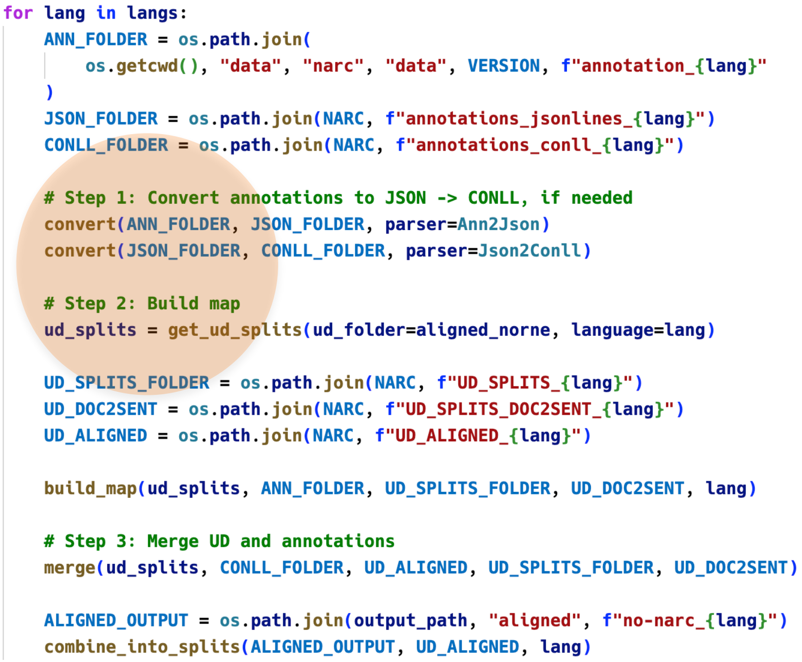
Table 2: Statistics of the Bokmål corpora



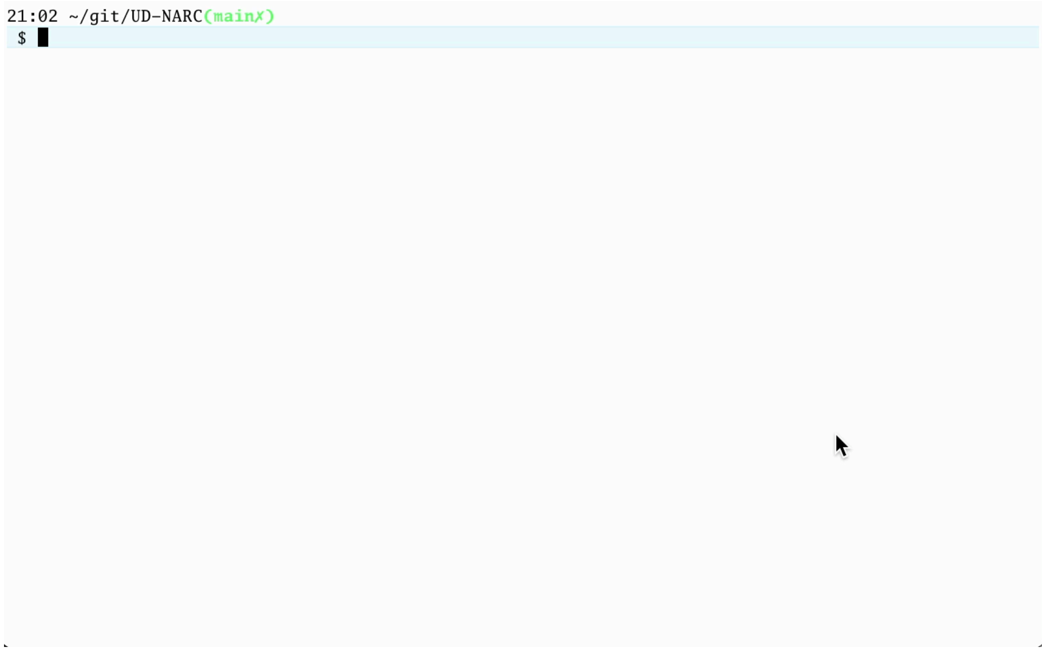
Nynorsk	**UD**	**NorNE**	**NARC**	**Aligned**
Sentences	17,575	17,575	12,762	12,481
Tokens	301,353	301,353	213,222	206,660
Entities	-	20,087	-	15,520
Markables	-	-	45,918	44,847
Mentions	-	-	63,137	61,615
SplitAnte	-	-	81	80
Bridging	-	-	868	841

