Tutorial on Universal Dependencies

Adding a new language to UD

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

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

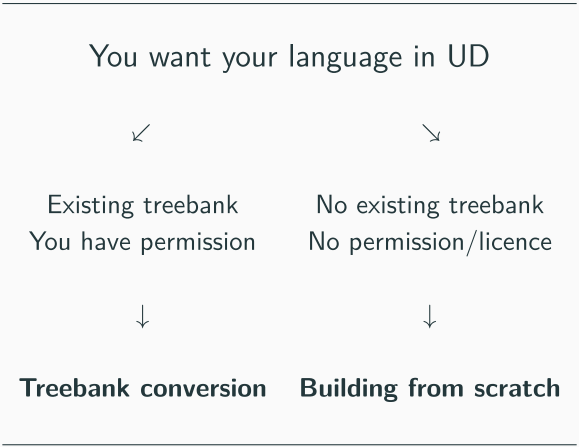
[2]Institute of Formal and Applied Linguistics, Charles University, Prague, Czechia

[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

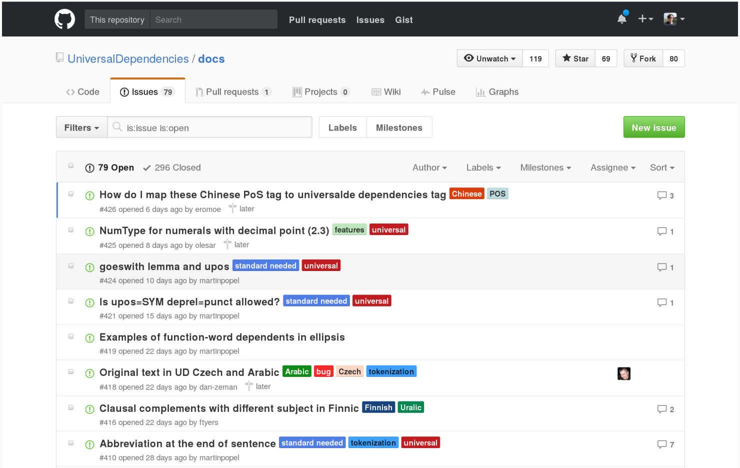
- Ask someone from the release team to set up a module
- Get in contact with any other teams working on your language, or a related one
- Register for the mailing list *

* `http://stp.lingfil.uu.se/mailman/listinfo/ud`

Release team

Linguistic Discussion

Linguistic discussion goes on under the *docs* module



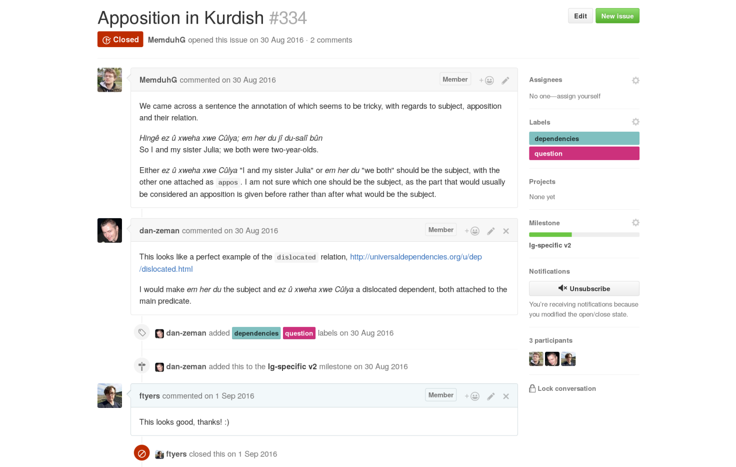
is:issue is:open

79 Open ✓ 296 Closed

- How do I map these Chinese PoS tag to universalde dependencies tag · Chinese · POS · 3
  #426 opened 6 days ago by eromoe · later
- NumType for numerals with decimal point (2.3) · features · universal · 1
  #425 opened 8 days ago by olesar · later
- goeswith lemma and upos · standard needed · universal · 1
  #424 opened 10 days ago by martinpopel
- Is upos=SYM deprel=punct allowed? · standard needed · universal · 1
  #421 opened 15 days ago by martinpopel
- Examples of function-word dependents in ellipsis
  #419 opened 22 days ago by martinpopel
- Original text in UD Czech and Arabic · Arabic · bug · Czech · tokenization
  #418 opened 22 days ago by dan-zeman · later
- Clausal complements with different subject in Finnic · Finnish · Uralic · 2
  #416 opened 22 days ago by ftyers
- Abbreviation at the end of sentence · standard needed · tokenization · universal · 7
  #410 opened 28 days ago by martinpopel

Linguistic Discussion

Annotation guidelines are discussed with examples



Apposition in Kurdish #334

Closed **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.

