OntoLex Module for Frequency, Attestation and Corpus Information (FrAC)

Supplemental Material

- Wiki: https://www.w3.org/community/ontolex/wiki/Frequency,_Attestation_and_Corpus_Information
- Concept draft: https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md, https://github.com/ontolex/frequency-attestation-corpus-information/ (src)
- Leipzig Face2Face slideshow : https://github.com/acoli-repo/ontolex-frac/blob/master/doc/ontolex-f2f-leipzig-may2019.pdf
- Minutes archive: https://docs.google.com/document/d/1EkhXH2Q-IPHuOX6S64v97BBjbj-WRCWIgmJiW08XLHU/edit?usp=sharing
- Multimodality discussion (evolved out of FrAC): https://docs.google.com/document/d/1-D7JUQhaZWLq7pS2ZIzSQZseq0zGoAISUrFReG527DU/edit?usp=sharing

Agenda 2024-01-31

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/ppz-qekb-vxj [check here for updates!]
- Second link (if needed): https://meet.google.com/moi-fqhj-ccf

Time: 16:00-17:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FS - Flavia Sciolette - ILC-CNR Pisa
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol
IK - Ilan Kernerman
JMc - John McCrae, Galway

DECIDED TO POSTPONE THE CALL

Some minor updates on post-Nexus activities

The draft agenda notes below is to be moved to next call

Status / Sprint

- **DONE@CC**: Diagram updated
- Publication procedure
    - Cf. Emails to CC by Jorge and John [**@CC email 2023-11-22**]
        - Apply HTML template
        - Put on github
        - Jorge would like draft to be put on a W3C page by the end of Nexus

- Sprint
    - **Current plan**
        - we edit the text first, then we convert to HTML new diagram, revision based on local forks/copies
        - coordinate via Slack that sprint is about to start
    - Status [from last call]
        - delays@CC (and others) because of post-Nexus activities (until Jan 24)
        - FS (overall => Dec 18, 2023: comments in)
        - FK (collocations => Jan 9: https://docs.google.com/document/d/148Mtlag7bvl-GCpOpXRxPPQUj1fSTBa7yZvHek0rPQY/edit)
        - KG (attestations)
        - GS (?)
        - Ciprian (similarity)
            - Slack Jan 12th: I managed to go over the Similarity section. From my point of view it is ok. I think that there is a typo, rdf:Bag is written in some places as rdfs:Bag.
        - embeddings?
        - Ilan: overall, towards the end (create comments/suggestions, not directly edit the text)
            - **TO-BE-DONE@CC**: create word document from html
        - John: can help with overall reviewing
    - Coordination via Slack
    - Old notes
        - Formulas are broken images! (**TODO@CC**: check)
        - Question: how to put formulas into HTML
            - https://www.mathjax.org/ ?

TODO@CC: consolidate sprint results, notify John about amnoutzher round of reviews.

- Not before Jan 26
- TBC: next call Jan 31?

Publications

- LREC papers: feedback?
    a. TEI+RDFa https://www.overleaf.com/read/bsvzmrwtmkcx subm
    b. Query proposal https://www.overleaf.com/9259751545ptffmxxswhpz subm
- **NOT YET**: LREC workshops papers (calls expected to be issued in Dec)
    a. Max + Zagreb team: follow-up on datathon, add frequencies and examples
        - RS; maybe for some of LREC workshops
- Journal
    a. Consensus: submit overall description/summary in a journal publication (**TODO@after Christmas**: set milestone, select venue)

- UniDive
    a. 2 abstracts submitted

-

frac:Frequency

To be applied: (no updates)

- Decide naming of current frac:unit (frac:metric, frac:measurement?) => **frac:measure**
- Decide whether frac:tokenFrequency is necessary => **not now**
- Example for relative frequencies of an mwe:

    :x a frac:Collocation; frac:relFreq “0.009888”; frac:head :xyz.
    :xyz a frac:Observable ; frac:frequency [a frac:Frequency; frac:measure “token count”; rdf:value “4356789”].
    Abnsolute frequency
    :x frac:frequency [frac:measure “token count”; rdf:value “5423”].

    Relative frequency of a single observable
    :x frac:frequency [frac:measure “relative frequency”; rdf:value “0.00000000000546”];
    frac:observedIn [dcmi:DataSet; frac:total [a frac:Frequency; frac:measure “token count”; rdf:value “3456789098765434567890”]].

Queries/Patterns

- See notes from earlier meetings, dropped here
- Anything from LREC reviewers?

