Tutorial on Universal Dependencies

Infrastructure, resources and tools for UD

Joakim Nivre[1] Daniel Zeman [2] **Filip Ginter**[3] Francis M. Tyers[45]

[1]Department of Linguistics and Philology, Uppsala University, Sweden

[2]Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic

[3]Department of Information Technology, University of Turku, Finland

[4]Giela ja kultuvrra instituhtta, UiT Norgga árktalaš universitehta, Tromsø, Norway

[5]Arvutiteaduse instituut, Tartu Ülikool, Estonia

UD as of Version 2.0 — Treebanks

How many?

- Languages: **50**
- Treebanks: **70**
- Trees: **630,000**
- Words: **12,103,000**

Can I use them?

- Creative Commons and GPL-like: **28**
- Creative Commons Non-Commercial: **42**

Where from?

- http://universaldependencies.org
- Official release preferred over GitHub

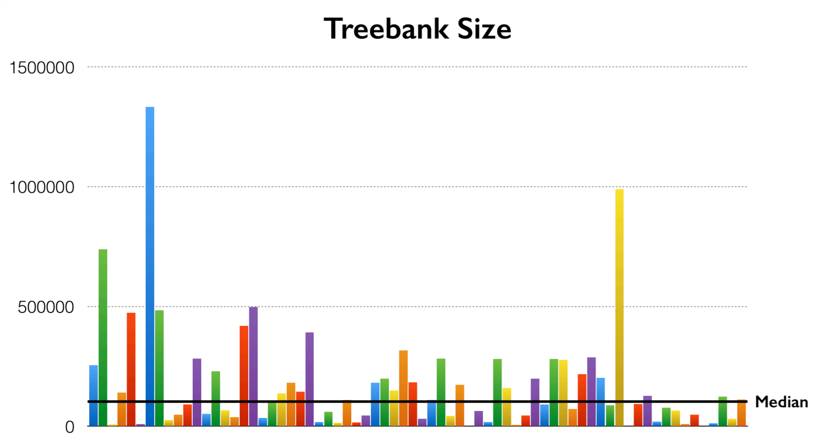
UD Treebanks Come in Many Flavors and Sizes

Annotation:

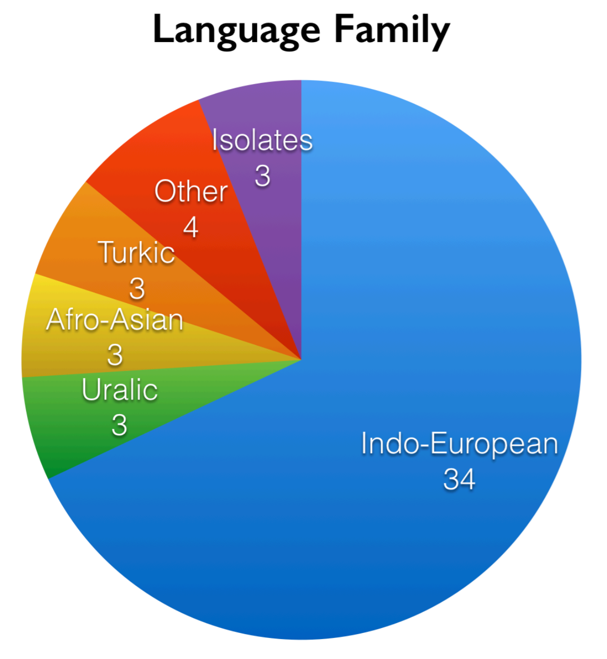
- POS and base dependency relations compulsory
- ...and additionally:
  - Features: XXX
  - Lemmas: XXX
  - Features + Lemmas: XXX
  - Features + Lemmas + Enhanced Relations: XXXX

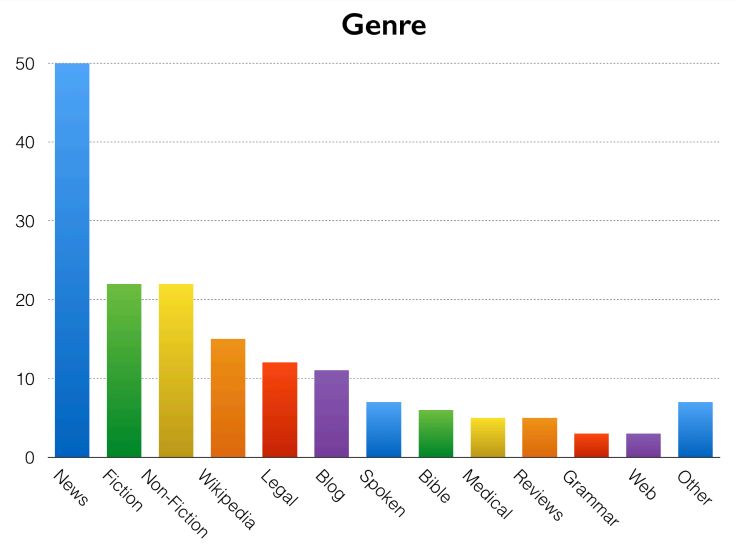
Size:

