**Meeting December 4 (Dan, Markéta, Šárka, Zdeňka)**

**Next meeting: Monday December 11, 13:00, room 510 (reserved)**

***Alignment:***

New validation reports:

1. words aligned to multiple nodes;

But: Overlapping alignment seems necessary for e.g. coordinated structures with ellipses (as in *Půjdeme do školy a školky*, with *do* aligned to both place).

2. words (non-punctuation) not aligned to any node.

But: We will align as many words as possible, even „být“ = identity-91.

(Both of the reports probably cannot be strict tests (just warnings, can be turned off).)

***Tokens:*** How to give tokens at the beginning of the sentence … ***TODO:*** Dan will decide.

***:refer-number:*** English data seems to treat singular as default, we will enter it explicitly even if singula***r*** (entities, named entities (their abstract concepts), personal pronouns).

But: What about, e.g., a NE denoting 1 entity with plural :name (as *Zbabělci*)

***TODO:*** The problem with ***relative clauses introduced by a possessive pronoun*** still not solved yet :-((

***Functors-to-UMR mapping***: Problems in mapping Prague functors to PropBank argument numbers (instead of using the excellsheet, try new version of synsemclass: <https://lindat.mff.cuni.cz/services/SynSemClassSearch/?version=synsemclass5.0>, then lemma search).

***Functors-to-UMR mapping***: We have (temporarily) introduced the new relation **:according-to** (corresponding to NORM and CRIT functors).

***:quant*** *vs.* ***have-quant-91:***

Where we should use the **:quant** attribute (not relation, see below!!) and where the **have-quant-91** predicate?

Possible solution: Reserve **have-quant-91**only for explicitly expressed comparisons and superlatives(or for cases annotated as comparison in PDT).See also <https://github.com/ufal/UMR/blob/main/doc/quantity.md>

***Wikification:*** We should reserve the **:wiki** relation to wiki instances of NEs (like Q217799, i.e. *Riigikogu* when we refer to the Estonian parliament).

Suggestion: We might introduce a new type of the :wiki relation for types, i.e. common nouns (appelatives) like *hlas*, *volby*, *parlament* … ***TODO:*** should be further discussed with Jan!

***Wikification and coreferential chains:*** We do not want to systematically repeat wikilinks within coreferential chains. However, if there are more links, they should be the same … ***TODO***: candidate to automatic validation?

**snt5:** nominovat = kandidovat-001? … ***TODO:*** Markéta will suggest a new structure.

**snt6:** completely different structure … ***TODO:*** Dan looks OK, Markéta will revise her annotation and compare it also with Šárka’s annotation.

**snt6:** **:quant** is an attribute, not a relation; thus the number/quantity is just a value (not node of the graph, thus it cannot be modified)! The same is valid for **:value**.

***TODO:*** Compare the last sentence (snt7).

***TODO:*** Prepare comparison with Šárka.

**Meeting November 27 (Dan, Markéta, Federica, Mišo)**

**Next meeting: Monday December 4, 13:00, room 510 (reserved)**

Mišo: The known issues with the visualizer have been fixed. For nodes that have the :wiki attribute with a Wikidata id in the value, the visualizer now also retrieves the title of the corresponding Wikipedia article and displays it. There are some wishes for new features (such as zoom in / zoom out; see the [issue tracker](https://github.com/ufal/UMRGraphViz/issues)), but in general, extending the visualizer is not a priority anymore; we will shift focus back to the conversion from the t-layer of PDT to the new u-layer to UMR (“we” = “Mišo + possibly Dan”). The next step in that direction will be a [Treex writer](https://github.com/ufal/UMR/issues/6) that can export the u-layer tree to the UMR file format.

Czech annotation (the Estonian elections article): Markéta and Dan continued comparing their annotations of the file.

* “jehož prezidentským kandidátem je Lennart Meri” – Dan uses implicit event nominovat-001 (and even aligns it with *jehož*). Markéta likes it.
* :wiki for “hlas” (missing in Dan’s file)
* :refer-number singular – Markéta takes it as default, annotates only plural; Dan annotates both. It seems that English UMR 1.0 annotates singular only for personal pronouns.
* “nejvíce hlasů” – Dan has :quant(n / nejvíce); Markéta uses :ARG3 in have-quant to show that *nejvíce* is a comparative word.
* Dan annotates *volby* using the event concept volit-001. Markéta uses so far a noun concept “volby” (including :wiki "Q40231") but it is a process, so it should probably use the verb concept.
* “podle předběžných výsledků” – Dan has (s5p2 / podle :op1 (s5v2 / výsledek :ARG0-of (s5p3 / předbíhat-001 … ))). Markéta has :source (s4v2 / výsledek :mod (s4p2 / předběžný)). The guidelines do not say what is better. :source seems to be used in a more physical sense (“to take something from a place”), not as source of information. In UMR English 1.0, when “according to someone” appears in the data, they create an implicit event with the concept say-01. In PDT t-layer, we have the functor CRIT (criterion). Perhaps we could tentatively define our own relation :crit for UMR annotation. It could be converted automatically to say-01 (or říkat-001) if needed; but we will lose information if we use the verb right away.
* :location vs. :place … Dan uses :location but it is obsolete (inherited from AMR). It should be changed to :place in the data. It should also be fixed in the UMR guidelines, which still mention :location at a few places.
* Something happened “ve volbách”. Dan had :location (=> :place), Markéta had :temporal. There does not seem to be means to annotate placement “in an event”. Temporal relation is probably better than locational but it is still not precise (something could happen at the time of the elections but not be part of the elections). The t-layer in PDT uses the TWHEN functor, so we will stick with :temporal for now.

All the above was found just in sentence 5 (mf920922-133-p4s2). Markéta will investigate differences in the subsequent sentences and bring them up at the next meeting.