Agenda 2024-01-17

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/ppz-qekb-vxj [check here for updates!]
- Second link (if needed): https://meet.google.com/moi-fqhj-ccf

Time: 16:00-17:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FS - Flavia Sciolette - ILC-CNR Pisa
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol
IK - Ilan Kernerman
JMc - John McCrae, Galway

Status / Sprint

- **DONE@CC**: Diagram updated
- Publication procedure
    - Cf. Emails to CC by Jorge and John [**@CC email 2023-11-22**]
        - Apply HTML template
        - Put on github
        - Jorge would like draft to be put on a W3C page by the end of Nexus
- Sprint
    - **Original plan**
        - **@CC**: reread and eliminate old examples (incl: check Dutch (Diamant) example, check Old English example), convert to HTML, then collect feedback from sprinters
    - **Dec 14, 2023@CC**
        - we edit the text first, then we convert to HTML new diagram, revision based on local forks/copies
        - coordinate via Slack that sprint is about to start
    - Status
        - delays@CC (and others) because of post-Nexus activities (until Jan 24)
        - FS (overall => Dec 18, 2023: comments in)
        - FK (collocations => Jan 9: https://docs.google.com/document/d/148Mtlag7bvl-GCpOpXRxPPQUj1fSTBa7yZvHek0rPQY/edit)
        - KG (attestations)
        - GS (?)
        - Ciprian (similarity)

            - Slack Jan 12th: I managed to go over the Similarity section. From my point of view it is ok. I think that there is a typo, rdf:Bag is written in some places as rdfs:Bag.
        - embeddings?
        - Ilan: overall, towards the end (create comments/suggestions, not directly edit the text)
            - **TO-BE-DONE@CC**: create word document from html
        - John: can help with overall reviewing
    - Coordination via Slack
    - Old notes
        - Formulas are broken images! (**TODO@CC**: check)
        - Question: how to put formulas into HTML
            - https://www.mathjax.org/ ?

TODO@CC: consolidate sprint results, notify John about amnoutzher round of reviews.

- Not before Jan 26
- TBC: next call Jan 31?

Publications

- LREC papers: feedback in about 2 weeks
    a. TEI+RDFa https://www.overleaf.com/read/bsvzmrwtmkcx subm
    b. Query proposal https://www.overleaf.com/9259751545ptffmxxswhpz subm
- **NOT YET**: LREC workshops papers (calls expected to be issued in Dec)
    a. Max + Zagreb team: follow-up on datathon, add frequencies and examples
        - RS; maybe for some of LREC workshops
- Journal
    a. Consensus: submit overall description/summary in a journal publication (**TODO@after Christmas**: set milestone, select venue)
- UniDive
    a. 2 abstracts submitted

Sprint

-

frac:Frequency

Discussed last time:

- Decide naming of current frac:unit (frac:metric, frac:measurement?) => **frac:measure**
- Decide whether frac:tokenFrequency is necessary => **not now**
- Example for relative frequencies of an mwe:

    :x a frac:Collocation; frac:relFreq “0.009888”; frac:head :xyz.
    :xyz a frac:Observable ; frac:frequency [a frac:Frequency; frac:measure “token count”; rdf:value “4356789”].
    Abnsolute frequency
    :x frac:frequency [frac:measure “token count”; rdf:value “5423”].

    Relative frequency of a single observable
    :x frac:frequency [frac:measure “relative frequency”; rdf:value “0.00000000000546”];
    frac:observedIn [dcmi:DataSet; frac:total [a frac:Frequency; frac:measure “token count”; rdf:value “3456789098765434567890”]].

Queries/Patterns [excluded for the moment]

- See notes from earlier meetings, dropped here

Agenda 2023-12-06

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/ppz-qekb-vxj [check here for updates!]
- Second link (if needed): https://meet.google.com/moi-fqhj-ccf

Time: 16:00-17:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FS - Flavia Sciolette - ILC-CNR Pisa
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble

RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol
IK - Ilan Kernerman

Status

- FrAC presentation at W3C day at LDK
    - **NOT DONE@CC**: upload slides (postponed, in order to keep anonymity for LREC paper)
- **TODO@CC**: Diagram needs to be updated
    - Revised frac:total, introduced frac:unit (already in text)
- Publication plan
    - Official draft by end of this year?
    - ask Jorge and John about procedure [**DONE@CC, Email 2023-11-22**]
    - **TODO@CC**: today-Friday:: reread and eliminate old examples
        - Old minutes on that:
            - check Dutch (Diamant) example, check Old English example
        - TODO: apply HTML template and send notification about that
    - **TODO@CC**: rend out reminders when sprint is about to start
    - Sprint weeks December 11th - January 10th (re-read, iron out issues), volunteers:
        - **FS**: next week(overall)
        - As of end of next week, in your own forks (edit directly)
            - **FK**: => Collocations
            - **KG:** => Attestations
            - **GS**: not before Christmas
            - **Ciprian**: => (Embedding) Similarity [maybe]
            - In Parallel: maintain slack for discussing larger changes
        - **Ilan**: overall reading after next week (after Flavia, but create comments/suggestions, not directly edit the text
            - Using track changes in a Google doc
            - **TODO@CC**: create word document from html
        - Post-call updates
          Dec 14, 2023@CC: we edit the text first, then we convert to HTML
          Dec 18, 2023@Flavia: comments in
        - **Next call: Jan 10, status update**
        - Coordination via Slack
    - Formulas are broken images!

        - **TODO@CC**: check
        - Question: how to put formulas int HTML
            - https://www.mathjax.org/ ?

