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 Now — Treebanks

How many?

- Languages: **50**
- Treebanks: **72**
- Trees: **642,000**
- Words: **12,400,000**

Can I use them?

- Creative Commons and GPL-like: **30**
- Creative Commons Non-Commercial: **42**

Where from?

- http://universaldependencies.org
- Official release preferred over GitHub
- Currently officially released: **70 treebanks**
- Twist: test sets currently withheld

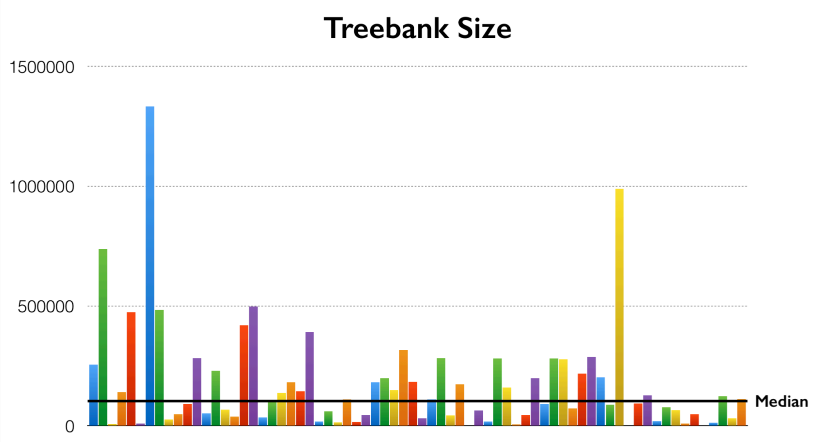
UD Treebanks Come in Many Flavors and Sizes

Annotation:

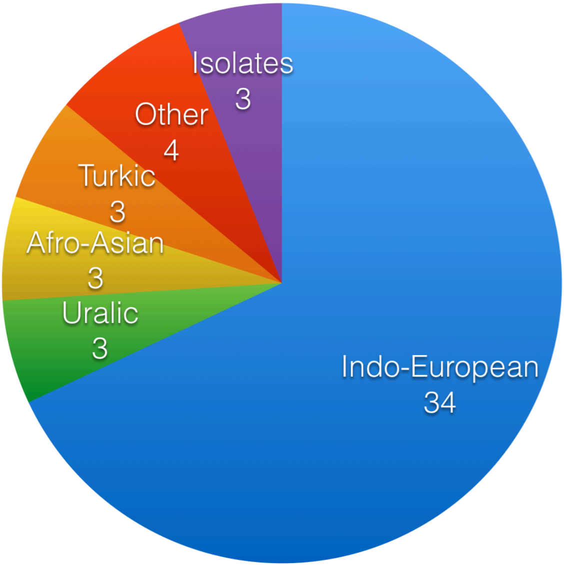
- POS and base dependency relations compulsory: 72 treebanks
- ...and additionally:
    - Forms + Features + Lemmas: 58
    - Forms - Features + Lemmas: 4
    - Forms - Features - Lemmas: 7
    - No Forms: 3 (Arabic-NYUAD, English-ESL, Japanese-KTC) — licensing

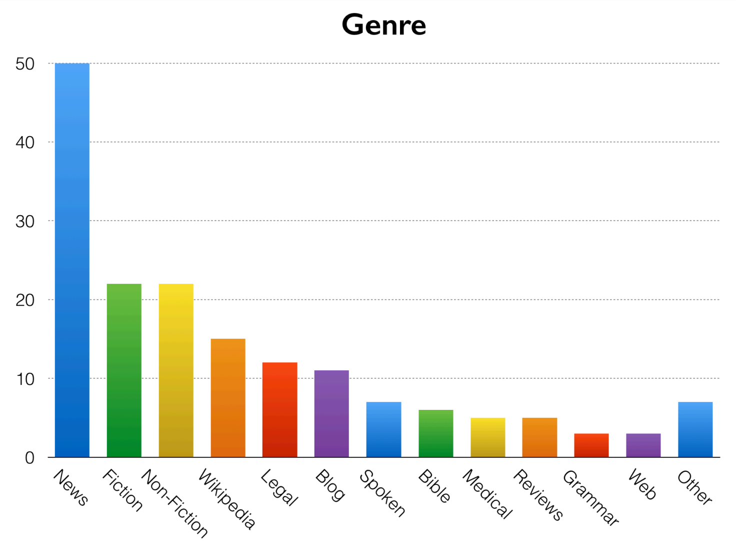
Size:

- Smallest: approx. 1000 words — Swedish Sign Language, Kazakh, Sanskrit
- Largest: Czech with 1.3M words, Russian with 980K words


Treebank Size
1500000
1000000
500000
0
Median

Language Family




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

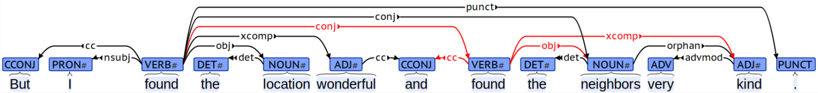
CoNLL-U Format

- 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:

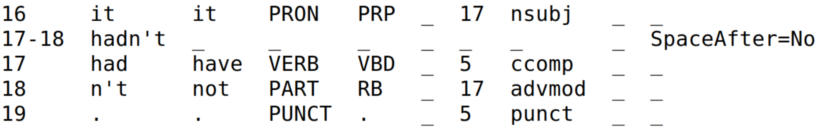
- Sentence-level metadata part of the format
- Explicit (and compulsory!) representation of the original text
- `DEPS` field encodes the enhanced dependencies (non-tree structure)
- `MISC` field allows arbitrary data stored for every word
- Empty nodes — only referred to from the enhanced representation
- Words as opposed to tokens

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    _   _
```

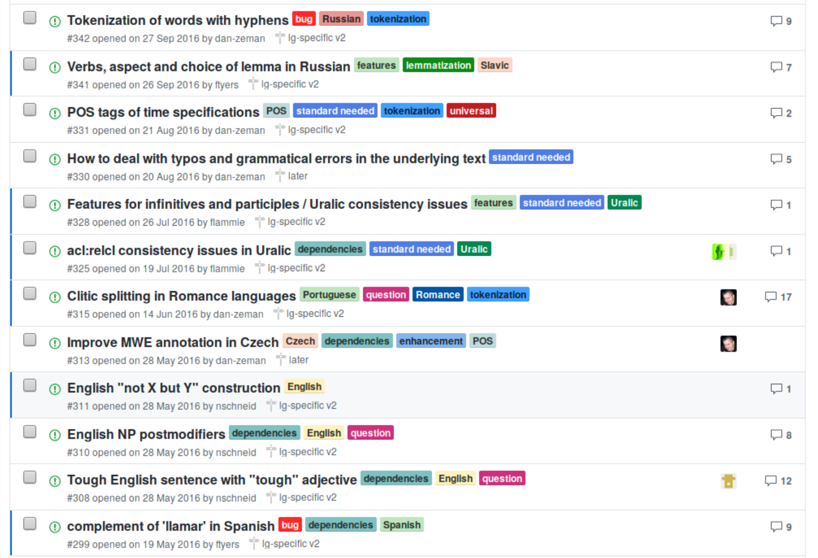
UD Infrastructure - Requirements

- 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

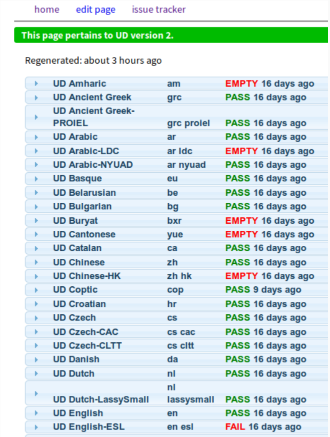
- One set of documentation for every language (not treebank)
- A GitHub repository holding mostly markdown pages
- Special care taken to make it easy to add tree visualizations and examples
- Stubs pre-generated when adding a new language
- 11,000+ commits from 80+ contributors
- Automatically regenerated 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
- Documentation system: `http://spyysalo.github.io/annodoc/`

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?



home edit page issue tracker

This page pertains to UD version 2.

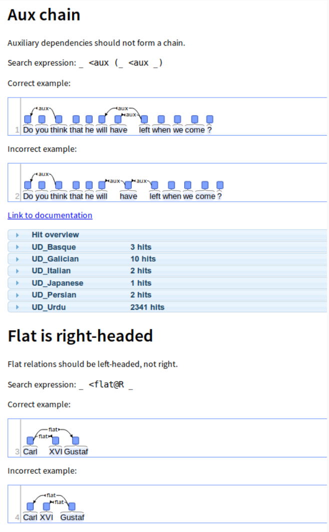
Regenerated: about 3 hours ago

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