Table 3: Statistics of the Nynorsk corpora

Specialized case: converting and aligning



```
for lang in langs:
    ANN_FOLDER = os.path.join(
        os.getcwd(), "data", "narc", "data", VERSION, f"annotation_{lang}"
    )
    JSON_FOLDER = os.path.join(NARC, f"annotations_jsonlines_{lang}")
    CONLL_FOLDER = os.path.join(NARC, f"annotations_conll_{lang}")

    # Step 1: Convert annotations to JSON -> CONLL, if needed
    convert(ANN_FOLDER, JSON_FOLDER, parser=Ann2Json)
    convert(JSON_FOLDER, CONLL_FOLDER, parser=Json2Conll)

    # Step 2: Build map
    ud_splits = get_ud_splits(ud_folder=aligned_norne, language=lang)

    UD_SPLITS_FOLDER = os.path.join(NARC, f"UD_SPLITS_{lang}")
    UD_DOC2SENT = os.path.join(NARC, f"UD_SPLITS_DOC2SENT_{lang}")
    UD_ALIGNED = os.path.join(NARC, f"UD_ALIGNED_{lang}")

    build_map(ud_splits, ANN_FOLDER, UD_SPLITS_FOLDER, UD_DOC2SENT, lang)

    # Step 3: Merge UD and annotations
    merge(ud_splits, CONLL_FOLDER, UD_ALIGNED, UD_SPLITS_FOLDER, UD_DOC2SENT)

    ALIGNED_OUTPUT = os.path.join(output_path, "aligned", f"no-narc_{lang}")
    combine_into_splits(ALIGNED_OUTPUT, UD_ALIGNED, lang)
```

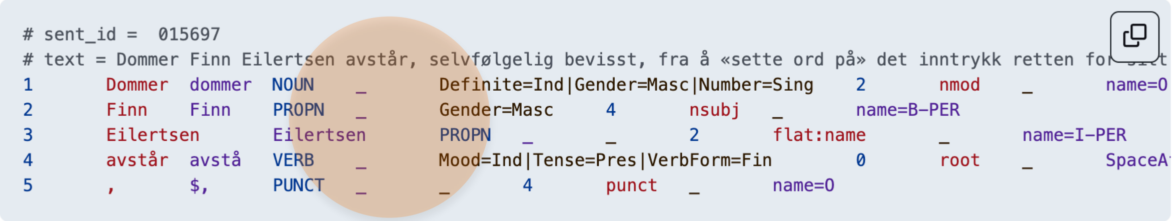


```
21:02 ~/git/UD-NARC(main✗)
 $ 
```

NTNU | Norwegian University of Science and Technology

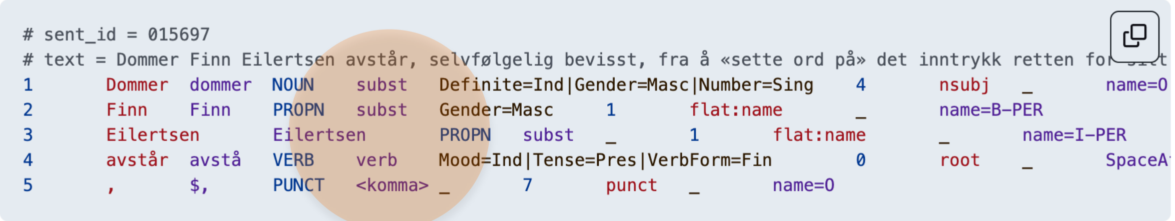
General case: updating outdated treebanks

