Minutes OntoLex F2F Leipzig

Agenda

- Intro by John
- Morph module by Bettina. Motivation, goal, and intro to the new elements proposed so far.
- Coffee break
- Frequency, Attestations and Corpus Information

Discussion

General remarks

- Decide on the name: ontolex, ontolex-lemon, lemon-ontolex, etc. Mailing list.
- Changes on some of the definitions of the Spec

Morphology module

- Problematic cases: Order with rdf:List properties & ontolex:writtenRep

- morph:Affix → inflectional affixes, ontolex:Affix → for derivational. But we still have the :derivational morph value to give people the freedom to describe this as they want.
- The paradigm may contain different patterns. The morph:Morphological Pattern, eg. common noun, will include all the patterns that show in the different paradigms (of common nouns, that is)
- The use of word paradigm vs. pattern leads to confusion. Paradigm as the table, the package of all inflections.
- Table vs. cell:
    - Table: PARADIGM (--> change definition: *a set is not structured*). (Although C Chiarcos thinks that InflectionalParadigm is less ambiguous)
        - Are PARADIGMS specific to the entry? What if what we have in the table is endings? (C Chiarcos) (e.g. example from the Sumerian Dictionary, only the endings provided. How do we represent the generic table?) Bettina: We don't know yet how to call that generic table yet, but IT IS NOT the morphological pattern.
    - Cell: INFLECTION
- [morph:meaning issue] Do we want to say that roots point (in terms of meaning) to something different than stems or affixes? (do we want to differentiate?, how fine-grained do we want to be?) What kind of meaning has a root?

    - Another option: to treat this as a datatype property using strings (we don't have data for this _yet_)
    - Can we "reconstruct" the Lexical Entry from the root? (and access the meaning through the lexical entry?

- [morph grammatical meaning]:
    - We can define grammaticalMeaning as a subproperty, or we can just use reification.
    - Or define everything under your own namespace (so, no grammaticalMeaning property)
    - Overlap with form properties in lexinfo, e.g. #accusativeCaseForm?

- Stems: restrict them to have one single pos? (vs. roots) (discussing the definitions of root and stems)
    - Do we want cardinality restrictions? (better not)
    - Root are not pos-tagged, from the root you can build stems (which do have pos). Roots as lexical entries in some etymological dicts.
    - (Discussion on stems vs. roots)
    - (Discussion) Root as the smallest semantic unit vs. root as the grouping in a dictionary
    - Stem as a concrete morph vs. Root as an abstract from from which we build stems (abstract in the sense that it does not occur in corpus data)
    - Root as the (semantic) nucleus, not further segmented vs. roots as the starting point for stem generation
    - Some affixes belong to the stem in some semitic languages (Ilya Khait)

    - Stem --- a root that has been characterized by another element,leave root definition the way it is
    - Attention: Semitic languages give rise to modelling/terminology issues → Semitic linguistics --- Indo European linguistics (future discussion)
    - What do we do with contractions? (e.g zur (zu + der, German), del (de + el, Spanish)
- [Morph] Order:
    - Alternative to rdf:_1, rdf:_2, → rdf:next

--- Coffee break ----

FrAC (Frequency, Attestations and Corpus Information)

- Logistics: Follow-up on the previous f2f. Development starting mid-July 2019 . Either after or alternating with morphology meetings
- NB: “Corpus” here — structured data collection, not necessarily linguistic corpus, also dictionary, collection of dictionaries, etc.

Frequency

- Frequency (of use of a term in a corpus) is lexicographically relevant, but not relevant only in lexicography (thus it has to be out of the *lexicog* module)
- Frequency is always relative to a corpus hence we need elements to describe these connections
- We want to store two different “types” of frequencies:
    - lexicographic frequencies as found in e.g. dictionaries
    - any extracted frequencies (e.g. as a result of corpus queries)
- Basically any ontolex element can be counted: entries, senses, etc.
    [Action point]. If morph is not a lexicalentry, it should be added to a list of elements on page 13. Maybe there should be a superelement for these
- Important: Normalisation for frequencies, relative counts/frequencies
    But: it can be derived if there is enough metadata
    Also, there can be subcorpora, and we don’t want to store every relative frequency → we want it to be derived by a toolchain
- We will probably want document frequencies as well.
- Method of counting can vary as well.
    CC: we can store it in a text description.
    John: using provenance for that. But it may be insufficient.
- *CorpusFrequency* should be subclassed for different scenarios.
- Is dc:source applicable for a link to a corpus? Maybe some other property should be created. The idea in the proposal was to reuse as much as possible.
- Reliable frequency information is hard to get. A lot of frequencies counted with e.g. SketchEngine is not reliable. See provenance