```
http://universaldependencies.org/validation.html
```

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



Aux chain

Auxiliary dependencies should not form a chain.

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

Correct example:

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

Incorrect example:

aux
aux
aux
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:

Incorrect example:

http://universaldependencies.org/svalidation.html

Tools and Resources

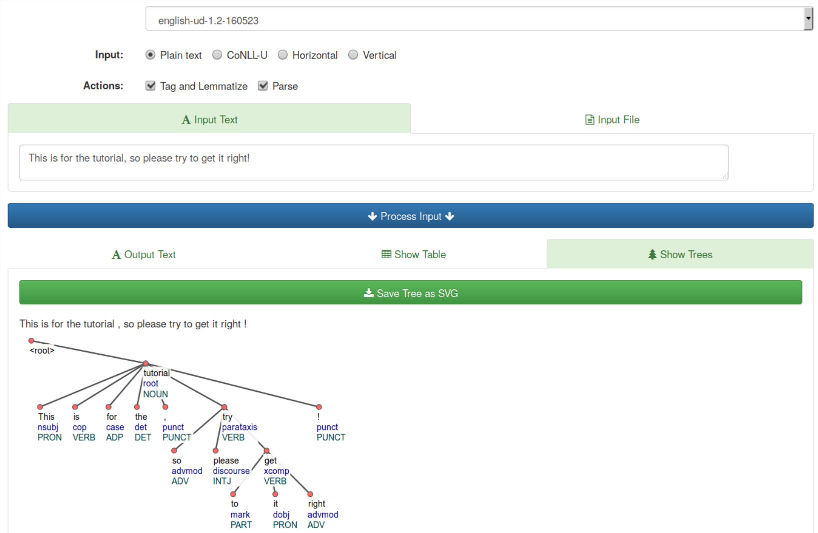
UD is **not just the treebanks**

- 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

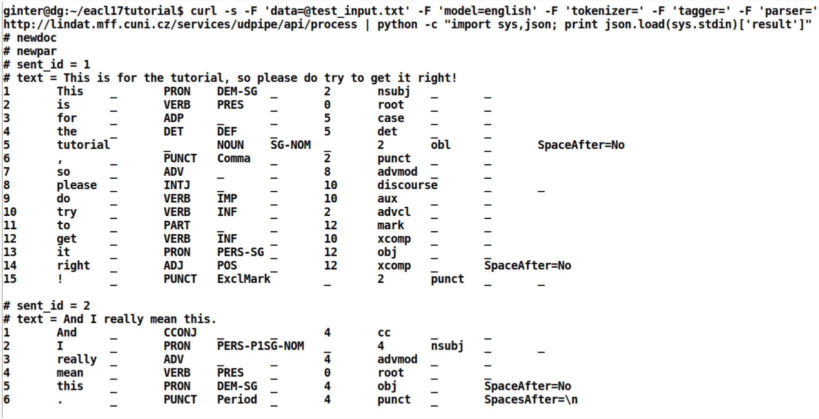
Input: ● Plain text ○ CoNLL-U ○ Horizontal ○ Vertical

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



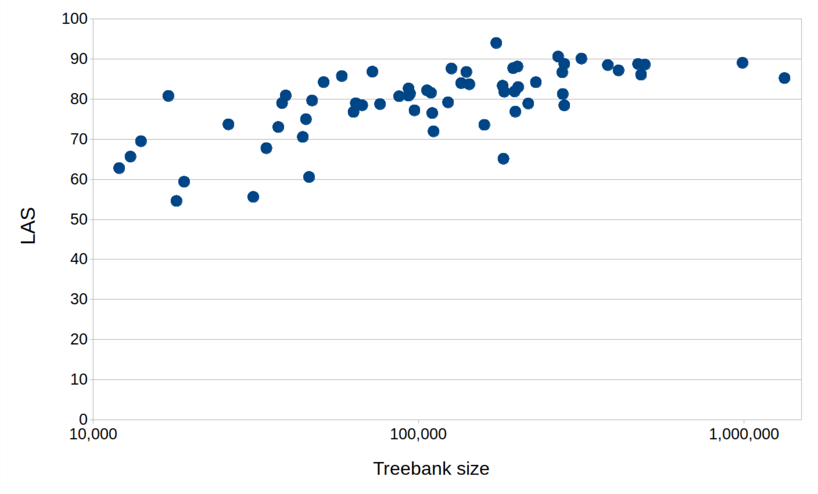
```
ginter@dg:~/eacl17tutorial$ curl -s -F 'data=@test_input.txt' -F 'model=english' -F 'tokenizer=' -F 'tagger=' -F 'parser='
http://lindat.mff.cuni.cz/services/udpipe/api/process | python -c "import sys,json; print json.load(sys.stdin)['result']"
# newdoc
# newpar
# sent_id = 1
# text = This is for the tutorial, so please do try to get it right!
1       This    _       PRON    DEM-SG  _       2       nsubj   _       _
2       is      _       VERB    PRES    _       0       root    _       _
3       for     _       ADP     _       _       5       case    _       _
4       the     _       DET     DEF     _       5       det     _       _
5       tutorial        _       NOUN    SG-NOM  _       2       obl     _       SpaceAfter=No
6       ,       _       PUNCT   Comma   _       2       punct   _       _
7       so      _       ADV     _       _       8       advmod  _       _
8       please  _       INTJ    _       _       10      discourse       _       _
9       do      _       VERB    IMP     _       10      aux     _       _
10      try     _       VERB    INF     _       2       advcl   _       _
11      to      _       PART    _       _       12      mark    _       _
12      get     _       VERB    INF     _       10      xcomp   _       _
13      it      _       PRON    PERS-SG _       12      obj     _       _
14      right   _       ADJ     POS     _       12      xcomp   _       SpaceAfter=No
15      !       _       PUNCT   ExclMark        _       2       punct   _       _