**Meeting November 20 (Dan, Markéta, Federica, Šárka)**

**Next meeting: Monday November 27, 13:00, room 510 (reserved)**

Markéta, Dan: the next step: comparison of their annotations

**Title – should it be annotated as “publication-91”? Conclusion for our internal guidelines:**

* Argument for: specific features of the title in the text structure
* Argument against: it has usually a standard sentence structure, there is nothing specific about that
* General conclusion: for metadata, we are going to use publication-91 label (title, location etc.)
* We are not sure whether the temporal and modal characteristics should be included in the document-level annotation
* English data (Philippines): first sentence (title) – no information about temporal or modal relations. The first sentence includes an event. Second sentence (just location – Tacloban, Philippines) – modality is described on the document level.

**Aspect: Estonci volili parlament**

* We decide to follow the morphology (saying this is imperfective, disregarding that the result is very important in this case)

**Wiki with specific events: Estonci volili parlament**

* There are articles there about Estonian parliamentary election, therefore “volit” is labelled with the wikipage
* (English manual: specific events can be named entities.)
* (English manual would probably not link this verb with a wiki link.)

**Zoom Meeting November 13 (Markéta, Dan, Jan, Federica, Mišo, Zdeňka)**

**Next meeting: Monday November 20, 13:00, room 510**

(And the room is now also booked for November 27, 13:00.)

We returned to the discussion of **what is event.** The new incentive is that according to the guidelines ([Part 3-3-1](https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md#part-3-3-1-Aspect)), every event node **should have the :aspect attribute.** The annotators need to know where to annotate aspect, and the validator may try to check that aspect is present if it can recognize an event automatically. As of today, the validator assumes that the presence of an :ARGx relation means that the parent node is event. This assumption is incorrect if the parent node represents a state (rather than process) and it is used in reference or modification (rather than predication). We decided to ignore this detail in the case of stative verbs such as *milovat* “love”: If they occur as concepts, they will be treated as events regardless whether they are used in predication *(on ji miluje)*, modification *(muž, který ji miluje)* or reference *(milovat ji je jeho osud)*. We also tentatively (pending feedback from Julia) decided that the abstract predicates (reifications), such as have-mod-91, will in general be treated as stative verbs, i.e., as events. Reifications are used less frequently in modification because then we can use the relation they reify, e.g., :mod. However (again pending feedback from Julia), abstract predicates that the [UMR spreadsheet](https://docs.google.com/spreadsheets/d/1PVxgXW3ED3OWLieie9scr6iq_xuQ5RAA8YJKwbLwJ2E/edit#gid=1927108453) lists under “discourse relation rolesets/reifications” will not be treated as events. In the future we may want to extend this exception to a few other abstract predicates, such as publication-91.

Summary: The validator will tentatively keep the current test but we may later modify it to exclude the discourse relations. Markéta will ask Julia whether it is the right approach.

Another question for Julia: Why are **discourse relations** annotated in three different ways?

* Using an abstract concept with :op1, :op2, … relations, e.g. and.
* Using an abstract roleset with :ARG1, :ARG2, …, e.g. contrast-91.
* Incorporating one event as child of another event via the relation :pure-addition (see example (4b) in [Part 3-1-6](https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md#part-3-1-6-discourse-relations)).

For information: Dan has created [our own fork of the UMR guidelines](https://github.com/ufal/umr-guidelines/blob/master/guidelines.md) Github repository. There are some things fixed in comparison to the original repository (e.g., the attribute :ref-number has been renamed to :refer-number right before releasing UMR 1.0 but the original guidelines still do not reflect the change).

**Meeting November 6 (Dan, Zdeňka, Federica, Markéta, Míša)**

**Next meeting: Monday November 13, 1pm , room 510**

(Šárka has vacations)

**(The meeting after that: Monday November 20, 1pm, room 510)**

**Dan - validator**

The validator has been updated. Now it checks for known relations and prints a warning if it finds an unknown relation.

Dan will check the [spreadsheet](https://docs.google.com/spreadsheets/d/1PVxgXW3ED3OWLieie9scr6iq_xuQ5RAA8YJKwbLwJ2E/edit#gid=1927108453) with lists of relations/abstract predicates/... for the UMR tool to see if his validation list is exhaustive, and possibly expand it.

As of now, expected values have been specified in the validation script only for aspect and modstr. Attributes are not checked yet.

**:quot**

The :quot relation indicates that an event is being reported. Cf. snt3 of the first English file:

*About 200 people were believed killed and 1,500 others were missing in the central Philippines on Friday when a landslide buried an entire village , the Red Cross said .*

N.B.: in the released data it's *quote*, while in the guidelines it is called *quot*.

**Alignment**

We need more specific guidelines about alignment. Several issues:

1. Alignment of Named Entities

Cf. *Estonians* file. The alignment of the first sentence *Estonsko* is s1c: 1-1, all the other nodes are unaligned (0-0). It appears that in the case of named entities it is the parent node *country* which gets an alignment, not the name of the country (*Estonsko).*

Marketa: maybe it would be better to align the name of the NE itself.

2. Negative alignment

Dan created a GitHub issue today about negative alignment (e.g., -1:-1), which may be an error or it needs to be explained. (<https://github.com/cu-clear/UMR-Annotation/issues/2>)

3. Discontinuous alignment

* Discontinuous alignment is needed, we need to define a way to annotate it.

As of now, there is nothing similar neither in the guidelines nor in the released data.

* Dan: the validator accepts multiple alignment ranges separated by commas (e.g., 1-1, 4-4). However, it appears only in our data.
* For instance, prepositions: we may align them to the noun in the same way it is done in the tectogrammatical layer, but the alignment could still be discontinuous - for example if an adjective occurs between the preposition and the noun.
* Also, the validator complains if it finds alignments like 1-1, 2-2, because it assumes that the correct format should be 1-2.

4. Numbering tokens to make alignment easier

It is possible to add a commented line after the sentence line to store token ids and make the alignment process easier.

Ideally, it should be standardized. Currently, sometimes the token line appears before the sentence, sometimes after, often it is not present at all. Sometimes it starts with *tk*, sometimes not, etc.

**Meeting October 30 (Dan, JH, Zdeňka, Federica, Šárka)**

**Next meeting: Monday November 6, 1pm , room 510**

(Šárka has vacations, two weeks.)

**(The meeting after that: Monday November 13, 1pm, room 510)**

(Šárka has vacations, the other week.)