- Smallest: XXX words
- Largest: XXX words


Treebank Size
1500000
1000000
500000
0
Median

XXX - replace with world map if I have the time to make it happen




Genre
50
40
30
20
10
0
News
Fiction
Non-Fiction
Wikipedia
Legal
Blog
Spoken
Bible
Medical
Reviews
Grammar
Web
Other

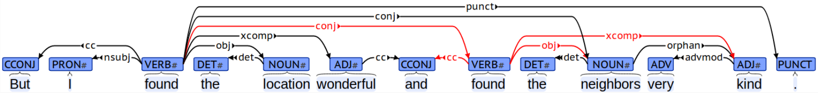
CoNLL-U Format

- Derived from CoNLL-X, overall logic same, details differ
- `ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC`
- Only `ID UPOS HEAD DEPREL` compulsory

Distinguishing features:

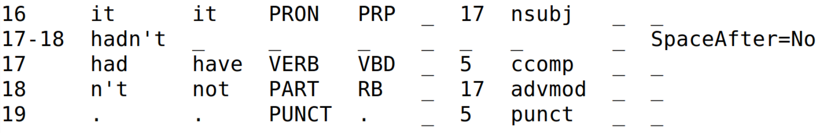
- Sentence-level metadata part of the format
- Explicit (and compulsory!) representation of the original text
- `DEPS` field encodes the enhanced *parse graph*
- `MISC` field allows arbitrary data stored for every word
- Empty nodes — only referred to from the enhanced representation
- Words — only referred to from the enhanced representation

CoNLL-U Format



```
# sent_id = reviews-044427-0003
# text = But I found the location wonderful and the neighbors very kind.
1    But       but       CCONJ       CC   _                                       3    cc      _          _
2    I         I         PRON        PRP  Case=Nom|Number=Sing|Person=1|PronType=Prs 3 nsubj   _          _
3    found     find      VERB        VBD  Mood=Ind|Tense=Past|VerbForm=Fin        0    root    _          _
4    the       the       DET         DT   Definite=Def|PronType=Art               5    det     _          _
5    location  location  NOUN        NN   Number=Sing                             3    obj     _          _
6    wonderful wonderful ADJ         JJ   Degree=Pos                              3    xcomp   _          _
7    and       and       CCONJ       CC   _                                       6    cc      _          _
7.1  found     find      VERB        VBD  Mood=Ind|Tense=Past|VerbForm=Fin        _    _       3:conj     _
8    the       the       DET         DT   Definite=Def|PronType=Art               9    det     _          _
9    neighbors neighbor  NOUN        NNS  Number=Plur                             3    conj    7.1:obj    _
10   very      very      ADV         RB   _                                       11   advmod  _          _
11   kind      kind      ADJ         JJ   Degree=Pos                              9    orphan  7.1:xcomp  SpaceAfter=No
12   .         .         PUNCT       .    _                                       3    punct   _          _
```

CoNLL-U Format — Tokens vs. Words



```
16      it      it    PRON   PRP  _  17  nsubj   _  _
17-18   hadn't  _     _      _    _  _   _       _  SpaceAfter=No
17      had     have  VERB   VBD  _  5   ccomp   _  _
18      n't     not   PART   RB   _  17  advmod  _  _
19      .       .     PUNCT  .    _  5   punct   _  _
```

...more on this later...