# sent_id = 2
# text = And I really mean this.
1       And     _       CCONJ   _       _       4       cc      _       _
2       I       _       PRON    PERS-P1SG-NOM   _       4       nsubj   _       _
3       really  _       ADV     _       _       4       advmod  _       _
4       mean    _       VERB    PRES    _       0       root    _       _
5       this    _       PRON    DEM-SG  _       4       obj     _       SpaceAfter=No
6       .       _       PUNCT   Period  _       4       punct   _       SpacesAfter=\n
```

ParseySaurus

- Major improvement upon *SyntaxNet's Parsey's cousins*
- Considerably improved models released mid-March 2017
- `http://tiny.cc/psaurus` — description
- `http://tiny.cc/psaurus-base` — numbers

ParseySaurus



Average=78% Median=81%

Parsebanks

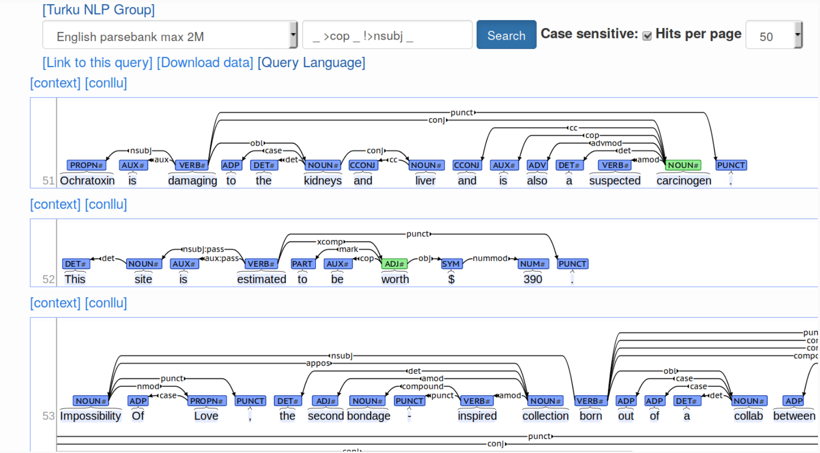
- UD-parsed corpora for 45 languages
- Data: CommonCrawl + Wiki + Perseus
- Parses: UDPipe
- Over 90B words total, 630GB zipped CoNLL-U files

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]

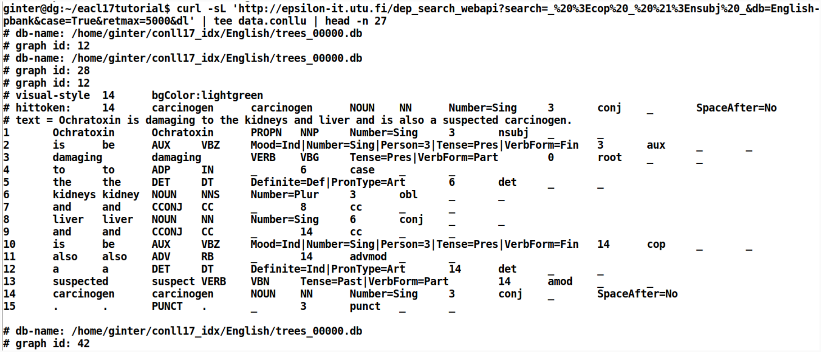
52 This site is estimated to be worth $ 390 .

[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
cop
advmod
amod
DET# NOUN# AUX# VERB# PART AUX# ADJ# SYM NUM# PUNCT
det
nsubj:pass
aux:pass
xcomp
mark
obj
nummod
NOUN# ADP PROPN# PUNCT DET# ADJ# NOUN# PUNCT VERB# NOUN# VERB# ADP ADP DET# NOUN# ADP
appos
nmod
compound
pun
cor
cor
compo

Syntactic Query



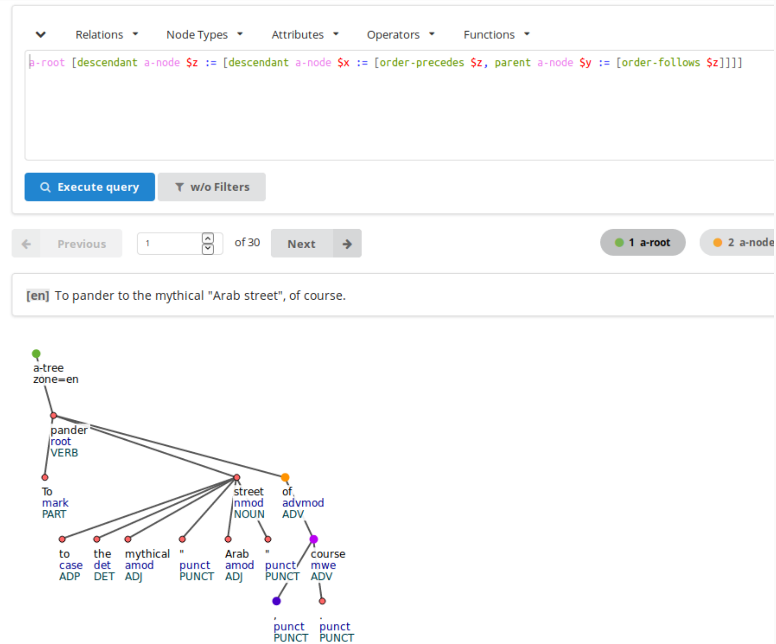
```
ginter@dg:~/eacl17tutorial$ curl -sL 'http://epsilon-it.utu.fi/dep_search_webapi?search=_%20%3Ecop%20_%20%21%3Ensubj%20_&db=English-
pbank&case=True&retmax=5000&dl' | tee data.conllu | head -n 27