**Míša: Visualizer**

<https://colab.research.google.com/drive/1pbmJ3k3_qFuVM44neVHikiJKe81xsAHD?usp=sharing>

It can be used online, does not need to be installed.

How to:

1. Initialize the environment – press the button (triangle in a circle)
2. Move to Title below – press the button (triangle in a circle). It will run till you browse there and choose a file.
3. Move to Import below – press the button (triangle in a circle). You will find the visualization below.

The issues and questions should be written here:

<https://github.com/ufal/UMRGraphViz>

<https://github.com/ufal/UMRGraphViz/issues>

**Šárka: the first part of the Estonian file is annotated on Github**

* Some mistakes found by Dan’s validator must be corrected.

**Dan: Validator – first steps for the validation what is an event and what is not**

* Searching for things showing that something is an event / entity
* Not finished yet
* At present it can print all nodes, attributes and events found in the document. This is done on demand only: use the option “--print-relations” (or run validate.py -h and it will tell you what to do).

**Federica: went through the PDT grammatemes searching for correspondence.** Saved on Github. Almost ready, just the Numertype needs a bit more explanation.

* Github: doc/grammatemes-attributes
* <https://github.com/ufal/UMR/blob/main/doc/grammatemes-attributes.md>
* **Polarity (positive – negative):** unclear: unknown polarity, like in “Did you see him?”
  + Dan: one possibility – the polarity of yes-no questions is always “unknown”.
  + Jan Hajič: ask Julia about the value “unknown”? In the data, it seems connected to questions. (Where do you… - the argument for Where is “unknown”).
  + Dan: so, it is not connected to yes/not questions only, but to questions in general
  + Federica: + to indirect questions: I don’t know if that's important or not
  + Dan: can we see in the annotation the interrogative mode, elsewhere?
  + Zdeňka: in AMR manual, they say “unknown” is used to mark questions. Used specifically for yes/no questions (what is the truth about the assertion).
    - With relative dependent clauses, instead of “unknown”, the “truth-value” should be used.
    - Federica: this is not in accordance with the example from the UMR manual: I don’t know if that's important or not (using “unknown” value).
* **Tense:** the PDT recognizes the following values: simultaneous, preceding, subsequent. They cover absolute as well as relative tenses
  + Šárka: this can be misleading when transferring this annotation of tenses to the UMR
  + Dan: we have to check the tense annotation and know exactly what can be transferred to the UMR.

**Meeting October 23, 2023 (Markéta, Dan, Federica, Šárka)**

**Next meeting: Monday October 30, 1pm , eastern corridor**

(Markéta is not coming.)

**Marking gender, semantic features of gramatemes in the PDT**

Šárka: sending the answer about gender by Julia

Markéta: What about going through PDT grammatemes and deciding which one of them are semantic

TASK go through the list of grammatemes and mark when and how they are already reflected in the UMR scenario.

**Annotation – Estonian file**

Markéta has the full annotation, many questions for Dan

Dan: there was a google link about concepts, entities etc. UMR is using – Markéta: yes, it being updated, probably by Julia, it is extremely useful.

<https://docs.google.com/spreadsheets/d/1PVxgXW3ED3OWLieie9scr6iq_xuQ5RAA8YJKwbLwJ2E/edit#gid=1927108453>

* On Github, under Useful links

TASK Šárka – finish the Estonian file

Dan (to Markéta and Šárka): collect the problems in a list, they will be solved step-by-step

**The difference between UMR manual and released data**

Markéta: there are the same nine sentences in both of them, but the annotations are not the same. The data is inconsistent, there just some halves of sentences annotated there. The manual is changing gradually.

**Comments on the annotation manual**

Markéta went through the whole sentence-level part, except for the aspect. She commented on problems and particular problems there.

On the main page <https://github.com/ufal/UMR/tree/main> - under Documentation (below)

**Aspect with the verbs in the description of people (teacher – a person who teaches)**

Šárka: a teacher teaches habitually vs. for a murderer, it is enough to commit just one murder to be a murderer. Do we want to mark the aspect at the verb to\_teach, to\_kill?

Dan, Markéta, Šárka – solution of the discussion: we want to mark these verbs in a specific way (e.g., aspect: 0) so that we are able to find them later again and mark the aspect there consistently later.

**Relative clauses with state verbs + adjective as a property (being X) vs pure adjective (modification)**

Markéta: doesn’t like that UMR annotates in a different way “a man who is tall” vs “a man who is coming”

Dan: what do we annotate with verbs? Aspect, temporal relations, modality, Arg+number

Markéta: temporal relations may be annotated at the document level

Dan: we can extend the validator in such a way it would understand that the concept was treated as an event. (All the subfeatures should be in accordance, if they are not, we would find these cases.)

Markéta: loving husband – love is a state verb, therefore the phrase would not be treated as an event. Does it mean that the aspect should not be marked there?

Markéta: other questions concerning verbs

Are we going to treat every verb as a verb? (That would make the things much easier.)

Dan: we need examples from our data (Czech, Latin) and solve the question about state and action verbs.

**Meeting Oct 16, 2023 (Dan, Markéta, Mišo, Šárka)**

**Graph visualizer**

Mišo tried to help Markéta install the UMR graph visualizer on Windows. Unfortunately, it requires Python library pygraphviz, which cannot be installed without Microsoft Visual C++ (meaning you invest several gigabytes of your disk space, wait an hour or so for the installation to complete, without a guarantee that it will then actually work). Dan tried the same on his laptop but gave up. It should be easier to run the visualizer on ÚFAL Linux network (can be done even remotely through the Putty SSH terminal), then put the result to $HOME/public\_html and view it through <http://ufallab.ms.mff.cuni.cz/~$user/>.

**Validator**

Dan fixed a bug Markéta found in the validator while applying it to data/czech/mf020922\_estonsko-ML.txt.

**Meeting Oct 9, 2023 (Dan, Markéta, Jan H., Zdeňka, Federica, Šárka)**

**Validator**

Dan had modified Python validator, it can read more languages now.

