Tutorial on Universal Dependencies

Infrastructure, resources and tools for UD

Joakim Nivre\textsuperscript{1}  Daniel Zeman\textsuperscript{2}  Filip Ginter\textsuperscript{3}  Francis M. Tyers\textsuperscript{4,5}

\textsuperscript{1}Department of Linguistics and Philology, Uppsala University, Sweden
\textsuperscript{2}Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic
\textsuperscript{3}Department of Information Technology, University of Turku, Finland
\textsuperscript{4}Giela ja kultuvrra instituhtta, UiT Norgga árktalaš universitehta, Tromsø, Norway
\textsuperscript{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
Language Family

Indo-European: 34
Uralic: 3
Afro-Asian: 3
Turkic: 3
Other: 4
Isolates: 3
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  
10  very  very  ADV  RB  11  advmod  
11  kind  kind  ADJ  JJ  Degree=Pos  9  orphan  
12  .  .  PUNCT  .  3  punct  

# CoNLL-U Format — Tokens vs. Words

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>it</td>
<td>it</td>
<td>PRON</td>
<td>PRP</td>
<td>_</td>
<td>17</td>
<td>nsubj</td>
</tr>
<tr>
<td>17</td>
<td>hadn't</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>18</td>
<td>had</td>
<td>have</td>
<td>VERB</td>
<td>VBD</td>
<td>5</td>
<td>ccomp</td>
<td>_</td>
</tr>
<tr>
<td>18</td>
<td>n't</td>
<td>not</td>
<td>PART</td>
<td>RB</td>
<td>17</td>
<td>advmod</td>
<td>_</td>
</tr>
<tr>
<td>19</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>.</td>
<td>5</td>
<td>punct</td>
<td>_</td>
</tr>
</tbody>
</table>
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
- 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
<table>
<thead>
<tr>
<th>Issue</th>
<th>Description</th>
<th>Tags</th>
<th>Created</th>
<th>Reporter</th>
<th>Language</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>#342</td>
<td>Tokenization of words with hyphens</td>
<td>bug, Russian, tokenization</td>
<td>27 Sep 2016</td>
<td>dan-zeman</td>
<td>Ig-specific v2</td>
<td>9</td>
</tr>
<tr>
<td>#341</td>
<td>Verbs, aspect and choice of lemma in Russian</td>
<td>features, lemmatization, Slavic</td>
<td>26 Sep 2016</td>
<td>flyers</td>
<td>Ig-specific v2</td>
<td>7</td>
</tr>
<tr>
<td>#331</td>
<td>POS tags of time specifications</td>
<td>POS, standard needed, tokenization, universal</td>
<td>21 Aug 2016</td>
<td>dan-zeman</td>
<td>Ig-specific v2</td>
<td>2</td>
</tr>
<tr>
<td>#330</td>
<td>How to deal with typos and grammatical errors in the underlying text</td>
<td>standard needed</td>
<td>20 Aug 2016</td>
<td>dan-zeman</td>
<td>later</td>
<td>5</td>
</tr>
<tr>
<td>#328</td>
<td>Features for infinitives and participles / Uralic consistency issues</td>
<td>features, standard needed, Uralic</td>
<td>26 Jul 2016</td>
<td>flammie</td>
<td>Ig-specific v2</td>
<td>1</td>
</tr>
<tr>
<td>#325</td>
<td>acl:relcl consistency issues in Uralic</td>
<td>dependencies, standard needed, Uralic</td>
<td>19 Jul 2016</td>
<td>flammie</td>
<td>Ig-specific v2</td>
<td>1</td>
</tr>
<tr>
<td>#315</td>
<td>Clitic splitting in Romance languages</td>
<td>Portuguese, question, Romance, tokenization</td>
<td>14 Jun 2016</td>
<td>dan-zeman</td>
<td>Ig-specific v2</td>
<td>17</td>
</tr>
<tr>
<td>#313</td>
<td>Improve MWE annotation in Czech</td>
<td>Czech, dependencies, enhancement, POS</td>
<td>28 May 2016</td>
<td>dan-zeman</td>
<td>later</td>
<td></td>
</tr>
<tr>
<td>#311</td>
<td>English &quot;not X but Y&quot; construction</td>
<td>English</td>
<td>28 May 2016</td>
<td>nschnaid</td>
<td>Ig-specific v2</td>
<td>1</td>
</tr>
<tr>
<td>#310</td>
<td>English NP postmodifiers</td>
<td>dependencies, English, question</td>
<td>28 May 2016</td>
<td>nschnaid</td>
<td>Ig-specific v2</td>
<td>8</td>
</tr>
<tr>
<td>#308</td>
<td>Tough English sentence with &quot;tough&quot; adjective</td>
<td>dependencies, English, question</td>
<td>28 May 2016</td>
<td>nschnaid</td>
<td>Ig-specific v2</td>
<td>12</td>
</tr>
<tr>
<td>#299</td>
<td>complement of 'llamar' in Spanish</td>
<td>bug, dependencies, Spanish</td>
<td>19 May 2016</td>
<td>flyers</td>
<td>Ig-specific v2</td>
<td>9</td>
</tr>
</tbody>
</table>
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
• 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?

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

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

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

And I really mean this.
- 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
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, Chinese, 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
- 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

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
# visual-style 14
# hittoken: 14
carcinogen  carcogen  NOUN  NN  Number=Sing  3  conj   SpaceAfter=No
1  Ochratoxin  Ochratoxin  PROPN  NNP  Number=Sing  3  nsubj   
2  is  be  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
- A very expressive query language
- Indexed: official UD releases
Syntactic Query

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](https://udapi.github.io)
From ADP case
the DET det
AP PROPN obl
comes VERB root
det
this DET det
story NOUN nsubj
: PUNCT punct

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