# db-name: /home/ginter/conll17_idx/English/trees_00000.db
# graph id: 12
# db-name: /home/ginter/conll17_idx/English/trees_00000.db
# graph id: 28
# graph id: 12
# visual-style  14      bgColor:lightgreen
# hittoken:     14      carcinogen      carcinogen      NOUN    NN      Number=Sing     3       conj    _       SpaceAfter=No
# text = Ochratoxin is damaging to the kidneys and liver and is also a suspected carcinogen.
1       Ochratoxin      Ochratoxin      PROPN   NNP     Number=Sing     3       nsubj   _       _
2       is      be      AUX     VBZ     Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   3       aux     _       _
3       damaging        damaging        VERB    VBG     Tense=Pres|VerbForm=Part        0       root    _       _
4       to      to      ADP     IN      _       6       case    _       _
5       the     the     DET     DT      Definite=Def|PronType=Art       6       det     _       _
6       kidneys kidney  NOUN    NNS     Number=Plur     3       obl     _       _
7       and     and     CCONJ   CC      _       8       cc      _       _
8       liver   liver   NOUN    NN      Number=Sing     6       conj    _       _
9       and     and     CCONJ   CC      _       14      cc      _       _
10      is      be      AUX     VBZ     Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   14      cop     _       _
11      also    also    ADV     RB      _       14      advmod  _       _
12      a       a       DET     DT      Definite=Ind|PronType=Art       14      det     _       _
13      suspected       suspect VERB    VBN     Tense=Past|VerbForm=Part        14      amod    _       _
14      carcinogen      carcinogen      NOUN    NN      Number=Sing     3       conj    _       SpaceAfter=No
15      .       .       PUNCT   .       _       3       punct   _       _