* Dan: the American data has a lot of errors (found by the validator). How to communicate this with the US team
* JH: solve it via Github issues; discrepancies between manual and data should be sent to Julia
* Dan: it is our interest, too, that the American data is good.
* It doesn’t help too much to write Github issues – the US side probably has no capacity to answer.
* JH: keep adding issues to the guidelines, hopefully someone will find it

**Omission of gender in the manual**

* Markéta, Šárka: there is no mention about gender in the manual
* Šárka: we should have an overview of Czech grammatical categories and their semantic relevance
* JH: maybe, gender is omitted intentionally, because of the team capacity (?). Many of these questions have been solved in the PDT (modality, person etc.).
* JH: send an e-mail to Julia and ask about the gender [julia.bonn@colorado.edu](mailto:julia.bonn@colorado.edu) (Šárka sent the e-mail right now.)

**Next steps:**

* Continue the annotation of the Estonian file – with a text editor. Check it with the Python validator.

**Dan:** we need to talk with the US side about the validator and the reliability of the data.

**Meeting Oct 2, 2023 (Míša, Federica, Markéta, Dan)**

**Dan – first version of the validator!**

UFAL/UMR Github, new folder tools/validate.py

**Miša, Dan – problems with data consistency even within the already released data!**

Inconsistencies also between the released data and the guidelines (e.g., different names of attributes/values)

No formal data format - only what can be inferred from the released data :-(

Not sure how strict the validator should be!

**Python for Windows users:**

***Instalation:***

* download Python 3.11.5 (or higher) from

<https://www.python.org/downloads/>

* install it (double click) – "basic" installation (not customized), follow the instructions
* online tutorial available: <https://docs.python.org/3.11/tutorial/index.html>
* online documentation: <https://docs.python.org/3.11/index.html>
* user guide for Python on Windows: <https://docs.python.org/3.11/using/windows.html>

***How to run the validator (with Windows PowerShell):***

My recommendation is to start in a sandbox folder you will create somewhere on your PC  :

* copy the validate.py script from the github repository to your sandbox folder
* collect your files for validation in the same folder
* open Windows PowerShell (from the Windows Menu)
* move to the sandbox folder (use the cd command)
* write the command "python validate.py "
* it will tell you that "No module named 'regex'."
* write the command "pip" … it serves for installing missing libraries and will give you basic help (which you probably will not understand:-)
* nevertheless, write the command "pip install regex" … hopefully, in this way, the regex library will be installed (Successfully installed regex-2023.xxx)

Now it should be prepared!

* write the command "python .\validate.py <name\_of\_file\_you\_want\_to validate>" (tip: start with one of the released files)

You can ask for less strict options:

* + --allow-trailing-whitespace (do not report extra spaces)
  + --allow-forward-references (do not report forward node references within a sentence level graph)
* the final command looks as the following:

python .\validate.py --allow-trailing-whitespace --allow-forward-references .\english\_umr-0001.txt

(output: \*\*\*FAILED \*\*\* with 3 errors for english\_umr-0001.txt :-((

***Alternative (if there are issues with the Windows PowerShell )***

* copy the validate.py script from the github repository to your sandbox folder
* collect your files for validation in the same folder

use the Windows command line (press the Windows key + R, then type "cmd.exe" and press Enter)[[1]](#footnote-1)

* in Windows Settings, you must define the path to your python app
  + Start  Settings  System  About (last item in the left menu)  Advanced system settings (scroll down)
  + It opens "System properties" dialog box, click on "Environment Variables"



* + It opens the "Environment Variables" dialogue box, focus on the "User variables for <your profile>" (upper part):
    - mark "Path" (by single L click)k and then open "Edit"
      * Click on New and enter the path where your Python is installed[[2]](#footnote-2)
      * confirm by OK
  + Back in the "Environment Variables" dialogue box, "User variables for <your profile>" (upper part):
    - Click on "New"
      * Insert Variable name PYTHONPATH
      * Insert Variable Value … the path where your Python is installed (as above)
      * confirm by OK

Now your computer knows from where python should be launched.

* move to the sandbox folder (use the cd command)
* write the command "python validate.py "
* it will tell you that "No module named 'regex'."
* write the command "python -m pip install regex"

… hopefully, in this way, the regex library will be installed (Successfully installed regex-2023.xxx)

Now everything should be prepared!

* write the command "python .\validate.py <name\_of\_file\_you\_want\_to validate>" (tip: start with one of the released files)

You can ask for less strict options:

* + --allow-trailing-whitespace (do not report extra spaces)
  + --allow-forward-references (do not report forward node references within a sentence level graph)
* the final command looks as the following:

python .\validate.py --allow-trailing-whitespace --allow-forward-references .\english\_umr-0001.txt

(output: \*\*\*FAILED \*\*\* with 3 errors for english\_umr-0001.txt :-((

**Federica – raised a problem NE Romani in UMR**

… quam Romani gramaticam vocaverunt. (… which the Romans called grammar.)

suggestion: annotate as NE, type ethnic-group

:wiki "Q1651392"

:name (n/ name :op1 "Romanus") We decided to at least normalise *Romani* (nomin. plur.) to the singular form *Romanus*.

BUT: American businessman as NE America ex. 1(2)?

The annotation as ethnic-group was confirmed, as the context here does not make any reference to the citizenship of Romans (not treated as a political group, which would have justified the annotation of p/ person from a certain c/ country), but the focus is on the fact that they speak the same language (i.e. ethnic/cultural feature).

:refer-number plural

see also <https://github.com/ufal/UMR/blob/main/doc/entities.md>

**Meeting Sep 25, 2023 (Cristina, Federica, Šárka, Dan, Mišo)**

**Federica – representation of Latin in UMR**

Latin annotation is a little behind the Czech one.

Latin has 2 Vallexes

1. Collected from Latin texts annotated in PDT style, derived from textogrammatics; 1.000 entries
2. No examples, but including definitions derived from WordNet; ca 4.000 entries

Míša: can some Latin annotations be visualized with his tool?

