

## UniDive Training School Course 3.6

# Documentation and Discussion on Git



Daniel Zeman, Agata Savary

[zeman@ufal.mff.cuni.cz](mailto:zeman@ufal.mff.cuni.cz)

<https://unidive.lisn.upsaclay.fr/>

# Outline

- 1 Documenting a UD language on GitHub
- 2 Documenting a UD treebank in README
- 3 Documenting a PARSEME corpus in README
- 4 UD GitHub issues
- 5 PARSEME GitLab issues

# Documenting a UD language

- **One** set of language-specific guidelines **per language** (not per treebank!)
  - ▶ Multiple treebanks/teams? **Talk to each other!**
  - ▶ Not possible? Document differences, do not make them invalid (at least initially).
  - ▶ Long term:
    - ★ Consensus in UD / language family group / several teams of one language?
    - ★ ⇒ **New rule in the validator** even if some data providers are not reachable
    - ★ ⇒ Their treebanks become “LEGACY”, must be fixed in 4 years, can be adopted by new maintainer

# Documenting a UD language

- **Mandatory:** One page summary of language-specific guidelines
  - ▶ E.g. <https://universaldependencies.org/cs/index.html>
  - ▶ Template available
  - ▶ Certain minimum size required
  - ▶ See links from <https://universaldependencies.org/guidelines.html>
  - ▶ Ideally all the other language-specific pages should be linked from this one

# Documenting a UD language

- **Mandatory:** One page summary of language-specific guidelines
  - ▶ E.g. <https://universaldependencies.org/cs/index.html>
  - ▶ Template available
  - ▶ Certain minimum size required
  - ▶ See links from <https://universaldependencies.org/guidelines.html>
  - ▶ Ideally all the other language-specific pages should be linked from this one
- **Mandatory:**
  - ▶ A page for each language-specific **feature**
    - ★ That is: Feature not already documented globally
    - ★ Or feature that is documented globally but we need it with **extra values**
  - ▶ A page for each dependency relation **subtype**
    - ★ Unless already documented globally

# Documenting a UD language

- **Mandatory:** One page summary of language-specific guidelines
  - ▶ E.g. <https://universaldependencies.org/cs/index.html>
  - ▶ Template available
  - ▶ Certain minimum size required
  - ▶ See links from <https://universaldependencies.org/guidelines.html>
  - ▶ Ideally all the other language-specific pages should be linked from this one
- **Mandatory:**
  - ▶ A page for each language-specific **feature**
    - ★ That is: Feature not already documented globally
    - ★ Or feature that is documented globally but we need it with **extra values**
  - ▶ A page for each dependency relation **subtype**
    - ★ Unless already documented globally
- Optional: Language-specific version of a globally documented UPOS / feature / relation
- Optional: Other language-specific pages as needed

# Where to Edit Docs?

- `universaldependencies.org` automatically generated from the **docs** repository (branch **pages-source**) on GitHub
- Treebank contributors have push access there
- Website typically regenerated in 2–5 minutes after push
- *Unfortunately this week it is broken and we are waiting for the maintainer to fix it*
  - ▶ *Fortunately the mandatory parts are checked directly in the docs repo, so no problem for validation*

# Where to Edit Docs?

- `universaldependencies.org` automatically generated from the **docs** repository (branch **pages-source**) on GitHub
- Treebank contributors have push access there
- Website typically regenerated in 2–5 minutes after push
- *Unfortunately this week it is broken and we are waiting for the maintainer to fix it*
  - ▶ *Fortunately the mandatory parts are checked directly in the docs repo, so no problem for validation*
- Etiquette:
  - ▶ Do not directly edit universal guidelines (create pull requests if necessary)
  - ▶ Same for languages you do not work on
  - ▶ Within your language(s), edit directly if there is consensus
  - ▶ **Be careful:** Some parts must be **machine-readable** (recognizable by the validator)



# Documenting a UD language

- Optional: Language-specific version of a globally documented UPOS / feature / relation
- **Never ever create** files named **AUX.md** or **aux.md**!
  - ▶ Illegal on some operating systems
  - ▶ ⇒ people could not clone the docs repo there!
  - ▶ Instead, **AUX\_.md** (aux\_.md) with redirect directive inside

# How to Edit Docs?

- **Overview:** <https://universaldependencies.org/contributing.html>
- **Markdown syntax:**  
<https://daringfireball.net/projects/markdown/syntax>
- **Style guidelines:** <https://universaldependencies.org/contributing.html#style-guidelines>
- **Examples with dependency trees:**  
<https://universaldependencies.org/visualization.html>
- **Lang-spec mandatory:** [https://universaldependencies.org/contributing\\_language\\_specific.html](https://universaldependencies.org/contributing_language_specific.html)
  - ▶ Copy an existing page
  - ▶ Pay attention to **headings** of feature values and to formatting of **examples**

# Registering Features, Relations, Auxiliaries

- Only documented features can be registered
- Only registered features will be accepted by the validator
  - ▶ **Error in documentation will block the feature!**
- **Features:** `https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_feature.pl`
- **Dependency relations:** `https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_deprel.pl`
- **Auxiliaries and copulas:** `https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_auxiliary.pl`
- **Enhanced case-marked relations:** `https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_edeprel.pl`

# Trebank Specific Pages

- Automatically generated at release time
  - ▶ Do not edit them, it will be overwritten
  - ▶ Edit trebank README instead
- Trebank hub page generated from **README.md** and from the data
- Thousands of statistics and examples generated from data
- Comparison of multiple trebanks of one language