Publications

- **NO FEEDBACK YET:** LREC papers
    a. TEI+RDFa https://www.overleaf.com/read/bsvzmrwtmkcx subm
    b. Query proposal https://www.overleaf.com/9259751545ptffmxxswhpz subm
- **NOT YET**: LREC workshops papers (calls expected to be issued in Dec)
    a. Max + Zagreb team: follow-up on datathon, add frequencies and examples
        - RS; maybe for some of LREC workshops
- Journal
    a. Consensus: submit overall description/summary in a journal publication (**TODO@after Christmas**: set milestone, select venue)

DataCube

- https://www.w3.org/TR/vocab-data-cube/
- Vocabulary for matrix data, mostly used for geodata, some overlap in intent
    - **TBD@CC+GS**: discuss/prep example, WHEN
- Doc: https://docs.google.com/document/d/1VQ40Lq-X24xc7mY1BkNOsyf53D7HoEdY2y72anQR5DY/edit?usp=sharing
- As it became clear last time that the mapping with DataCube isn't fully straight-forward, we do not revise the model but put a DataCube interpretation into a non-normative section. Maybe along with SPARQL rules to generate a DataCube view from FrAC data.
- GS: agreed

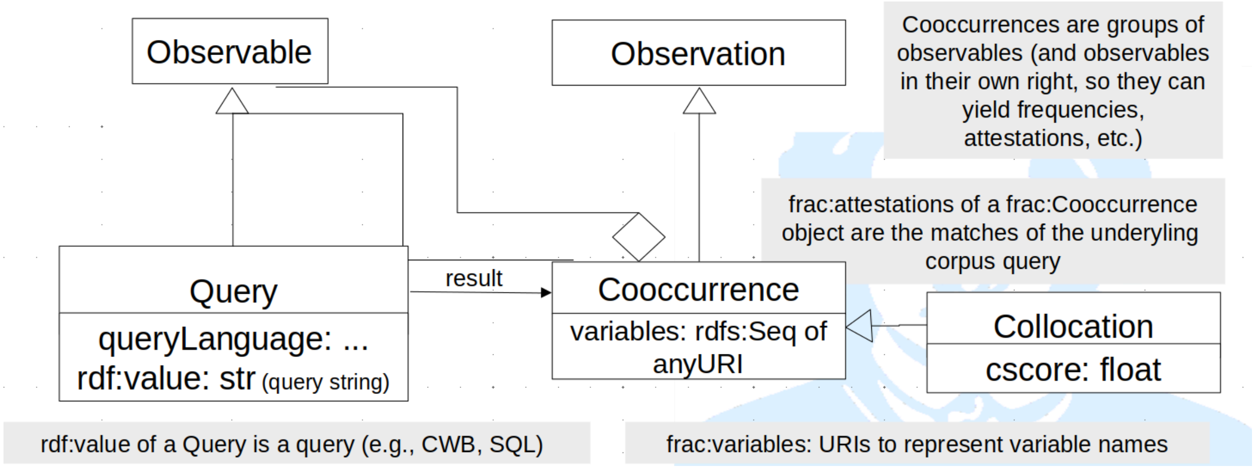
frac:Frequency

Discussed last time:

- Feature request: slightly extend the scope of frac:unit to also be able to provide frequency scores (like TF/IDF)
    - otherwise, there is an asymmetry in that we can provide such scores for MWEs but not for single words
    - Could be done with frac:unit, *if its definition is extended accordingly and the naming is adjusted*
    - Would eliminate the restriction of frac:Frequency to absolute frequencies
    - Possible naming: not decided yet, should cover both (units for) absolute frequencies and derived scores

        - Frac:metric
        - Frac:measurement
        - Frac:measure => this one
        - NB: to replace unit, we need to write token frequency, then
- GS: freqs without blank node?
    - CC: not with OWL semantics, different property frac:tokenFrequency?
- **TODO@TODAY**:
    - Decide naming of current frac:unit (frac:metric, frac:measurement?) => **frac:measure**
    - Decide whether frac:tokenFrequency is necessary => **not now**
        - GS: maybe as a recipe, if OWL/SHex transformation is possible
        - Postponed into official review period

Queries/Patterns [excluded for the moment]

- consolidated in paper proposal, to be discussed here after Oct 20
- Paper authors: discuss?
- LDK2023 version:
- 

Cooccurrences are groups of observables (and observables in their own right, so they can yield frequencies, attestations, etc.)

frac:attestations of a frac:Cooccurrence object are the matches of the underlying corpus query

rdf:value of a Query is a query (e.g., CWB, SQL)

frac:variables: URIs to represent variable names

TODO@RS: rename QUeryResult to Query ?
Other changes?