Federica:

* Yes, 4 sentences were annotated (sentence-level annotation, without the document annotation)
* From Sallust (war description; we have the tectogrammatical annotation)

Dan: maybe different data? Obvious difficulties with Vallexes

Federica: so far stick to these data; these weeks – technical solutions of the annotations

* Cooperation with Vallex people, using editor like for the Czech Vallex and adaptation for Latin
* Question of grammatical number:
  + UMR: entities have number (derived from morphology)
    - Questionable: Philippines (1 country) – no number is given in the example
    - Dan: proposed solution: if the form says you the number and you agree with that from the semantic point of view, write there the number as you feel it. If you are not sure, write there “?” so that these examples can be found later.

Míša: prepared a visualization of Federica’s annotation

<https://ufallab.ms.mff.cuni.cz/~mnovak/umr/graphs/ufal/latin/sample.txt/visualization.html>

Šárka: are we going to connect each occurrence of mentioning a person in a text to Wikipedia, or will we do it each time? (Coreferential chains.)

Dan: how would we say which one should be connected to Wikipedia and which one not?

Šárka: maybe just the first occurrence in the coreferential chains should be connected to Wikipedia?

Míša: looking into English data.

Federica: proposing checking the text about Edmund Pope (file 5)

* Name – American businessman – name (there is a coreference chain like in the PDT)
* In the first / third / four sentence there is a wiki-link (i.e. from the Named Entity)

Míša: this could be solved automatically or semi-automatically

Šárka: so, it is important to mark the coreference correctly and make at least one wiki-link.

Dan, Míša: we can check it then automatically: e.g., when we have one coreference chain with two wiki-links, it should be the same wiki-link. If we have two coreferential chains with the same wiki-link, the coreferential chains should be probably connected.

Šárka: what about marking the gender (with the people and maybe animals?). It can be important for coreference annotation.

Dan: nevertheless, we don’t need it once we have the annotation of coreference. So the question is: do we need a gender as a semantic feature?

Šárka: yes, we (in Czech) need it to be able to imagine the situation

Dan: yes, if we know the biologic gender, we should annotate it.

Dan: overview of the tools: Míša’s visualization, convertor AMR to UMR, …

We should put more emphasize on the validation of the data. – going to work on that for the next weeks.

Míša: going to work on the conversion and on the population of the conversion rules.

**Meeting Sep 11, 2023 (Prof. Hajič, prof. Hajičová, Míša, Federica, Eva F., Markéta, Šárka, Zdeňka)**

Comparison of 3 variants of annotation (Estonian file), with the visualisation by Míša

* visualizations:
  + Markéta: <http://ufallab.ms.mff.cuni.cz/~mnovak/umr/graphs/ufal/czech/mf020922_estonsko-ML.txt/visualization.html>
  + Dan: <http://ufallab.ms.mff.cuni.cz/~mnovak/umr/graphs/ufal/czech/mf020922_estonsko-DZ.txt/visualization.html>
* Annotation Markéta:
  + Predicate „publication“ – from the English published data
  + Predicates in the list of predicates and arguments (publish-91)
  + Argument structure – from the English released data
* Grammatic number with named entities:
  + In English data, the number is not given with named entities
  + Maybe if we have a link to wikidata, it is clear that it is one person or more

Federica:

* Country – our states (like Estonsko)
* State – US states

Šárka:

* Grammatic number with NE: vote for writing the number in the annotation (both Germanies; 1 Estonec – všichni Estonci; 1 Novák – všichni Novákovi)
* Estonia – is it an event or not? It dependes on the interpretation of the ellipsis (either following of the English annotation, or based on the PDT manual)

Zdeňka: publication-91, ARG0: author, ARG1: title, ARG2: …

Markéta: Estonci volili parlament – the whole sentence is a title, it should be ARG1 of publication-91

Šárka: what about **number marking with verbs** – Dan: no, this is a feature of nouns

What about **gender** – Dan: just the semantic gender should be marked

What about **aspect** – Šárka: the aspect classification in UMR is more detailed than in PDT. Markéta: if we want to take the aspect from PDT automatically, we can distinguish perfective and imperfective aspects only.

**Temporal** annotation: it is a part of document-level annotation

**Meeting Sep 4, 2023 (Markéta, Zdeňka, Federica, Šárka, J. Hajič, Míša, Cristina)**

**Míša:**

* short introduction of the new release of the UMR data, as described in the email.
* Visualization based on this corpus. TASK any comments and suggestions are welcome – how to distinguish any relations, concepts etc.
* Another project will be a conversion of Czech files to UMR
* We can let the UMR team know this tool exists. It is not publishable yet.

Michal’s report about the tool in the e-mail:

Hi all,

I adjusted the AMR Reader tool (<https://github.com/panx27/amr-reader>) to be able to read UMR files and visualize them. Following our discussion two weeks ago, I also changed the visualization so that atomic values are displayed as nodes' attributes, which makes the final visualization less messy even for large graphs. I will publish UMRGraphViz on UFAL GitHub soon.

I used the tool to visualize the documents from the recent UMR 1.0 release. You can browse the documents here: <https://ufallab.ms.mff.cuni.cz/~mnovak/umr/graphs/umr-v1.0/> (open visualization.html for a given language and document to visualize it).

Known issues:

- it displays only sentence graphs. Alignment and document-level annotation are ignored.

- original sentence in its surface form is not displayed for some languages (e.g. Arapaho, Navajo). The reason is that the format of representing this information has slightly changed in the official release (compared to <https://github.com/cu-clear/UMR-Annotation/tree/main>).

- there is no visualization for arapaho\_umr-0001.txt and other Arapaho documents may be corrupted, too. Parsing of that file failed from yet unknown reason.

Best

M.

* Plan: small adjustments, then release

Markéta: eventive concepts

* What about words like *válka, příchod*? (Either entities – conservative approach, or events – based on the Vallex, they have a valency frame. Problem: not consistent, some similar words in PDT Vallex are not assigned a valency frame.)
* JH: it would be good to use our sources. Nominal valency does not depend on the distinction between states, events etc. E.g., a book can have an author as a position in a valency frame, nevertheless, it does not mean that a book is an event.

Thus, even if the NP has a frame, it does not mean it is an event.