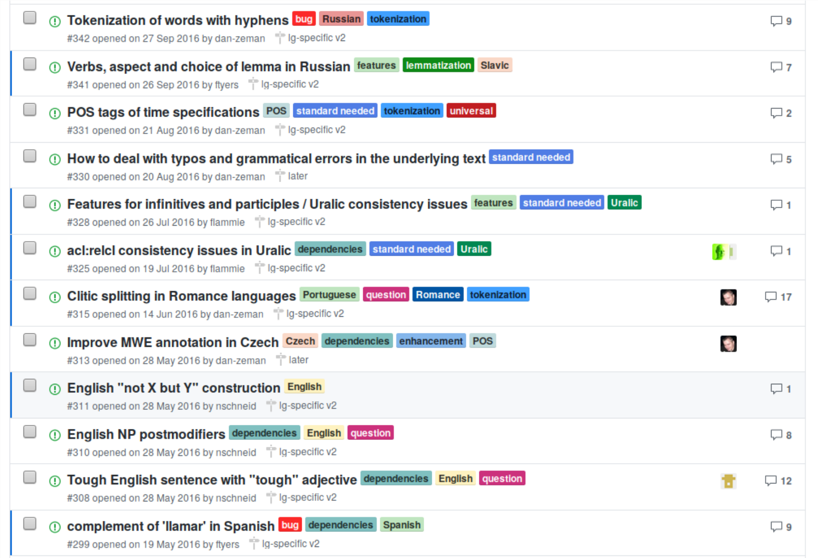
UD Infrastructure - Requirements

- 83 treebank repositories
- 100+ contributors
- Online documentation consisting of roughly 14,000 web-pages
- Guidelines, universal and language-specific
- Discussions, decision making, validation
- Regular, carefully checked official releases
- A comparatively small group of core “staff” running the show
- Budget: $0

UD Infrastructure - GitHub

- GitHub in use from Day 1
- Documentation and data first
- Followed exclusive use of the issue tracker for discussions and proposals
  - Before: many email chains — chaos
- Practically *everything* happens openly

UD is open



□ ⓘ **Tokenization of words with hyphens** bug Russian tokenization 💬 9
#342 opened on 27 Sep 2016 by dan-zeman lg-specific v2

□ ⓘ **Verbs, aspect and choice of lemma in Russian** features lemmatization Slavic 💬 7
#341 opened on 26 Sep 2016 by ftyers lg-specific v2

□ ⓘ **POS tags of time specifications** POS standard needed tokenization universal 💬 2
#331 opened on 21 Aug 2016 by dan-zeman lg-specific v2

□ ⓘ **How to deal with typos and grammatical errors in the underlying text** standard needed 💬 5
#330 opened on 20 Aug 2016 by dan-zeman later

□ ⓘ **Features for infinitives and participles / Uralic consistency issues** features standard needed Uralic 💬 1
#328 opened on 26 Jul 2016 by flammie lg-specific v2

□ ⓘ **acl:relcl consistency issues in Uralic** dependencies standard needed Uralic 💬 1
#325 opened on 19 Jul 2016 by flammie lg-specific v2

□ ⓘ **Clitic splitting in Romance languages** Portuguese question Romance tokenization 💬 17
#315 opened on 14 Jun 2016 by dan-zeman lg-specific v2

□ ⓘ **Improve MWE annotation in Czech** Czech dependencies enhancement POS
#313 opened on 28 May 2016 by dan-zeman later

□ ⓘ **English "not X but Y" construction** English 💬 1
#311 opened on 28 May 2016 by nschneid lg-specific v2

□ ⓘ **English NP postmodifiers** dependencies English question 💬 8
#310 opened on 28 May 2016 by nschneid lg-specific v2

□ ⓘ **Tough English sentence with "tough" adjective** dependencies English question 💬 12
#308 opened on 28 May 2016 by nschneid lg-specific v2

□ ⓘ **complement of 'llamar' in Spanish** bug dependencies Spanish 💬 9
#299 opened on 19 May 2016 by ftyers lg-specific v2

Data

- A GitHub repository for every treebank
  - UD_{Language}-{Treebank}
  - **master** branch holds the most recent official release
  - **dev** branch holds development data, not guaranteed to be valid
  - Some teams use GitHub for development, others only to “submit” their data prior to the release
  - No strict requirements on the workflow
- **Official release:** LINDAT, May & November, all treebanks which contain valid data

Docs

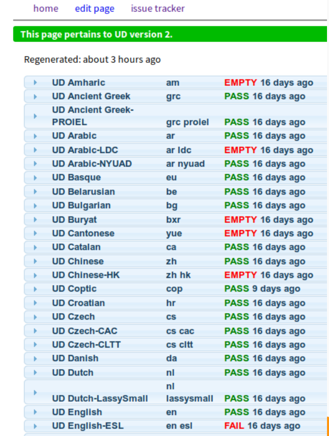
- One set of documentation for every language (not treebank)
- A GitHub repository holding mostly markdown pages
- Special care taken to make it easy to add tree visualizations and examples
- Stubs pre-generated when adding a new language
- 11,000+ commits from 80+ contributors
- Automatically regenerated on every push and published on GitHub pages
- The issue tracker for the *docs* repository is where all the UD activity is happening
    - Hundreds of issues, thousands of replies