# db-name: /home/ginter/conll17_idx/English/trees_00000.db
# graph id: 42
```

Syntactic Query

- PML Tree Query
- `http://lindat.mff.cuni.cz/services/pmltq/`
- A very expressive query language
- Indexed: official UD releases

Syntactic Query



```
a-root [descendant a-node $z := [descendant a-node $x := [order-precedes $z, parent a-node $y := [order-follows $z]]]]
```

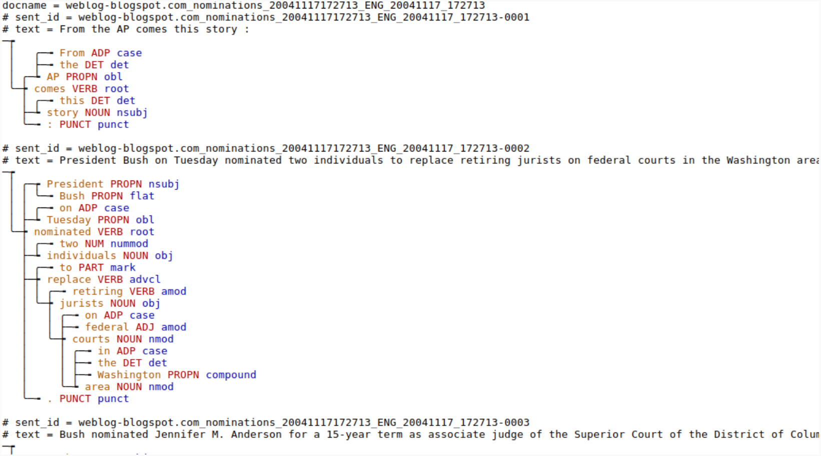
[en] To pander to the mythical "Arab street", of course.

Udapi

- A library and command line tool for processing UD data
  - **Python**, Java, Perl
- Format conversions
- Initial v1-v2 conversion
- Validation tests
- Evaluation, filtering, statistics
- Tree visualization
- `https://udapi.github.io`

Tree Visualization Tools