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

Either *ez û 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 û 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 **lg-specific v2** milestone on 30 Aug 2016

ftyers commented on 1 Sep 2016

This looks good, thanks! :)

ftyers closed this on 1 Sep 2016

Assignees

No one—assign yourself

Labels

dependencies

question

Projects

None yet

Milestone

lg-specific v2

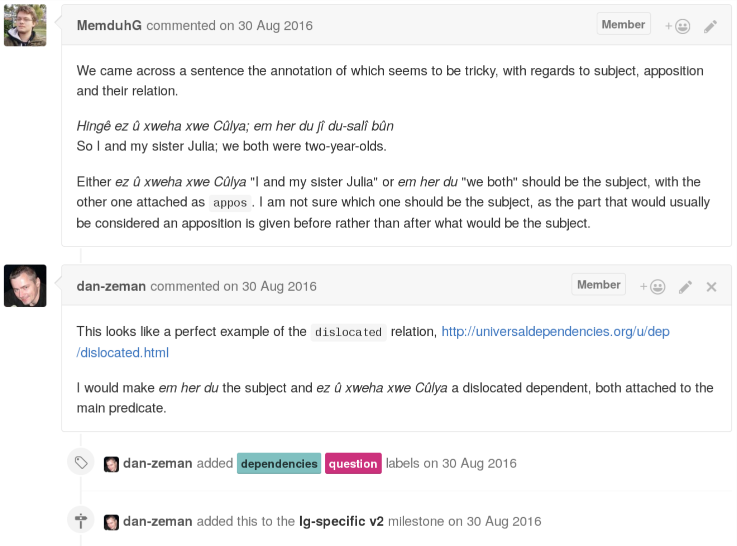
Notifications

You're receiving notifications because you modified the open/close state.

3 participants

Linguistic Discussion

Annotation guidelines are discussed with examples



MemduhG commented on 30 Aug 2016 Member

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

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

Either *ez û 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 Member

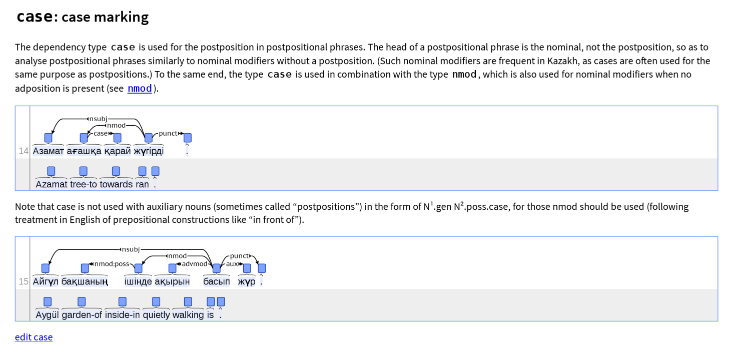
This looks like a perfect example of the `dislocated` relation, http://universaldependencies.org/u/dep/dislocated.html

I would make *em her du* the subject and *ez û xweha xwe 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 **lg-specific v2** milestone on 30 Aug 2016

Write Documentation



`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`).

14 Азамат ағашқа қарай жүгірді .

Azamat tree-to towards ran .

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”).

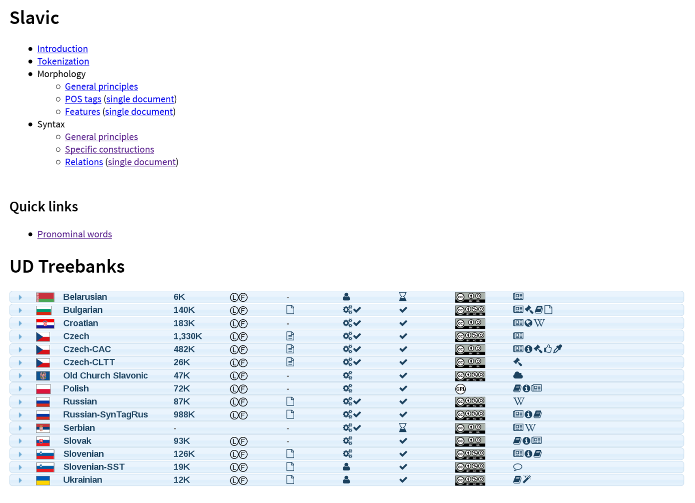
15 Айгүл бақшаның ішінде ақырын басып жүр .

Aygül garden-of inside-in quietly walking is .

edit case

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

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

Belarusian	6K	ⓁⒻ	-	👤	⌛	CC BY SA	📰
Bulgarian	140K	ⓁⒻ	🗋	⚙✓	✓	CC BY NC SA	📰🔨📘🗋
Croatian	183K	ⓁⒻ	-	⚙✓	✓	CC BY SA	📰🌍W
Czech	1,330K	ⓁⒻ	🗎	⚙✓	✓	CC BY NC SA	📰
Czech-CAC	482K	ⓁⒻ	🗎	⚙✓	✓	CC BY SA	📰ℹ🔨👍🖉
Czech-CLTT	26K	ⓁⒻ	🗎	⚙✓	✓	CC BY SA	🔨
Old Church Slavonic	47K	ⓁⒻ	-	⚙	✓	CC BY NC SA	☁
Polish	72K	ⓁⒻ	-	⚙	✓	GPL	📘ℹ📰
Russian	87K	ⓁⒻ	🗋	⚙✓	✓	CC BY NC SA	W
Russian-SynTagRus	988K	ⓁⒻ	🗋	⚙✓	✓	CC BY NC SA	📰ℹ📘
Serbian	-		-	⚙✓	⌛	CC BY SA	📰W
Slovak	93K	ⓁⒻ	-	⚙	✓	CC BY SA	📘ℹ📰
Slovenian	126K	ⓁⒻ	🗋	⚙	✓	CC BY NC SA	📰ℹ📘
Slovenian-SST	19K	ⓁⒻ	🗋	👤	✓	CC BY NC SA	💬
Ukrainian	12K	ⓁⒻ	🗋	👤	✓	CC BY NC SA	📘🪄