Workflow and Organization

- Highly ~~chaotic~~ distributed
- All contributors given broad edit rights to all data, docs, and tools repositories
- Fully trust-based setup, `git` giving a safety net
- Joakim holds the honorary title of *Chief Cat Herder* and looks after the project as a whole — is obeyed unconditionally

Validation

- Script to validate treebank data
- Passing is compulsory
- Format validation
- Runs automatically every time a treebank is updated
- Indispensable especially close to an official release date
- Contributors: do we validate?
- Release team: whom to help next?
- `http://universaldependencies.org/validation.html`



home edit page issue tracker

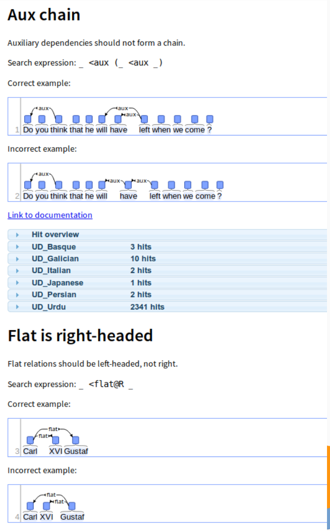
This page pertains to UD version 2.

Regenerated: about 3 hours ago

▸	UD Amharic	am	EMPTY 16 days ago
▸	UD Ancient Greek	grc	PASS 16 days ago
▸	UD Ancient Greek-PROIEL	grc proiel	PASS 16 days ago
▸	UD Arabic	ar	PASS 16 days ago
▸	UD Arabic-LDC	ar ldc	EMPTY 16 days ago
▸	UD Arabic-NYUAD	ar nyuad	PASS 16 days ago
▸	UD Basque	eu	PASS 16 days ago
▸	UD Belarusian	be	PASS 16 days ago
▸	UD Bulgarian	bg	PASS 16 days ago
▸	UD Buryat	bxr	EMPTY 16 days ago
▸	UD Cantonese	yue	EMPTY 16 days ago
▸	UD Catalan	ca	PASS 16 days ago
▸	UD Chinese	zh	PASS 16 days ago
▸	UD Chinese-HK	zh hk	EMPTY 16 days ago
▸	UD Coptic	cop	PASS 9 days ago
▸	UD Croatian	hr	PASS 16 days ago
▸	UD Czech	cs	PASS 16 days ago
▸	UD Czech-CAC	cs cac	PASS 16 days ago
▸	UD Czech-CLTT	cs cltt	PASS 16 days ago
▸	UD Danish	da	PASS 16 days ago
▸	UD Dutch	nl	PASS 16 days ago
▸	UD Dutch-LassySmall	nl lassysmall	PASS 16 days ago
▸	UD English	en	PASS 16 days ago
▸	UD English-ESL	en esl	FAIL 16 days ago

Content Validation

- Runs automatically every time a treebank is updated
- Reports "suspicious" syntactic constructions
- Passing not compulsory at the moment
- Contributors: Is there anything odd-looking in my data?
- Release team: Overview of guideline adoption
- `http://universaldependencies.org/svalidation.html`



Aux chain

Auxiliary dependencies should not form a chain.

Search expression: `_ <aux (_ <aux _)`

Correct example:

1 Do you think that he will have left when we come ?

Incorrect example:

2 Do you think that he will have left when we come ?

Link to documentation

Hit overview	
UD_Basque	3 hits
UD_Galician	10 hits
UD_Italian	2 hits
UD_Japanese	1 hits
UD_Persian	2 hits
UD_Urdu	2341 hits

Flat is right-headed

Flat relations should be left-headed, not right.