Agenda 2023-11-22

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/ppz-qekb-vxj [check here for updates!]
- Second link (if needed): https://meet.google.com/moi-fqhj-ccf

Time: 16:00-17:00 CET

Draft: Github

Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FS - Flavia Sciolette - ILC-CNR Pisa
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Status

- FrAC presentation at W3C day at LDK
    - **NOT DONE@CC**: upload slides (postponed, in order to keep anonymity for LREC paper)
- **TODO@CC**: Diagram needs to be updated
    - Revised frac:total, introduced frac:unit (already in text)
- **TODO:** Find reviewers/readers/guinea pigs for reading FrAC draft
    - LiLa people?
- Publication plan
    - Official draft by end of this year?
    - ask Jorge and John about procedure [**DONE@CC, Email 2023-11-22**]
    - **TODO@CC**: after Wed Nov 29: reread and eliminate old examples
        - Old minutes on that:
            - check Dutch (Diamant) example, check Old English example
    - Sprint week Dec 4th-8th (re-read, iron out issues), volunteers:
        - **FK, FS, (KG)**
        - Coordinartion via Slack
        - Progress report by Dec 6th

Publications

- Status MWE chapter
    a. Submitted with minor revisions (by Max, Nov 15)

- LREC workshop (LDL)
    - a. **ACCEPTED** Nov 20 (Details to Max, only, fullday)
        - confirmation email copied to overleaf (unidive.tex)
        - Coordinate via Slack
- **NOT YET:** LREC papers
    - a. TEI+RDFa https://www.overleaf.com/read/bsvzmrwtmkcx subm
    - b. Query proposal https://www.overleaf.com/9259751545ptffmxxswhpz subm
- **UNIDIVE** call > https://unidive.lisn.upsaclay.fr/doku.php?id=meetings:general_meetings:2nd_unidive_general_meeting_call_for_abstracts
    - a. Leveraging Linked Data, NIF, and CONLLU for Enhanced Annotation in Aligned Parallel Corpora, https://www.overleaf.com/8127479297nzrmtngdfjdf#620f76
- **NOT YET**: LREC workshops papers (calls expected to be issued in Dec)
    - a. Max + Zagreb team: follow-up on datathon, add frequencies and examples
        - RS; maybe for some of LREC workshops
- Journal
    - a. Consensus: submit overall description/summary in a journal publication (**TODO@after Christmas**: set milestone, select venue)

DataCube

- https://www.w3.org/TR/vocab-data-cube/
- Vocabulary for matrix data, mostly used for geodata, some overlap in intent
    - **TBD@CC+GS**: discuss/prep example, WHEN
- Doc: https://docs.google.com/document/d/1VQ40Lq-X24xc7mY1BkNOsyf53D7HoEdY2y72anQR5DY/edit?usp=sharing
- As it became clear last time that the mapping with DataCube isn't fully straight-forward, we do not revise the model but put a DataCube interpretation into a non-normative section. Maybe along with SPARQL rules to generate a DataCube view from FrAC data.
- GS: agreed

TF/IDF for single words

Proposed by Flavia:

- For MWEs, could be done with frac:cscore
- For single words? (not a cooccurrence)
- Idea: use frac:unit for that and rename it to frac:metric ?
- Example:

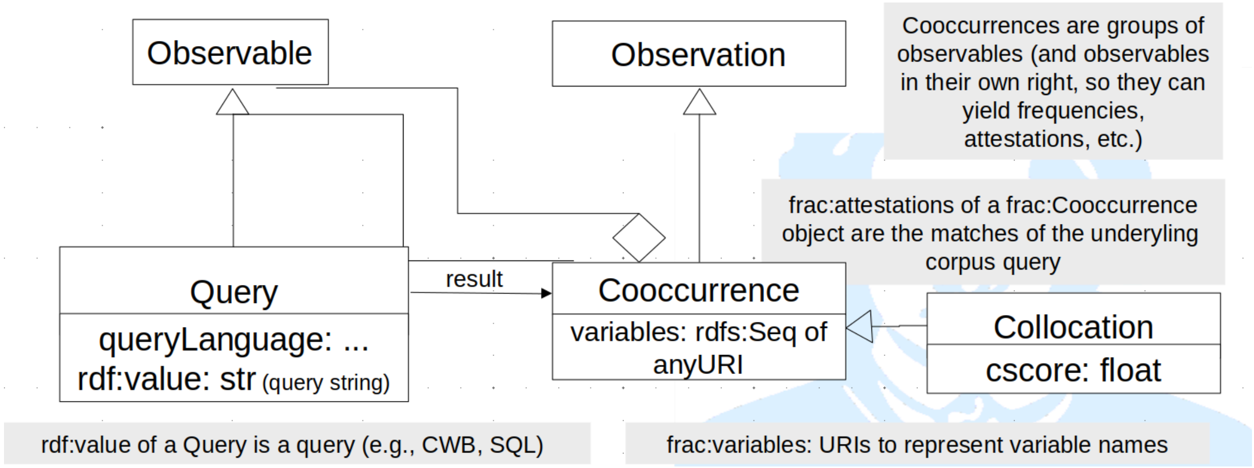
```
:shema a ontolex:lexicalEntry ;
```

```
        frac:frequency[
                a frac:Frequency;
                rdf: value "206"^^xsd:int;
                frac: observedIn <#traduco_berakhot>
        ] ;
        frac:frequency[
                a frac:Frequency; frac:unit "TF-IDF" ;
                rdf:value "0.0204743790070761"
                ].
:shema_lemma_att_1 a frac:Attestation ;
          frac:quotation "Questo primo versetto è la lettura dello Shemà." .
```

