Tutorial on Universal Dependencies

Adding a new language to UD

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

$^1$Department of Linguistics and Philology, Uppsala University, Sweden

$^2$Institute of Formal and Applied Linguistics, Charles University, Prague, Czechia

$^3$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
Two Scenarios

You want your language in UD

✔

Existing treebank
You have permission

No existing treebank
No permission/licence

↓

Treebank conversion

Building from scratch
Common Steps

First steps

- Get an account in Github
  - All development goes on here

Get in contact

- Ask someone from the release team to set up a module
- Get in contact with any other teams working on your language, or a related one
- Register for the mailing list *

* [http://stp.lingfil.uu.se/mailman/listinfo/ud](http://stp.lingfil.uu.se/mailman/listinfo/ud)

Release team
Linguistic Discussion

Linguistic discussion goes on under the *docs* module

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>How do I map these Chinese PoS tag to universal dependencies tag?</td>
<td>Chinese POS</td>
</tr>
<tr>
<td>NumType for numerals with decimal point (2.3)</td>
<td>features universal</td>
</tr>
<tr>
<td>goeswith lemma and upos</td>
<td>standard needed universal</td>
</tr>
<tr>
<td>Is upos=SYM deprel=punct allowed?</td>
<td>standard needed universal</td>
</tr>
<tr>
<td>Examples of function-word dependents in ellipsis</td>
<td></td>
</tr>
<tr>
<td>Original text in UD Czech and Arabic</td>
<td>Arabic bug Czech tokenization</td>
</tr>
<tr>
<td>Clausal complements with different subject in Finnic</td>
<td>Finnish Uralic</td>
</tr>
<tr>
<td>Abbreviation at the end of sentence</td>
<td>standard needed tokenization universal</td>
</tr>
</tbody>
</table>
Linguistic Discussion

Annotation guidelines are discussed with examples

Apposition in Kurdish #334

MemduhG opened this issue on 30 Aug 2016 · 2 comments

MemduhG commented on 30 Aug 2016

We came across a sentence the annotation of which seems to be tricky, with regards to subject, apposition and their relation.

Hîynê ez û xweha xwe Cûyû; em her du ji du-sall bûn
So I and my sister Julia; we both were two-year-olds.

Either ez û xweha xwe Cûyû “I and my sister Julia” or em her du “we both” should be the subject, with the other one attached as appos. I am not sure which one should be the subject, as the part that would usually be considered an apposition is given before rather than after what would be the subject.

dan-zeman commented on 30 Aug 2016

This looks like a perfect example of the dislocated relation, http://universaldependencies.org/u/dep/dislocated.html

I would make em her du the subject and ez û xweha xwe Cûyû a dislocated dependent, both attached to the main predicate.

dan-zeman added dependencies question labels on 30 Aug 2016

dan-zeman added this to the Ig-specific v2 milestone on 30 Aug 2016

ftyers commented on 1 Sep 2016

This looks good, thanks! :)

ftyers closed this on 1 Sep 2016
Linguistic Discussion

Annotation guidelines are discussed with examples

MemduhG commented on 30 Aug 2016

We came across a sentence the annotation of which seems to be tricky, with regards to subject, apposition and their relation.

Hingê ez ù xweha xwe Cûlya; em her du ji du-sali bûn
So I and my sister Julia; we both were two-year-olds.

Either ez ù xweha xwe Cûlya "I and my sister Julia" or em her du "we both" should be the subject, with the other one attached as appos. I am not sure which one should be the subject, as the part that would usually be considered an apposition is given before rather than after what would be the subject.

dan-zeman commented on 30 Aug 2016

This looks like a perfect example of the dislocated relation, http://universaldependencies.org/u/dep/dislocated.html

I would make em her du the subject and ez ù xweha xwe Cûlya a dislocated dependent, both attached to the main predicate.

dan-zeman added dependencies question labels on 30 Aug 2016

dan-zeman added this to the Ig-specific v2 milestone on 30 Aug 2016
**case: case marking**

The dependency type *case* is used for the postposition in postpositional phrases. The head of a postpositional phrase is the nominal, not the postposition, so as to analyse postpositional phrases similarly to nominal modifiers without a postposition. (Such nominal modifiers are frequent in Kazakh, as cases are often used for the same purpose as postpositions.) To the same end, the type *case* is used in combination with the type *nmod*, which is also used for nominal modifiers when no adposition is present (see *nmod*).

Note that case is not used with auxiliary nouns (sometimes called “postpositions”) in the form of N¹.gen N².poss.case, for those *nmod* should be used (following treatment in English of prepositional constructions like “in front of”).

