cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by
the European Union

UniDive Training School Course 3.6

Documentation and Discussion on Git

Daniel Zeman, Agata Savary

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`
  - 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`

Treebank Specific Pages

- Automatically generated at release time
  - Do not edit them, it will be overwritten
  - Edit treebank README instead
- Treebank hub page generated from README.md and from the data
- Thousands of statistics and examples generated from data
- Comparison of multiple treebanks 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`

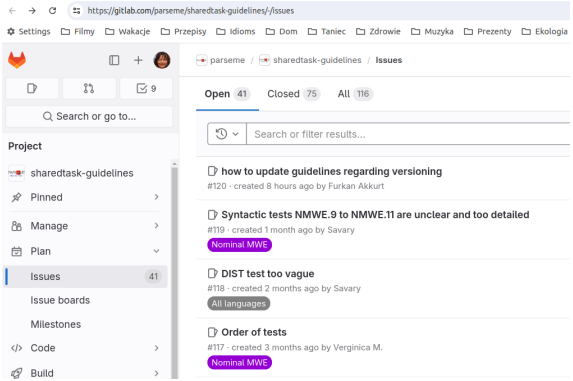
Documenting a PARSEME corpus in README

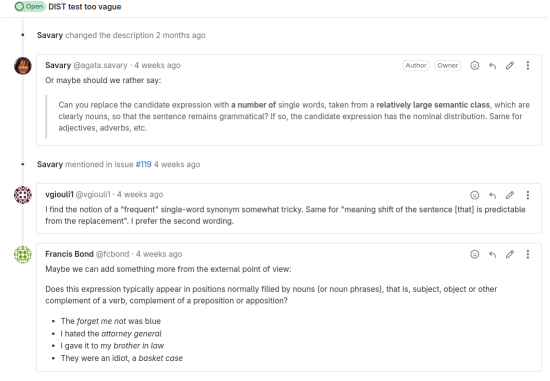
- Each PARSEME corpus contains a `README.md` file, maintained by the **Language Leaders**
- Compulsory items in `README.md`
  - Source corpora and text genres
  - Format
    - For each CUPT column: was the annotation automatic or manual?
    - For the UPOS, FEATS and DEPREL columns: reference to the 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`

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`

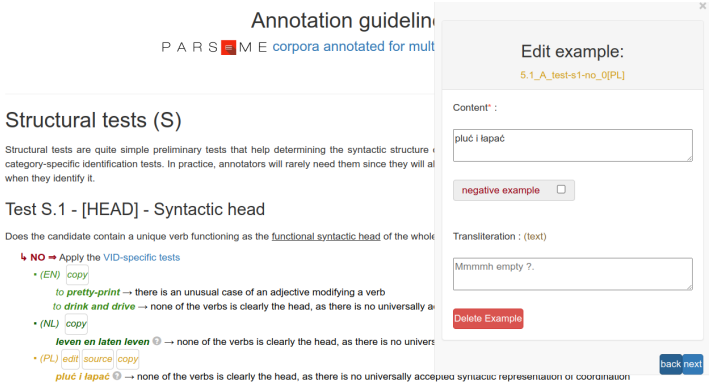
PARSEME Gitlab issues





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



UD GitHub Issues

- Each repo has an Issue Tracker

- General guidelines questions ⇒ **docs** issue tracker
- Language-specific guidelines ⇒ 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