```
cat en-ud-dev.conllu | udapy -T | less -R
```



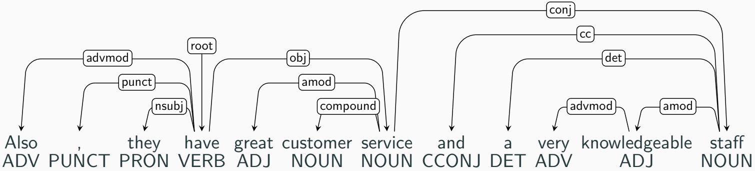
```
docname = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713
# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0001
# text = From the AP comes this story :
─┮
 │ ╭─╼ From ADP case
 │ ├─╼ the DET det
 │ ╭─┶ AP PROPN obl
 ╰─┾ comes VERB root
   │ ╭─╼ this DET det
   ├─┶ story NOUN nsubj
   ╰─╼ : PUNCT punct

# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0002
# text = President Bush on Tuesday nominated two individuals to replace retiring jurists on federal courts in the Washington area
─┮
 │ ╭─┮ President PROPN nsubj
 │ │ ╰─╼ Bush PROPN flat
 │ │ ╭─╼ on ADP case
 │ ├─┶ Tuesday PROPN obl
 ╰─┾ nominated VERB root
   │ ╭─╼ two NUM nummod
   ├─┶ individuals NOUN obj
   │ ╭─╼ to PART mark
   ├─┾ replace VERB advcl
   │ │ ╭─╼ retiring VERB amod
   │ ╰─┾ jurists NOUN obj
   │   │ ╭─╼ on ADP case
   │   │ ├─╼ federal ADJ amod
   │   ╰─┾ courts NOUN nmod
   │     │ ╭─╼ in ADP case
   │     │ ├─╼ the DET det
   │     │ ├─╼ Washington PROPN compound
   │     ╰─┶ area NOUN nmod
   ╰─╼ . PUNCT punct

# sent_id = weblog-blogspot.com_nominations_20041117172713_ENG_20041117_172713-0003
# text = Bush nominated Jennifer M. Anderson for a 15-year term as associate judge of the Superior Court of the District of Colum
─┮
```

Tree Visualization Tools

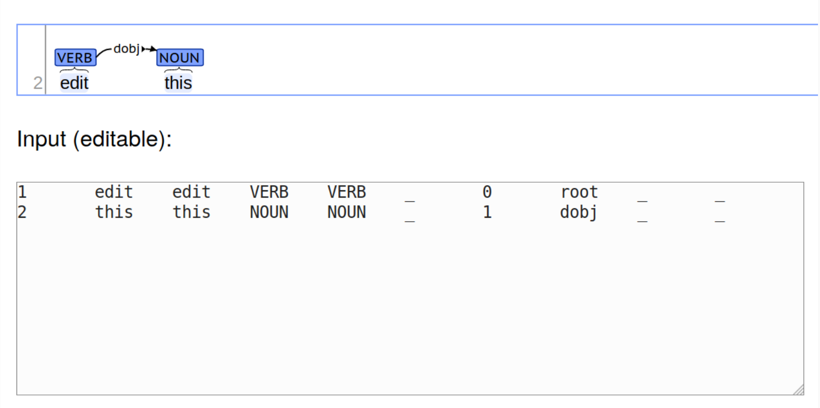
```
cat en-ud-dev.conllu | udapy write.Tikz
```



Tree Visualization Tools

http://spyysalo.github.io/conllu.js/

http://spyysalo.github.io/annodoc/sdparse.html



Input (editable):

```
1       edit     edit     VERB     VERB     _        0        root     _        _
2       this     this     NOUN     NOUN     _        1        dobj     _        _
```

Annotation Tools

- No official annotation tool (yet)
- A list of tools: `http://universaldependencies.org/tools.html`
- At present, none downright outstanding

Questions?