# Documenting a UD Treebank in README

- Partially prescribed structure
- Mandatory metadata at the end
- See the **Release Checklist**
  - ▶ `https://universaldependencies.org/release\_checklist.html#the-readme-file`

# Documenting a PARSEME corpus in README

- Each PARSEME corpus contains a `README.md` file, maintained by the **Language Leaders**
- Compulsory items in `README.md`
  - ▶ Source corpora and text genres
  - ▶ Format
    - ★ For each CUPT column: was the annotation automatic or manual?
    - ★ For the UPOS, FEATS and DEPREL columns: reference to the UD version (e.g. 2.11) or the UDPipe model (e.g. `greek-gdt-ud-2.5-191206`)
  - ▶ License
  - ▶ Authors and contact to the Language Leader
  - ▶ Paper to cite when using the treebank
  - ▶ Known issues and future work
  - ▶ **Change log** – changes introduced in each version with respect to the previous version
- Example: [https://gitlab.com/parseme/parseme\\_corpus\\_pt](https://gitlab.com/parseme/parseme_corpus_pt)

## PARSEME Gitlab issues

- The PARSEME guidelines were conceived to be universal, that is, all languages should be able to use the **same guidelines**
- Annotator teams should try to **follow them strictly**, regardless of their previous background or a particular linguistic theory; there should be as little personal interpretation as possible
- Some languages may have phenomena that are not correctly or **not** at all **covered in the guidelines**
- Doubts may arise about **unclear tests**, borderline cases, etc.
- Language Leaders and annotators can participate in enhancing the guidelines
- The PARSEME guidelines **Gitlab issues** are meant for this (note the **labels**):  
<https://gitlab.com/parseme/sharedtask-guidelines/-/issues>

# PARSEME Gitlab issues

https://gitlab.com/parseme/sharedtask-guidelines/-/issues

Settings Filmy Wakacje Przepisy Idiomy Dom Taniec Zdrowie Muzyka Prezenty Ekologia

parseme / sharedtask-guidelines / Issues

Open 41 Closed 75 All 116

Search or filter results...

- how to update guidelines regarding versioning #120 · created 8 hours ago by Furkan Akkurt
- Syntactic tests NMWE.9 to NMWE.11 are unclear and too detailed #119 · created 1 month ago by Savary **Nominal MWE**
- DIST test too vague** #118 · created 2 months ago by Savary **All languages**
- Order of tests #117 · created 3 months ago by Verginica M. **Nominal MWE**

Project: sharedtask-guidelines

- Pinned
- Manage
- Plan
- Issues 41
- Issue boards
- Milestones
- Code
- Build

Open DIST test too vague

- Savary changed the description 2 months ago

Savary @agata.savary · 4 weeks ago  
Or maybe should we rather say:

Can you replace the candidate expression with a **number** of single words, taken from a **relatively large semantic class**, which are clearly nouns, so that the sentence remains grammatical? If so, the candidate expression has the nominal distribution. Same for adjectives, adverbs, etc.

- Savary mentioned in issue #119 4 weeks ago

vgiouli1 @vgiouli1 · 4 weeks ago  
I find the notion of a "frequent" single-word synonym somewhat tricky. Same for "meaning shift of the sentence ([that] is predictable from the replacement)". I prefer the second wording.

Francis Bond @fcbond · 4 weeks ago  
Maybe we can add something more from the external point of view:

Does this expression typically appear in positions normally filled by nouns (or noun phrases), that is, subject, object or other complement of a verb, complement of a preposition or apposition?

- The *forget me not* was blue
- I hated the *attorney general*
- I gave it to my *brother in law*
- They were an idiot, a *basket case*



# Editing examples in the PARSEME guidelines

- You may create an account in the PARSEME guidelines page (inform Agata)
- If you are logged in, you can add, delete or edit examples in your language

The image shows a screenshot of the PARSEME guidelines website. The main content area is titled "Structural tests (S)" and describes "Structural tests are quite simple preliminary tests that help determining the syntactic structure of a sentence." Below this, it lists "Test S.1 - [HEAD] - Syntactic head" and asks "Does the candidate contain a unique verb functioning as the functional syntactic head of the whole sentence?".

There are three examples listed:

- (EN) `copy` to **pretty-print** → there is an unusual case of an adjective modifying a verb to **drink and drive** → none of the verbs is clearly the head, as there is no universally accepted syntactic representation or coordination
- (NL) `copy` **leven en laten leven** → none of the verbs is clearly the head, as there is no universally accepted syntactic representation or coordination
- (PL) `edit` `source` `copy` **pluć i łapać** → none of the verbs is clearly the head, as there is no universally accepted syntactic representation or coordination

Overlaid on the right side of the screenshot is a modal window titled "Edit example:" with the ID "5.1\_A\_test-s1-no\_0[PL]". The modal contains a "Content\*" field with the text "pluć i łapać", a "negative example" checkbox which is currently unchecked, a "Transliteration : (text)" field with the text "Mmmmh empty ?", a red "Delete Example" button, and "back next" buttons at the bottom right.

# UD GitHub Issues

- Each repo has an **Issue Tracker**
- General guidelines questions  $\Rightarrow$  **docs** issue tracker
- Language-specific guidelines  $\Rightarrow$  still **docs** issue tracker!
  - ▶ ... even if there is only one treebank for the language
- Treebank-specific issue trackers are only for reporting bugs in (or questions about) the particular treebank