NEW! Documentation by language family

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 editted, 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

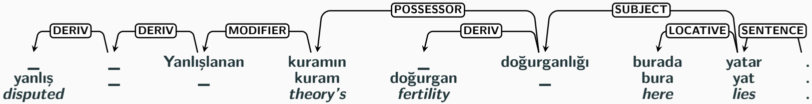
Method:

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

Tools:

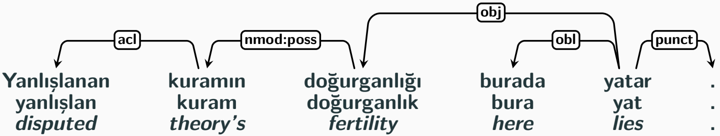
- 6,000 lines of Java
  - Morphological synthesis
  - Collapse derivations
  - Remove multiwords
  - Distinguish clause from non-clause

Original Treebank



1	_	yanlış	Adj	_	_	2	DERIV
2	_	_	Verb	_	Acquire|Pos	3	DERIV
3	Yanlışlanan	_	_	APresPart	_	4	MODIFIER
4	kuramın	kuram	Noun	_	A3sg|Pnon|Gen	6	POSSESSOR
5	_	doğurgan	_	Adj	_	6	DERIV
6	doğurganlığı	_	_	Noun	Ness|A3sg|P3sg|Nom	8	SUBJECT
7	burada	bura	Noun	_	A3sg|Pnon|Loc	8	LOCATIVE
8	yatar	yat	Verb	_	Pos|Aor|A3sg	9	SENTENCE
9	.	.	Punc	_	_	0	ROOT

UD Turkish



<table>
<tr><td>1</td><td>Yanlışlanan</td><td>Yanlışlan</td><td>VERB</td><td>_</td><td>Tense=Pres|VerbForm=Part|...</td><td>2</td><td>acl</td></tr>
<tr><td>2</td><td>kuramın</td><td>kuram</td><td>NOUN</td><td>_</td><td>Case=Gen|...</td><td>3</td><td>nmod:poss</td></tr>
<tr><td>3</td><td>doğurganlığı</td><td>doğurganlık</td><td>NOUN</td><td>_</td><td>Case=Nom|...</td><td>5</td><td>nsubj</td></tr>
<tr><td>4</td><td>burada</td><td>bura</td><td>NOUN</td><td>_</td><td>Case=Loc|...</td><td>5</td><td>obl</td></tr>
<tr><td>5</td><td>yatar</td><td>yat</td><td>VERB</td><td>_</td><td>Tense=Aor|VerbForm=Fin|...</td><td>0</td><td>root</td></tr>
<tr><td>6</td><td>.</td><td>.</td><td>PUNCT</td><td>_</td><td>_</td><td>5</td><td>punct</td></tr>
</table>

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:

Language	Annotators	Tokens	Months
Kazakh	2	4,500	1
Buryat	1	10,000	3
Irish	1	23,600	12

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

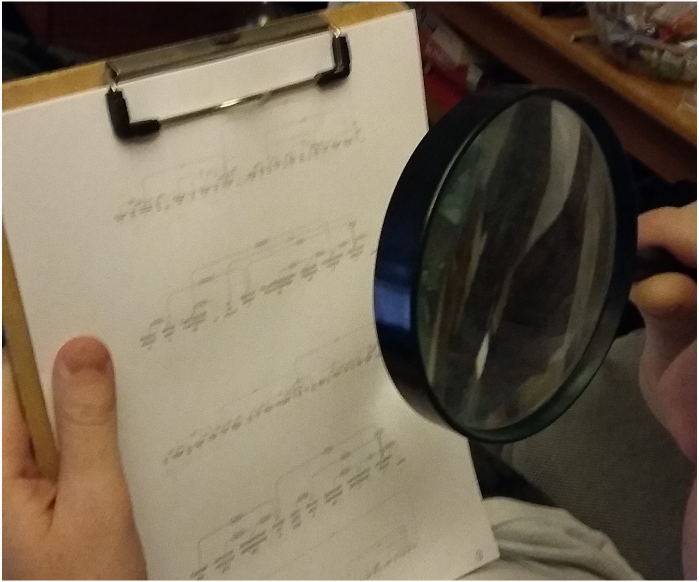
Resources:

- 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?