- Documentation is written in Markdown and converted to HTML
- Not mandatory, but highly recommended
- Document as you write conversion rules/the annotation scheme
NEW! Documentation by language family

Language-Family Documentation

Slavic

- Introduction
- Tokenization
- Morphology
  - General principles
  - POS tags (single document)
  - Features (single document)
- Syntax
  - General principles
  - Specific constructions
  - Relations (single document)

Quick links

- Pronominal words

UD Treebanks

<table>
<thead>
<tr>
<th>Language</th>
<th>Size</th>
<th>POS</th>
<th>Morphology</th>
<th>Syntax</th>
<th>UD Treebank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belarusian</td>
<td>8K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Bulgarian</td>
<td>140K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Croatian</td>
<td>183K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>1,330K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Czech-CAC</td>
<td>482K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Czech-CLTT</td>
<td>26K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Old Church Slavonic</td>
<td>47K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td>72K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>97K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Russian-SynTagRus</td>
<td>988K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Serbian</td>
<td>-</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Slovak</td>
<td>93K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Slovenian</td>
<td>126K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Slovenian-SST</td>
<td>19K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
<tr>
<td>Ukraininan</td>
<td>12K</td>
<td>🟡🟠</td>
<td>🟡</td>
<td>🟡</td>
<td></td>
</tr>
</tbody>
</table>
Treebank conversion

Case study: Turkish
Original Treebank

METU-Sabancı Treebank

- Started in 2003
- Converted to CoNLL format for the 2006 shared task

İTÜ-METU-Sabancı Treebank

- 2016 reannotation of the METU-Sabancı treebank
- Morphology edited, dependencies from scratch
- 60k tokens in 2 months with 5 annotators (Sulubacak et al., 2016)
Co-ordination

A loosely co-ordinated effort between:

- Çağrı Çöltekin (U. Tübingen)
- A team from İTÜ
  - Umut Sulubacak
  - Memduh Gökırmak
  - Gülşen Eryiğit
- Hüner Kaşıkara (U. Boğaziçi)
- Joakim Nivre
- Francis Tyers

Kickoff meeting in Uppsala (November, 2015)
Conversion Process

Method:

- Go through reference grammar (Göksel & Kerslake, 2011)
- Document phenomena
- Convert treebank according to documentation

Tools:

- 6,000 lines of Java
  - Morphological synthesis
  - Collapse derivations
  - Remove multiwords
  - Distinguish clause from non-clause
yanlışlanan kuramın doğurganlığı burada yatar.

Disputed theory's fertility here lies.
Yanlışlanan Yanlışlan VERB _ Tense=Pres|VerbForm=Part|... 2 acl
2 kuramın kuram NOUN _ Case=Gen|... 3 nmod:poss
3 doğurganlığı doğurganlık NOUN _ Case=Nom|... 5 nsubj
4 burada bura NOUN _ Case=Loc|... 5 obl
5 yatar yat VERB _ Tense=Aor|VerbForm=Fin|... 0 root
6 . . PUNCT _ _ 5 punct
From scratch

Case study: Kazakh
Four questions

- Which annotation scheme?
- Where to get the data?
- How much data?
- How long will it take?
Which Annotation Scheme?

- **Non-UD:**
  - Perhaps there are existing treebanks for your language and you want to retain compatibility

- **UD:**
  - No need for any special conversion
  - ...at least until v3.0 ;)

- **Mixed:**
  - Follow UD guidelines
  - Add information where you think it is useful
  - ...providing it is easily convertible

**At the end of the day:** Do what is best for your language and your application
Where To Get The Data?

Free text:

- Plenty of options:
  - WikiMedia projects: Wikipedia, Wikinews, ...
  - Public domain texts (varies by country)
    - Out of copyright (e.g. old literature, folktales)
    - Laws/state administrative texts

Non-free text:

- Contact copyright holders early on
How Much Data?

- No minimum size
  - Smallest treebank: 1K tokens
  - Biggest treebank: 1.3M tokens
- CoNLL-2006, smallest treebank: 29K tokens

« You can actually train a parser and get over 50% accuracy for many languages with just about 100 sentences. » — Dan Zeman
How Long Will It Take?

- How long is a piece of string?
- Some approximate numbers:

<table>
<thead>
<tr>
<th>Language</th>
<th>Annotators</th>
<th>Tokens</th>
<th>Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kazakh</td>
<td>2</td>
<td>4,500</td>
<td>1</td>
</tr>
<tr>
<td>Buryat</td>
<td>1</td>
<td>10,000</td>
<td>3</td>
</tr>
<tr>
<td>Irish</td>
<td>1</td>
<td>23,600</td>
<td>12</td>
</tr>
</tbody>
</table>

In all the above cases, annotation guidelines were developed from scratch by people with no prior exposure to UD.
How We Made A Kazakh Treebank

Two people:

- Francis Tyers: Computational linguist, interests in Turkic languages and morphosyntax
- Jonathan North Washington: Phonologist, interests in Turkic languages, fluent speaker of Kazakh and Kyrgyz

One month:

- Summer holiday in Bloomington, Indiana

Resources:

- Morphological analyser and constraint grammar

The whole thing would have been impossible without the UD project.

- Guidelines were straightforward to apply
- Community was exceptionally helpful and welcoming
Annotation Process
Summary
Summary

What you need to do

- Join the project
- Start annotating or converting
- Ask if you get stuck!

Can’t wait to get started?

- Come and talk to us!
Questions?