Original NorNE



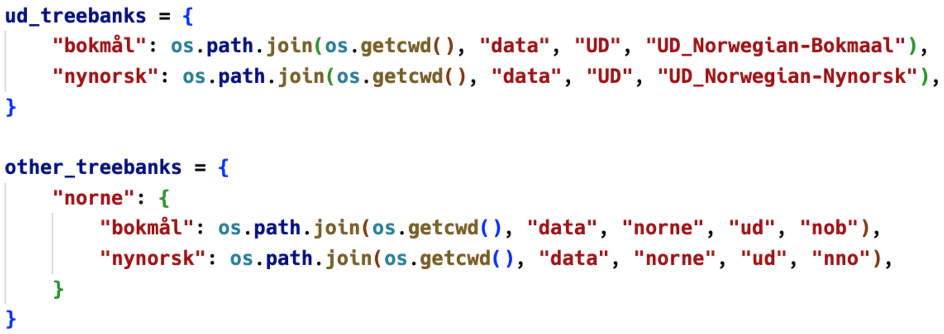
```
# sent_id =  015697
# text = Dommer Finn Eilertsen avstår, selvfølgelig bevisst, fra å «sette ord på» det inntrykk retten for sitt
1       Dommer  dommer  NOUN    _       Definite=Ind|Gender=Masc|Number=Sing    2       nmod    _       name=O
2       Finn    Finn    PROPN   _       Gender=Masc     4       nsubj   _       name=B-PER
3       Eilertsen       Eilertsen       PROPN   _       _       2       flat:name       _       name=I-PER
4       avstår  avstå   VERB    _       Mood=Ind|Tense=Pres|VerbForm=Fin        0       root    _       SpaceA
5       ,       $,      PUNCT   _       _       4       punct   _       name=O
```

Aligned NorNE-UD



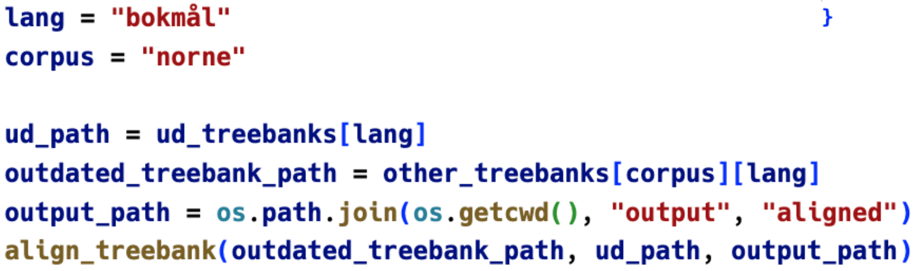
```
# sent_id = 015697
# text = Dommer Finn Eilertsen avstår, selvfølgelig bevisst, fra å «sette ord på» det inntrykk retten for sitt
1       Dommer  dommer  NOUN    subst   Definite=Ind|Gender=Masc|Number=Sing    4       nsubj   _       name=O
2       Finn    Finn    PROPN   subst   Gender=Masc     1       flat:name       _       name=B-PER
3       Eilertsen       Eilertsen       PROPN   subst   _       1       flat:name       _       name=I-PER
4       avstår  avstå   VERB    verb    Mood=Ind|Tense=Pres|VerbForm=Fin        0       root    _       SpaceA
5       ,       $,      PUNCT   <komma> _       7       punct   _       name=O
```

General case: updating outdated treebanks



```
ud_treebanks = {
    "bokmål": os.path.join(os.getcwd(), "data", "UD", "UD_Norwegian-Bokmaal"),
    "nynorsk": os.path.join(os.getcwd(), "data", "UD", "UD_Norwegian-Nynorsk"),
}

other_treebanks = {
    "norne": {
        "bokmål": os.path.join(os.getcwd(), "data", "norne", "ud", "nob"),
        "nynorsk": os.path.join(os.getcwd(), "data", "norne", "ud", "nno"),
    }
}
```

Your data goes here



```
lang = "bokmål"
corpus = "norne"

ud_path = ud_treebanks[lang]
outdated_treebank_path = other_treebanks[corpus][lang]
output_path = os.path.join(os.getcwd(), "output", "aligned")
align_treebank(outdated_treebank_path, ud_path, output_path)
```

NTNU | Norwegian University of Science and Technology

Summary

- A system for aligning UD treebanks
  - Both special/complex cases and more general merging procedures
- Tools for handling coreference data
- Inspection tools for treebank data (split overlaps, unmatched sentences, ...)

- Acknowledgements:
  - Thanks to Michal Novák and Daniel Zeman, ÚFAL, for help with getting the system ready for CorefUD.