- Frac:unit allows to provides a normalized frequency, then
    - Would also include relative frequency
    - GS: weird name, better measurement
        - Frac:metric?
        - Frac:measurement?
- GS: freqs without blank node?
    - CC: not with OWL semantics, different property frac:tokenFrequency?
- **TODO@next call**:
    - Decide naming of current frac:unit (frac:metric, frac:measurement?)
    - Decide whether frac:tokenFrequency is necessary
        - GS: non-normative informatuion on how to model a datadpropertym that is to be interpreted as an object property of FrAC

Queries/Patterns [POSTPONED]

- consolidated in paper proposal, to be discussed here after Oct 20
- Paper authors: discuss?
- LDK2023 version:
-



Cooccurrences are groups of observables (and observables in their own right, so they can yield frequencies, attestations, etc.)

frac:attestations of a frac:Cooccurrence object are the matches of the underlying corpus query

rdf:value of a Query is a query (e.g., CWB, SQL)

frac:variables: URIs to represent variable names

TODO@RS: rename QUeryResult to Query ?
Other changes?

Agenda 2023-11-09

Canceled
Next call Wed, 2023-12-06 16:00 CET (Berlin time)

Agenda 2023-10-26

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/ppz-qekb-vxj [check here for updates!]
- Second link (if needed): https://meet.google.com/szp-kamy-ied

Time:	14:00–15:00 CET
Draft:	Github
Diagram	https://github.com/ontolex/frequency-attestation-corpus-information note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FS - Flavia Sciolette - ILC-CNR Pisa
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Status

- FrAC presentation at W3C day at LDK
    - **TODO@CC**: upload slides [not done yet, apparently, no upload of PDFs to W3C wiki]
- **TODO@CC**: Diagram needs to be updated
    - Revised frac:total, introduced frac:unit (already in text)

- **DONE@CC**: Doodle for new slot: https://doodle.com/meeting/participate/id/egk2APDe
    - New slot: Wed 16:00-17:00 Berlin time, until end of year
- **TODO@CC**:
    - send email about new slot
- **TODO:** Find reviewers/readers/guinea pigs for reading FrAC draft
    - LiLa people?

Publications

- Status MWE chapter?
    a. Accepted with minor revisions
    b. to be submitted on 31st of October….can the authors who haven't done so yet please proofread the text before that? (Katerina has already fixed the minor errors listed in the document but we need to check the grammar together)
    c. **TODO@BK:** proof-read text (grammar, readability) within the next 2 days
    d. **TODO@CC**: proof-read text (grammar, readability), no earlier than Monday night
    e. **TODO@MI**: submit final version (Fahad can't)
        - fallback : Kateria
        - **TODO@KG**: ask whether he'd do it ;)
            - **TBC**: how to submit (email?)
- LREC
    a. **TEI+RDFa** https://www.overleaf.com/read/bsvzmrwtmkcx subm
    b. Query proposal https://www.overleaf.com/9259751545ptffmxxswhpz subm
    c. **LDL workshop** proposal submitted
- LREC workshops?
    a. **Max + Zagreb team:** follow-up on datathon, add frequencies and examples
        - RS; maybe for some of LREC workshops
    b. **?** Middle Persian@Cologne to LREC?
        - MI: maybe using FrAC (and maybe Morph) for the Middle Persian
        - Potential contributors: MI, CC, FK + people from MPCD
- Journal
    a. **Consensus**: submit overall description/summary in a journal publication (milestone tbc. after next telco)

DataCube

- https://www.w3.org/TR/vocab-data-cube/
- Vocabulary for matrix data, mostly used for geodata, some overlap in intent
    - GS: library for export into pandas
    - CC: we need an exa,mple t properly judge
    - CC+GS: work out possible bridge, maybe in an LREC workshop paper

    - CC: interesting, if compatibility confirmed (or can be affirmed) diagram quite complicated, maybe in an appendix, if included in community report, but may guide (re-)design
    - MI: inclusion into FrAC requires some level of maturity on a technical level. How well supported?
    - CC+GS: discuss/prep example, not earlier than in 2 weeks
- Doc: https://docs.google.com/document/d/1VQ40Lq-X24xc7mY1BkNOsyf53D7HoEdY2y72anQR5DY/edit?usp=sharing
- BK: sketchengine might give relative frequencies only
    - GS: can be computed from absolute frequencies; relative: relative to what?
- frac:Observation ~ qb:Observation, but multiple metrics
    - One qb:observation can have multiple measures (e.g., one sensor for pressure and temperature)
- CC: more compact for rdf:values of frac:Frequency (unit incorporated in property)
- GS: reversed point of view
- Call ended,. To be continued next time ;)
    - CC: Unfortunately, I can’t today :([I’m sorry]
        - CC: See you soon!