Search expression: `_ <flat@R _`

Correct example:

3 Carl XVI Gustaf

Incorrect example:

4 Carl XVI Gustaf

Tools and Resources

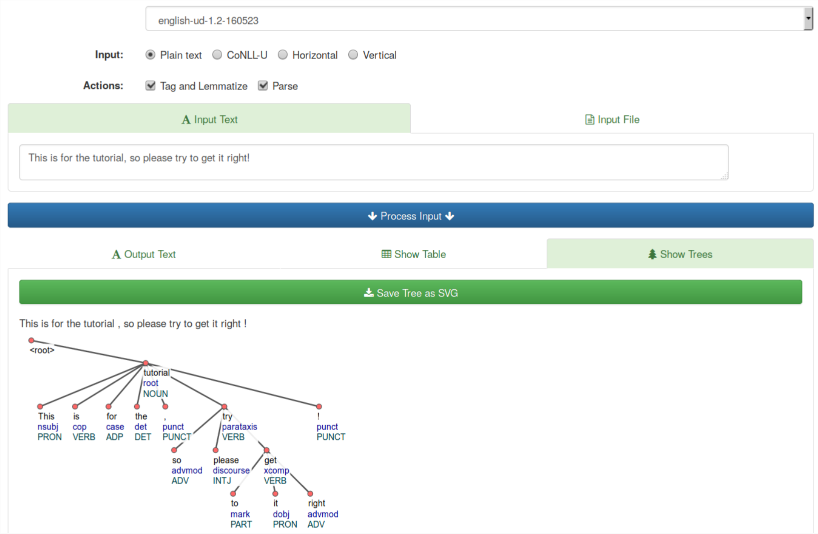
UD is **not just the treebanks**

- Parsers trained on UD data
- Large multilingual parsebanks
- Query tools for treebanks and parsebanks
- Libraries for handling CoNLL-U
- Tree visualization tools
- Annotation tools

Parsers

- UDPipe and SyntaxNet
- State-of-the-art parsers, free
- Full-stack parsers: raw text in - parses out
- Models trained on all of UD
- UDPipe — demo & Web API
- UDPipe Web API — get parsed text with a simple HTTP request

UDPipe



english-ud-1.2-160523

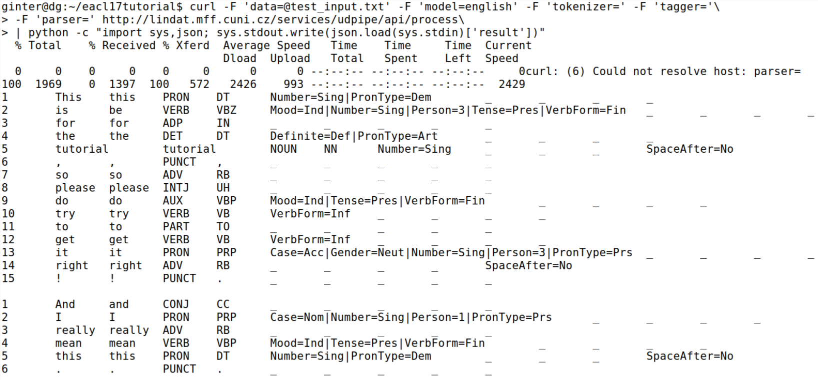
Input: ● Plain text ○ CoNLL-U ○ Horizontal ○ Vertical

Actions: ☑ Tag and Lemmatize ☑ Parse

This is for the tutorial, so please try to get it right!

This is for the tutorial , so please try to get it right !

UDPipe



```
ginter@dg:~/eacl17tutorial$ curl -F 'data=@test_input.txt' -F 'model=english' -F 'tokenizer=' -F 'tagger='\
> -F 'parser=' http://lindat.mff.cuni.cz/services/udpipe/api/process\
> | python -c "import sys,json; sys.stdout.write(json.load(sys.stdin)['result'])"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0curl: (6) Could not resolve host: parser=
100  1969    0  1397  100   572   2426    993 --:--:-- --:--:-- --:--:--  2429
1       This    this    PRON    DT      Number=Sing|PronType=Dem        _       _       _       _
2       is      be      VERB    VBZ     Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   _       _       _       _
3       for     for     ADP     IN      _       _       _       _       _
4       the     the     DET     DT      Definite=Def|PronType=Art       _       _       _       _
5       tutorial        tutorial        NOUN    NN      Number=Sing     _       _       _       SpaceAfter=No
6       ,       ,       PUNCT   ,       _       _       _       _       _
7       so      so      ADV     RB      _       _       _       _       _
8       please  please  INTJ    UH      _       _       _       _       _
9       do      do      AUX     VBP     Mood=Ind|Tense=Pres|VerbForm=Fin        _       _       _       _
10      try     try     VERB    VB      VerbForm=Inf    _       _       _       _
11      to      to      PART    TO      _       _       _       _       _
12      get     get     VERB    VB      VerbForm=Inf    _       _       _       _
13      it      it      PRON    PRP     Case=Acc|Gender=Neut|Number=Sing|Person=3|PronType=Prs  _       _       _       _
14      right   right   ADV     RB      _       _       _       _       SpaceAfter=No
15      !       !       PUNCT   .       _       _       _       _       _

1       And     and     CONJ    CC      _       _       _       _       _
2       I       I       PRON    PRP     Case=Nom|Number=Sing|Person=1|PronType=Prs      _       _       _       _
3       really  really  ADV     RB      _       _       _       _       _
4       mean    mean    VERB    VBP     Mood=Ind|Tense=Pres|VerbForm=Fin        _       _       _       _
5       this    this    PRON    DT      Number=Sing|PronType=Dem        _       _       _       SpaceAfter=No
6       .       .       PUNCT   .       _       _       _       _       _
```