Markéta: we need to know what this difference means for the annotation

JH: let us see real examples from Míša’s database

Míša: <https://ufallab.ms.mff.cuni.cz/~mnovak/umr/graphs/umr-v1.0/>

0001.txt, sentence 7 (roofing as a material on the roof)

Markéta: with some words, there can be connection to the Wiki (as entities) as well as to a valency frame (e.g., WW2). Can we connect one concept with both features (valency frame and Wiki)?

JH: what can prevent us to do that?

Markéta: We have not seen such an example in the manual.

JH: Most events do not have Wikipedia entries, it is rather entities.

Do we want to connect a general entry of war (a type) to Wikipedia, as well as special occurrences (WW2)? Do we want to have both connections with one phrase in our text?

JH: maybe it would be good to make a differences (and mark it) between a type and an occurrence (individuum). Maybe, we should always have a type (or valency frame, or later on Synsemclass, Propbank…).

Zdeňka: mé psaní (my writing) – how to interpret it? Either I am the author, or just an owner.

JH: when it is a letter, it is an entity.

JH: in UMR, there is just an unclear border between entities and events. But we have to follow the instructions. It would be good to put “type” everywhere where it is possible.

The primary database for us is “Wikidata”.

Example in UMR data: file 0001, sentence 4: Eyewitness is a person who witnesses.

Míša: should the types be linked to the Wikidata?

Federica: working on a file in Latin, still technical and scientific issues. Trying to prepare a PDT-like file that could be later converted to UMR-like file.

Šárka: can we annotate the Estonian file in the Míša’s tool?

Markéta: yes, she tried that. Dan did it too, in his way. Markéta proposes us to try the annotation separately. TASK try (before Hejnice) to annotate at least a part of Estonian part. In Hejnice, we can compare it. In a text format, then Míša can converse it to his visualization.

**Meeting Aug 28, 2023 (Markéta, Dan, Zdeňka, Federica, Šárka, Eva F., J. Hajič)**

**Markéta:**

Eventive concept document – TASK comment it

One or two concepts? – not ready yet for comments

Functors to AMR labels:

* Markéra: Table: conversion between functors and AMR labels (update) – not all functors are covered yet. Questions esp. about coordinative functors, they are solved in a different way in AMR.
* Eva F.: document – verb-specific mappings between functors and argument roles (from SynSemClass)

Based on these tables, Michal can prepare a conversion from PDT to AMR

**Plans:**

Dan – revise the annotation of the Estonian file

Šárka – meet Markéta, discuss what is needed now

Federica – merging two versions of Vallex

**J. Hajič:**

Martha Palmer published small data – English (300 sentences), Arapau – official UMR.

It will soon be in Lindat. Examples for us. The release is being prepared by Jin, communicating with Ondřej Košarko.

Probably the same version of data is on the UMR Github, or maybe it is smaller?

**Markéta:**

PropBank webpage – there were some abstract predicates there, but they somehow disappeared.

**Reservations for October:** 510 – ask Jirka Mírovský for reservation when it is open; or later we can try it ourselves.

# Next meetings

28/8 – 13:00 Dan +, Federica +, JH +, Markéta +, Míša +, Zdeňka +;

Cristina –, Eva +?, Šárka ?

# Meeting 2023-08-21 (Dan, Eva, Federica, Markéta, Michal)

**(0) cu-clear/UMR-Annotation repository**

invitation by Julia Bonn (July 26)

<https://github.com/cu-clear/UMR-Annotation>

Dan did not have an access to the repository – he should send an email to JH (he will ask Julia to add Dan as a member).

**(1) updated AMR viewer for UMR annotations**

Michal: visualization of the Eng data from the UMR-Annotation repository

just sentence-level annotation

problem: too complicated graph

relations displayed as edges between concepts, repeating mentions as single nodes … OK

BUT attribute-value pairs displayed as relations as well (inherited from the original viewer) … KO

todo (Miša):

attributes (with pre-defined set of values) should be represented as attributes of individual nodes

**(2) entities vs. events**

solution from the mast meeting (2023-07-31) discussed

todo (Markéta): update the Eventive concept notes accordingly

<https://github.com/ufal/UMR/blob/main/doc/eventive-concepts.md>

**(3) functors  UMR relations/attributes conversion**

Markéta: 2014 table updated and commited to the github repository (still provisional)

<https://github.com/ufal/UMR/tree/main/tecto2umr>

todo (Markéta): further refinements needed

**(4) verb specific functors  UMR relations conversion**

Eva: extracted conversions for individual SSC classes

todo (Eva): conversions for individual verb frames

# Next meetings

21/8 – 13:00 Cristina +, Dan +, Federica + (zoom); Markéta +, Míša + (?zoom);

JH in China – zoom +?, Zdeňka +?, Eva ?, Šárka ?

28/8 – 13:00 JH +, Dan +, Federica +, Markéta +, Zdeňka +

Eva ?, Šárka ?

Cristina –

# Meeting 2023-07-31 (JH, Cris, Eva, Markéta, Michal, Zdeňka)

**(0) cu-clear/UMR-Annotation repository**

invitation by Julia Bonn (July 26)

Does it arrive to everybody? If not, send an email to JH, he will ask her to add you as a member.

**(1) convertor from AMR**:

Michal:

update and plans:

* defined first draft of u-layer
* work on vizualization in Tred, now it works (with Dan)

**(2) coreference / co-occurences – visualization in TrEd:**

… keep both nodes (and merge them on demand / when converting u-layer to UMR

??? both for inter-sentence and intra-sentence relations

*Peter likes himself.*

UMR annotation: and its visualization in TrEd:

#Ref

l/like

p/Peter

:ARG0

:ARG1

l/like

:ARG0 (p/person

:name ( n/name :op1 "Peter" ))

:ARG1 (p)

**(3) relations vs. attributes:**

* relation between two concepts (two UMR nodes)
* attributes: only for one concept/UNR node, with (?close] set of values (like sg/pl or numerals)

BUT there are some labels defined as attributes in UMR guidelines which are later treated as relations (as well)

e.g., :quant in ex. 1 (1) *… more than eight months* :quant as attribute (with value 8)

:range (m/ more-than

:op1 (t3/ temporal-quantity :quant **8**

:unit (m2/ month)))

BUT 1 (6) … *the next several days* :quant as relation (*several* as a child)