AoB

Next call: see Doodle resultt

POSTPONED

- Queries/Patterns
- TF/IDF for single words (=> revising frac:uniq)
- Consolidation: attestation

Agenda 2023-10-12

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/ppz-qekb-vxj [check here for updates!]
- Second link (if needed): https://meet.google.com/szp-kamy-ied

Time:	14:00–15:00 CET
Draft:	Github
Diagram	https://github.com/ontolex/frequency-attestation-corpus-information note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FS - Flavia Sciolette - ILC-CNR Pisa
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Status

- FrAC presentation at W3C day at LDK
  - **TODO@CC**: upload slides [not done yet, apparently, no upload of PDFs to W3C wiki]
- **TODO@CC**: Diagram needs to be updated
  - Revised frac:total, introduced frac:unit (already in text)
- **DONE@CC**: Doodle for new slot: https://doodle.com/meeting/participate/id/egk2APDe

Publications

- Status MWE chapter?
  a. still under review
- LREC, deadline Oct 13 -> Oct 20

a. **TEI+RDFa** [FK, GS, CC, MI, contributors welcome, talk to FK]: https://www.overleaf.com/read/bsvzmrwtmkcx
    - Need input from other authors, esp. On RDFa
    - Todo: anonymize code excerpts (anonymous github repo)

b. **Max + Zagreb team:** follow-up on datathon, add frequencies and examples
    - RS; maybe for some of LREC workshops

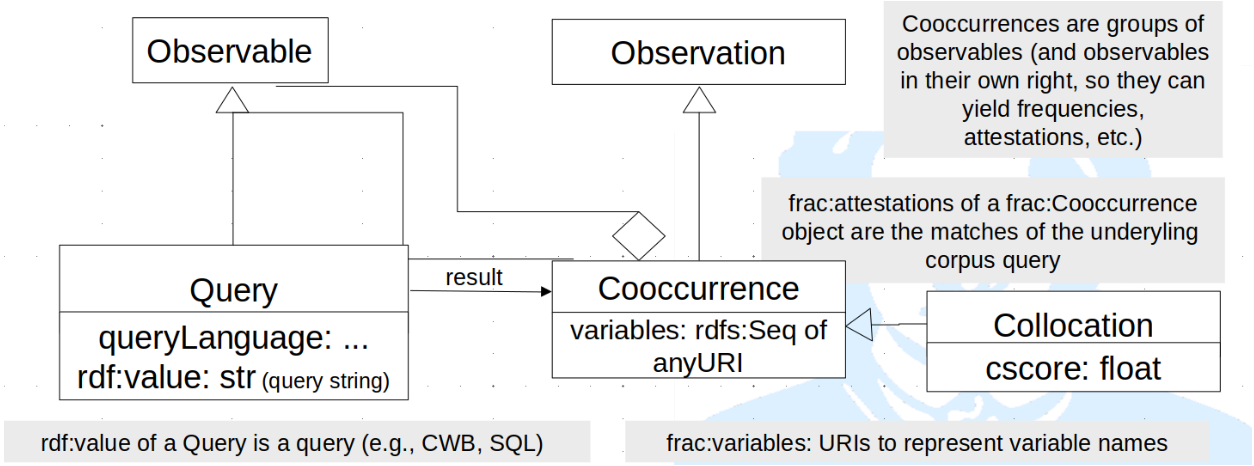
c. **[NOT]** Middle Persian@Cologne to LREC?
    - MI: maybe using FrAC (and maybe Morph) for the Middle Persian
    - Potential contributors: MI, CC, FK + people from MPCD

d. Query proposal
    - Contributors: CC, RS, MI; GS (for checking modelling)
    - https://www.overleaf.com/9259751545ptffmxxswhpz
    - RS: examples for att and fre; nosketch: API limited , details see below

- **Consensus**: submit overall description/summary in a journal publication (not before LREC-COLING)
- Find reviewers/readers/guinea pigs for reading FrAC draft
    a. LiLa people?

Queries/Patterns

- consolidated in paper proposal, to be discussed here after Oct 20
- Paper authors: discuss?
- LDK2023 version:
-