Parsebanks

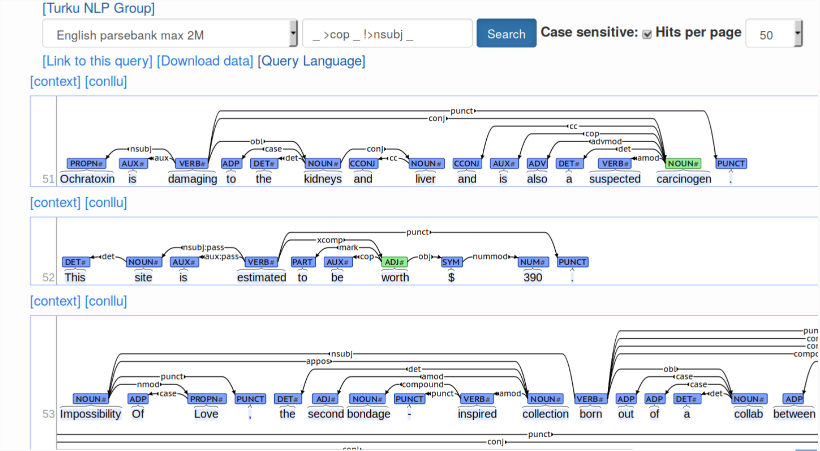
- UD-parsed corpora for 45 languages
- Data: CommonCrawl + Wiki + Perseus
- Parses: UDPipe
- Over 90B words total, 630GB zipped CoNLL-U files

Ancient Greek, Arabic, Basque, Bulgarian, Catalan, ChineseT, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Irish, Italian, Japanese, Kazakh, Korean, Latin, Latvian, Norwegian-Bokmaal, Norwegian-Nynorsk, Old Church Slavonic, Persian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Turkish, Ukrainian, Urdu, Uyghur, and Vietnamese

Syntactic Query

- dep_search `http://bionlp-www.utu.fi/dep_search`
- Relatively expressive query language, especially geared towards dependencies and rich morphology
- Indexed:
  - Latest UD official release
  - ‘dev‘ branches - reindexed on every push
  - Up to 2 million trees for every language from the UD Parsebanks
- Web and API access
- Used by some during annotation
- Also serves as content validation back-end

Syntactic Query



[Turku NLP Group]

English parsebank max 2M | _ >cop _ !>nsubj _ | Search | Case sensitive: ☑ Hits per page 50

[Link to this query] [Download data] [Query Language]

[context] [conllu]

51 Ochratoxin is damaging to the kidneys and liver and is also a suspected carcinogen .

[context] [conllu]

52 This site is estimated to be worth $ 390 .

[context] [conllu]

53 Impossibility Of Love , the second bondage - inspired collection born out of a collab between

PROPN AUX VERB ADP DET NOUN CCONJ NOUN CCONJ AUX ADV DET VERB NOUN PUNCT
nsubj aux obl case det conj cc punct conj cc cop advmod det amod
DET NOUN AUX VERB PART AUX ADJ SYM NUM PUNCT
det nsubj:pass aux:pass xcomp mark cop obj nummod punct
NOUN ADP PROPN PUNCT DET ADJ NOUN PUNCT VERB NOUN VERB ADP ADP DET NOUN ADP
nsubj appos punct nmod case det amod compound punct amod obl case case det pun cor cor compo punct conj conj

UDApi

(One slide about UDApi — will ask Martin — TODO)

Tree Visualization Tools

(todo)

Annotation Tools

- No official annotation tool yet
- Little in terms of what can be recommended
- `http://universaldependencies.org/tools.html`

Coffee Break