(t/ temporal-quantity

:quant (**s2/ several**)

:unit (d/ day

:mod (n/ next)))

***provisional solution for u-layer***:

* all UMR attributes that appear also as a relation provisionally represent as relations
* UMR attributes with only predefined (close) set of values represent as attributes

list of attributes/relations <https://github.com/ufal/UMR/blob/main/doc/relations-attributes.md>

todo: update whenever found a new attribute/relation (or change their description there)

**(4) functors**  **UMR relations/attributes** **conversion** … todo

**(5) entities vs. events** … unclear cases

**rule 1**: Whenever a concept has a roleset  represent it as event

(link to the lexicon with its roleset)

**rule 2**: JH: ***Be conservative (at least for the time being)!***

i.e.: We do not want to add a big number of nouns / adjectives to the lexicon OR represent them as verbs, unless we have a clear case of a deverbal noun (ending with -ní/-tí) / deverbal adjective (??) … YES - morphological criterion!!!

 ***verbs:***

**rule 1a**: JH: consider them as eventive concepts (both action and state/stative verbs)

***!! in conflict with some UMR examples !!***

TODO: to be changed/stressed in the internal guidelines <https://github.com/ufal/UMR/blob/main/doc/eventive-concepts.md>

 ***nouns:***

e.g., *jídlo* … should not be represented as ARG1of *jíst* but as an entity

(as it can be related also to other eventive concepts as, e.g., *vařit*)

BUT agentive nouns like *teacher*, *driver* … as ARG0 of the respective verb (*teach-01*, *drive-01*) even in context different than teaching, driving (intention of the speaker to use just these nouns) ***!! in conflict with some UMR examples !!***

TODO: to be changed in the internal guidelines <https://github.com/ufal/UMR/blob/main/doc/eventive-concepts.md>

 ***adjectives:***

**rule 1b**: clear deverbal cases should be represented as events

(mapped onto respective verbs in the lexicon, as ARGx-of …)

e.g., *podobný* ‘resembing’  *podobat se* ‘resembe’(morphology!)

**rule 2a**: primary adjectives (like *chytrý, vysoký, clever, high*) as non-eventive concepts

compare with UMR examples

3-1-1-3 (2a) The tall man...

3-1-1-3 (2b) The man, who is tall...

(m/ man

:mod(t/ tall)) (the UMR graph by ML, guidlines only says that this is not an event)

3-3-1-3 (1b)The doctor is tall.

(h/ have-mod-91

:ARG1 (d/ doctor

:ref-number Singular)

:ARG2 (t/ tall))

Compare also to AMR guidelines, sect. Adjectives that invoke predicates

<https://github.com/amrisi/amr-guidelines/blob/master/amr.md#adjectives-that-invoke-predicates>

(i)*the attractive man* = the man who is ARG0-of attract-01 … **OK for CZ** (morphology!)

the same solution for Eng. adjectives ending with -ed (acquainted)

the same solution for other types of endings, like -able (edible) -ful,

(ii) to be+adjectives … often exist natural corresponding verbal predicates … **NOT for CZ ??**

*The soldier was aware of the battle.* … *be aware (of X)*  *realize-01*

(iii) adjectives without natural verbal predicates … create predicate … **NOT for CZ ??**

*be responsible (for X)*  *responsible-01* (cause), *responsible-02* (trustworthy), *responsible-03* (duty)

*be nervous (about X)*  *nervous-01*

*be serious (about X)*  *serious-01* (no kidding), *serious-02* (grave)

(iv) adjectives like *sad, white, and free* … as predicates if there is an implied event or proces

… **NOT for Czech ??**

general question: How much inferencing we want to include in the annotation?

JH: Be conservative with ***reification*** (reification = converting a role into a concept is called reification)!

Reification is used, e.g., for the change of TFA in AMR guidelines (“One reason is to make a relation the focus of an AMR fragment.”)

more advanced solution: create more general predicates like *color-01* … **?? not now for CZ**

“ideal” solution: JH: all adjectives as predicates (like the proposal by Dan), but **not now for CZ**

TODO: if agreed, then to be changed in the internal guidelines <https://github.com/ufal/UMR/blob/main/doc/eventive-concepts.md>

# Meeting 2023-07-24 (Dan, Michal, Eva, Markéta, Federica)

**AMR guidelines**

We should consult also ***AMR*** guidelines!

<https://github.com/amrisi/amr-guidelines/blob/master/amr.md>

Frame Files … PropBank 3.4 should be used

<http://propbank.github.io/v3.4.0/frames/index.html>

**Convertor from AMR**:

Michal:

***relation*** … relation between two concepts (UMR nodes)

as, e.g., :ARG0 is understood as a relation between the predicate and its participant

However, in UMR guidelines, attributes and their values are also sometimes mentioned as relations

as, e.g., :polarity

 consequences for TrEd

update and plans:

* work on PDT-to-UMR conversion, now it works :-)
* now work on vizualization in Tred (with Dan)
* plan: focus on coreference (as it leads to structural changes)
* plan: start with verbs
* work on functors conversion (incomplete list of functors from 2014), based on AMR guidelines (and on sample English annotation, where necessary) (Markéta:)

Problem repeated throughout the UMR (and AMR) guidelines – uneasy to extract **an exhaustive list** of relations (between nodes/concepts) and **full list of values for individual attributes**.

**coreference**

Míša: 3 different ways in UMR

* inversed participant roles
* reference to an already specified concept variable (only within a single sentence)
* :coref attribute (definitely inter-sentential relations)

Suggestion: Use the :coref attribute for two (or more) mentions of the same entity within a single sentence only when necessary (i.e., when some of their values differ)

**ellipses**

Estonsko:

Dan suggests to annotate such fragments using the abstract predicate have-location-91

(h/ have-location-91

:ARG1 (e/ event)

:ARG2 (c/ country

:wiki "Q191"

:name (n/ name :op1 "Estonsko")))

How should coreference look like? All events in the article annotated as subset-of event?

**quantities**

:quant

* typically an attribute with a numeric value
* but sometimes understood as a relation … approximate cardinalities (*more than, most*),

How to annotate percentage?