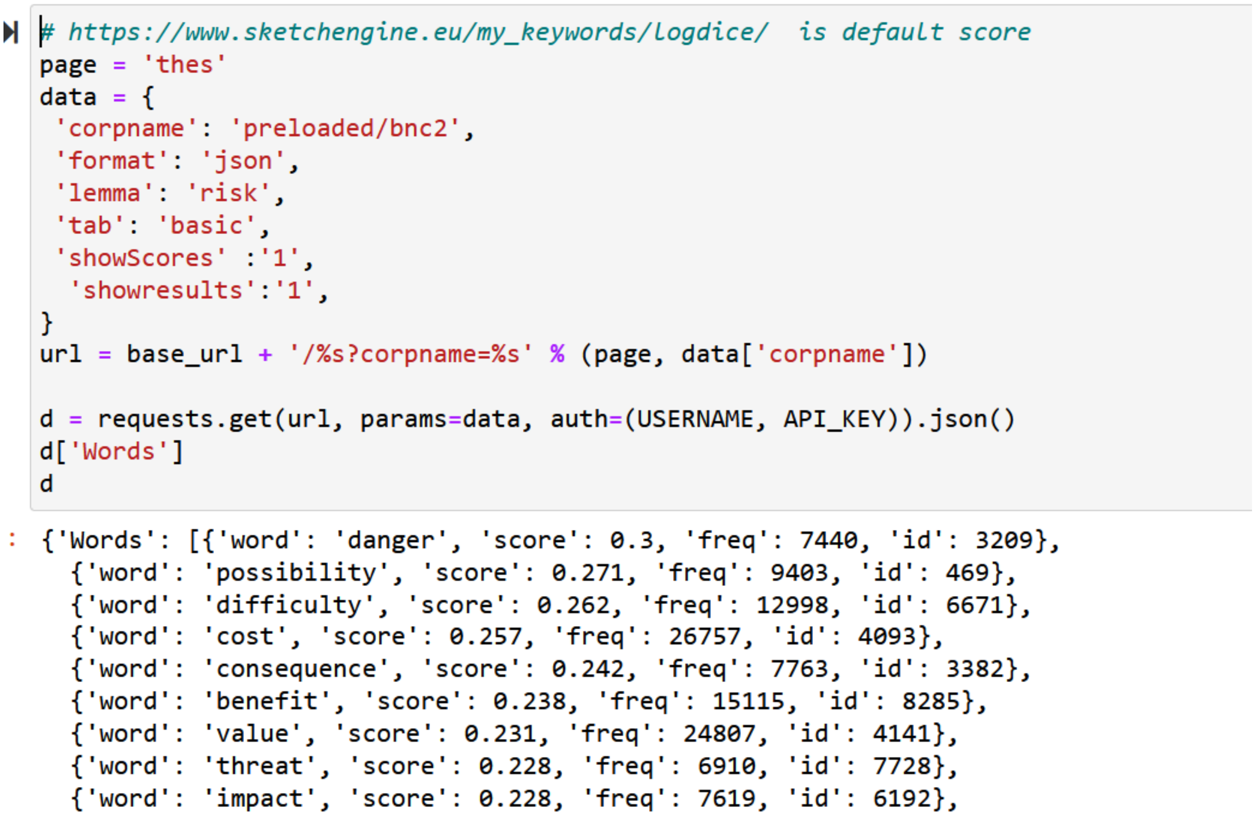
Cooccurrences are groups of observables (and observables in their own right, so they can yield frequencies, attestations, etc.)

frac:attestations of a frac:Cooccurrence object are the matches of the underlying corpus query

rdf:value of a Query is a query (e.g., CWB, SQL)

frac:variables: URIs to represent variable names

TODO@RS: rename QUery to QueryResult



```python
# https://www.sketchengine.eu/my_keywords/logdice/  is default score
page = 'thes'
data = {
 'corpname': 'preloaded/bnc2',
 'format': 'json',
 'lemma': 'risk',
 'tab': 'basic',
 'showScores' :'1',
  'showresults':'1',
}
url = base_url + '/%s?corpname=%s' % (page, data['corpname'])

d = requests.get(url, params=data, auth=(USERNAME, API_KEY)).json()
d['Words']
d
```

```
{'Words': [{'word': 'danger', 'score': 0.3, 'freq': 7440, 'id': 3209},
  {'word': 'possibility', 'score': 0.271, 'freq': 9403, 'id': 469},
  {'word': 'difficulty', 'score': 0.262, 'freq': 12998, 'id': 6671},
  {'word': 'cost', 'score': 0.257, 'freq': 26757, 'id': 4093},
  {'word': 'consequence', 'score': 0.242, 'freq': 7763, 'id': 3382},
  {'word': 'benefit', 'score': 0.238, 'freq': 15115, 'id': 8285},
  {'word': 'value', 'score': 0.231, 'freq': 24807, 'id': 4141},
  {'word': 'threat', 'score': 0.228, 'freq': 6910, 'id': 7728},
  {'word': 'impact', 'score': 0.228, 'freq': 7619, 'id': 6192},
```

Sketchengine, requires account (issue with replication?) (no query there), only freq and attestation