- One of the application: provide a vocabulary that can be used by an API for a web-service that returns OntoLex-compatible answer
- Alongside integer values for frequencies, it may be useful to say qualitative things (high, low, etc.). But it exists in *lexinfo*
- OntoLex Element (proposed): planned to add it in the new OntoLex specification (1.1)

Attestations

- Suggestion for now: following 2 existing proposals
    - Linking with existing editions
    - Lexicographic properties from lexinfo
- [The thread on the mailing list about this (during the development of the *lexicog* module (2018)), starts with this email: https://lists.w3.org/Archives/Public/public-ontolex/2018Jun/0000.html]
- Do not cover scientific citations
- Let existing models (e.g. WebAnno) deal with links to complicated links

- Clearing things up a little:
    - Citation: bibliographical reference. Within this, we make an Attestation — a specific quote from that citation. *dc:source* can point to this location.
    - We do not specify granularity

- Should a link be *dc:source*? CC wants it to be the same property as for frequencies

- Slides are on github: https://github.com/acoli-repo/ontolex-frac
- Changing an element from *dc:source* — not to mix this up with a source of information: *frac:source*?

- Maybe we need a way to specify a type/genre of citations
    - A characteristic of citation or a source element
    - We could create limitations based on these types for some frequency subclasses
    - *dc:subject* may be a good match. Definitely not a new element

- Attestation locus can be an attestation (IIIF image annotations use-case)
- Citation is not linked to the source, but attestation is. It is possible to link an attestation to a book, but the source is a corpus

Corpus information

- Embeddings:
    - the idea is not to store embeddings in RDF, but to have a vocabulary and be able to get an RDF view on CSV data (on request)
    - use-case: more order in versioning, store metadata
    - lexeme = LexicalEntry on slide 32

- Vectors can be compressed and we need a way to save information about the way they were compressed
- Vectors are always relative to a corpus, source need to be explicit
- The way to store the embedding: maybe JSON?
- Is storing embeddings in the scope of OntoLex at all? Maybe metashare ontology is a better fit?
    - Sense embeddings make sense, because they do represent lexicographical data
    - A lot here concerns metadata, not lexica

- Collocations:
    - Here: co-occurrence information, not lexical choice
    - We want to establish relations between *ontolex:Element*s

Preliminary participants for the telcos:

- John
- CC
- Julia
- Ilya
- Thierry
- Max

Further plans

- Lexicog module:
    - Status report. Final spec finished.
    - Guidelines and best practices → to be done (by community). Someone should take this as editors
        - Thierry will take the lead
        - More participants: Frances, Dorielle, Mustafa
    - The idea for the guidelines: to show when and how it should be used, e.g. where OntoLex-Core suffices
- Morph module
    - Telcos will continue
- FrAC
    - mid-July 2019, see above
- Etymology module (Fahad)
    - Depends on the morphology module
    - Hence should start discussing after it's ready
- Lexinfo
    - Needs a lot of work
        - Revise definitions, properties
        - How to work on that:

            - either with a Google Sheet — probably a better idea (less complex)
            - or smth like WebProtégé
        - Need to define a workflow
    - Updating it to Ontolex

- Ontolex 1.1
    - Non-breaking minor changes
    - The most up-to-date version: https://github.com/cimiano/ontolex

- Standartisation track (ISO or W3C). Is it needed?
- ELEXIS has plans about this
- During eLex there will be a joint OntoLex/TEI meeting (1-3 of October, Sintra)
- There should be 3 W3C members to support this. SAP is planning to do this, KDictionaries as well

Action Points

- [GENERAL] Decide on the name: ontolex, ontolex-lemon, lemon-ontolex, etc. Mailing list.
- [GENERAL] Changes on some of the definitions of the Spec
- [MORPH] How do we represent/what do we call the generic table? (e.g. only with endings) (see C Chiarcos' question above)
- [MORPH] Change definition of _Paradigm_ , remove "structured".
- [MORPH] morph:grammaticalMeaning → delete (proposal). You can create your own properties and values + extend lexinfo, keeping the model minimal
- [MORPH, GENERAL] [Future discussion] Attention: Semitic languages give rise to modelling/terminology issues → Semitic linguistics --- Indo European linguistics. FUTURE: Discussion on this focusing on Semitic languages.
- [MORPH] Proposal: stem --- a root that has been characterized by another element,leave root definition the way it is (J Bournes)
- [FRAC, GENERAL, CORE] **#Discussion #To-Decide** If morph is not a lexicalentry, it should be added to a list of elements on page 13 (elements that can be counted). Alternative: create a general super element (e.g. **ontolex:Element**).
- [FRAC] Start telcos with the preliminary list of participants (from mid-July on)
    - Some things to discuss: the goal, scope and aim (e.g. does this fit into OntoLex?)
    - Basic new elements (e.g. ontolex:Element for the core)