How to annotate **indefinite quantity**?

[la] *paululum commoratus* 'having waited for a while'

**English concepts in Czech annotation?**

where to use Czech concept words and where to stick to English ones

e.g., more-than

* abstract events … OK (distinguished by the \_91 suffix)
* named entities … governing organization (and as a node Ministerstvo … zemědělství)

?? Where is the borderline – where else we want to use Eng

* more than 3 annotated as the more-than relation … and for Czech?
* the abstract concept percentage-entity with the numeric attribute :value

**Lattices for individual attributes?**

Make use of values on different levels?

e.g., for number, we want to use *singular*, *dual* (***only when annotated as such in PDT***) and *plural*, even if they appear on different levels of the lattice (and plural in not a leaf node there)

The singular/plural values seem to be in compliance with English examples.

<https://github.com/umr4nlp/umr-guidelines/blob/master/Guidelines_figures/Number%20Lattice.jpg>

**UMR validator**

We will definitely need an UMR validator!

# Meeting 2023-07-17 (Federica, Cris, Zdeňka, Eva F., JH, Dan, Šárka, Michal)

**UMR Writer:**

Summary from the last meetings: the interface does not seem to be developed enough yet, for now, we are not going to use it.

JH: somebody (everybody) in our group should be able to work with it. The American group will develop the tool later and we yet need to stay in contact.

JH: UMR-writer 2.0 – [Sijia Ge](https://aclanthology.org/people/s/sijia-ge/) from ACL meeting (Boulder) is going to develop it. It should be better. This guy is from China, employed just til December 2023. JH has his e-mail address.

Paper on UMR writer 2.0. <https://aclanthology.org/2023.law-1.21/>

**Using TrEd**

JH: It would be good to work in TrEd. (edges in the graphs, based on the nodes; possibility to see the tree as a whole).

But we should be still able to edit it in UMR Writer

**Convertor from AMR**

Míša:

Three types how coreference is represented in UMR (described in the documentation on Github)

The original instructions from UMR about coreference are unclear.

JH: whenever we have a question, we should write to Julia – we need to set a frequent communication.

Dan: write it as an issue to Julia (official [UMR github](https://github.com/umr4nlp/umr-guidelines/issues), not our github)

**Events (processes) and entities**

Dan: Named events (named entities): should they have the abstract concept of entity?

Basically: event (ARG0, ARG1…), and then adding attributes which are usual for entities (name, link to Wikipedia)

Dan: **Basic question: why do we need to discern states and entities? Why are states not understood as events (with copulas)?**

JH: AMR classic problem: entity vs. type. It is clear with named entities (the type does not show up). IBM – name, type: company-entity.

JH: UMR is about the structure of the graph. They talk about the types only.

ŠZ: What is the difference between a concept and a type?

JH: more or less, they are the same. The most important difference is between specific cases (my running at Marathon) and concepts (running itself)

JH: *a clever man – he is clever*:

* In the ideal representation by JH it would be the same (predicate, eventive concept)
* In the UMR representation it would be different (modification – event / predicate)

Dan: still, what does it change in the annotation? Ask by Julia?

**Inferences**

JH: we want to make immediate inferences (if you listen to me, what you remember that I said), not more

**How far should our annotation be based on the word formation?**

Cf. CZ *jídlo* (food, from *jíst,* annotated as a part of the verb *jíst – to eat*) and EN *food* (not connected with *eat*) – different annotation?

Zdeňka: **How far should we match UMR?**

JH: within “UMR Light-Contact project” we should be as close as possible.

In other projects, we need UMR as a referential annotation – our next annotation can show what is wrong there and make it better.

For now, do it in the spirit of the guidelines.

Zdeňka: could we join internal Bolder etc. meeting?

JH: Julia said there are some files annotated for English. Could we get an access to some file? – JH can ask Julia.

Dan: Solution: until further notice, we will do it this way…

JH: We are preparing an example file; the proper annotation will follow.

Dan: **How is SynsemClass connected with our annotation?**

JH: Not yet, we will use rather PDT-Vallex for Czech.

**Next steps**

JH: try to annotate a short text

Federica: tried the annotation for Latin, comparing it to the Czech valency files on github (frames)

ŠZ: proposed points of the meetings:

* Presentation of a part of guidelines (as before)
* Presentation of a part of the annotation (Estonian file, Latin by Federica…)
* Discussion on the problems of the annotation

# Meeting 2023-07-10 (Markéta, Dan, Zdeňka, Federica, Cris, Šárka)

1. **Meeting with Julia – summary:**

* Dan: Julia is not using the UMR writer either
* Markéta: the structures disappear there
* Summary: we are not going to master the UMR writer, for the time being

1. **Eventive concepts (a document by Markéta) – comments on the document on Github**

* Markéta: not found any references to semantic type / information packaging
* Dan: information packaging – typological term, very general definition
* Markéta: there are books on information packaging by Bill Croft; Federica has the book Radical Construction Grammar
* The man who is tall = the tall man
* Please, go through the document and comment on the examples
* References about basic terms: Jackendoff, William Croft
* Zdeňka: we can ask Mirjam Fried about construction grammar
* Dan: concentrate on data, not so much on the theories

1. **Entities (a document by Dan) – comments on the document on Github**

* Dan: there are categorial named entities (name of a product – many things); what about the Estonians – nationality in UMR; Novákovi – family in UMR;
* Dan: proper noun – morphological term Dan is using
* When NE should be used?
  + Estonian
  + A general affiliation to a country rather than nationality etc.

If possible, please, go through the document and comment on the examples

* Is connection to a valency lexicon AND Wikipedia possible (like a specific war – a fight would be in a valency lexicon, the name of the war, as a named entity, would be in Wikipedia)

**In general:** we need to collect examples in Czech and other languages and ask Julia

Find a nice sentence with the WW2 (classification event / entity) and ask JH, Julia…

1. Warning: there might be an issue with proper display of Unicode characters. [↑](#footnote-ref-1)
2. In my case … C:\Users\lopatkova\AppData\Local\Programs\Python\Python312

   Tip: when you find the place on our PC, R click to the address field and select copy address :-) [↑](#footnote-ref-2)