Below: NoSketchEngine example, incl. query

```
:q_att_SrFudKo_fm_fudbal_4378892 a frac:QueryResult;
                        rdf:value """<s/> containing [lc =\"fudbal\" & tag=\"N.*\"]"""^^xsd:string;
                        rdfs:annotation
"""https://noske.rgf.rs/run.cgi/first?corpname=SrFudKo&tab=advanced&queryselector=cql&attrs=word&viewmode=sen&attr_allpos=all&refs_up=0&shorten_refs=1&glue=1&gdex_enabled=1&gdexcnt=10&show_gdex_scores=1&itemsPerPage=20&structs=s,g&refs=doc&default_attr=lemma&cql=<s/> containing [lc =\"fudbal\" & tag=\"N.*\"]&showresults=1&showTBL=0&tbl_template=&gdexconf=&f_tab=basic&f_showrelfrq=1&f_showperc=0&f_showreldens=0&f_showreltt=0&c_customrange=0&operations=[{\"name\":\"cql\",\"arg\":<s/> containing [lc =\"fudbal\" & tag=\"N.*\"],\"query\":{\"queryselector\":\"cqlrow\",\"cql\":<s/> containing [lc =\"fudbal\" & tag=\"N.*\"],\"default_attr\":\"lemma\"},\"id\":1750}]""";
                        owl:sameAs
<https://noske.rgf.rs/run.cgi/first?corpname=SrFudKo&tab=advanced&queryselector=cql&attrs=word&viewmode=sen&attr_allpos=all&refs_up=0&shorten_refs=1&glue=1&gdex_enabled=1&gd
```

```
excnt=10&show_gdex_scores=1&itemsPerPage=20&structs=s,g&refs=doc&default_attr=lemma
&cql=%3Cs/%3E%20containing%20[lc%20=%22fudbal%22%20&%20tag=%22N.*%22]&showr
esults=1&showTBL=0&tbl_template=&gdexconf=&f_tab=basic&f_showrelfrq=1&f_showperc=0&
f_showreldens=0&f_showreltt=0&c_customrange=0&operations=[%7B%22name%22:%22cql%
22,%22arg%22:%3Cs/%3E%20containing%20[lc%20=%22fudbal%22%20&%20tag=%22N.*%2
2],%22query%22:%7B%22queryselector%22:%22cqlrow%22,%22cql%22:%3Cs/%3E%20conta
ining%20[lc%20=%22fudbal%22%20&%20tag=%22N.*%22],%22default_attr%22:%22lemma%
22%7D,%22id%22:1750%7D]>;
                    frac:queryLanguage
<https://www.sketchengine.eu/documentation/corpus-querying/>;
                    frac:result [a frac:Cooccurrence ;
                            frac:attestation :att_1_SrFudKo_fm_fudbal_4378892 ;
                            frac:attestation :att_2_SrFudKo_fm_fudbal_4378892 ;
                            frac:attestation :att_3_SrFudKo_fm_fudbal_4378892].
```

A bit too clumsy for paper

- CC: decode query string?
- MI: looks into that in paper
- motivATION DIG LEX: MI+CC?
- RS: use case context (1x sketch endinge on BNC), project context (1x nosketch and serbian)

DataCube

- https://www.w3.org/TR/vocab-data-cube/
- Vocabulary for matrix data, mostly used for geodata, some overlap in intent
    - GS: library for export into pandas
    - CC: we need an exa,mple t properly judge
    - CC+GS: work out possible bridge, maybe in an LREC workshop paper
    - CC: interesting, if compatibility confirmed (or can be affirmed) diagram quite complicated, maybe in an appendix, if included in community report, but may guide (re-)design
    - MI: inclusion into FrAC requires some level of maturity on a technical level. How well supported?
    - CC+GS: discuss/prep example, not earlier than in 2 weeks

Consolidation

- Check vs:status
- frac:Observable: approved
- frac:observedIn
    - Gilles: cf. DataCube vocabulary for observations => “I belong to the datacube that takes this kind of dimensions”, e.g., lexical entries x corpora

        - NB: also related to frac:unit, frac:total
        - DataCube: set of observations. Several data cubes for the same corpus
    - **TODO@Gilles**: propose a modelling of some FrAC use cases for comparison

Attestations

- CC: Diamant example is incomplete and not reconstructable
    - Needs to be replaced
- **TO-BE-DONE@FK**: look into example after https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#gloss-property
    - Dug out literature referecnes for old english examples, see comments *below*

AoB

Next call: see Doodle results

Agenda 2023-09-28

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/ppz-qekb-vxj [check here for updates!]
- Second link (if needed):

Time: 14:00–15:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FS - Flavia Sciolette - ILC-CNR Pisa
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Intros

- Flavia Sciolette, https://www.ilc.cnr.it/en/people/flavia-sciolette/

Status

- Was there a call on 2023-07-20? No, cancelled
- FrAC presentation at W3C day at LDK
    - TODO@CC: upload slides
    - Feedback:
        - Disc whether queries? Consensus: if major delay, we don't do it
        - Recording to be checked ;)
- **TODO@CC**: Diagram needs to be updated
    - Revised frac:total, introduced frac:unit (already in text)
- **DONE@Ciprian**: standardize cscore subproperties for lowercase camelCase

- **TODO@CC**: Doodle for new slot
    - Ranka: not Wed 12-14, Thu 12-14
    - Ciprian: not Tue, Wed, Thu, 13-19:00, resp.
    - Gilles: a lot ;) Maybe not Mon, Tue, Fri
    - Christian: not Monday, different slots
    - Flavia: teaching slots not confirmed, yet, pref. Thu, Fri
    - Elena: can wed, thu morning

Publications

- Status MWE chapter?
    a. First round of reviews, resubmitted (by Max?), under review
- Use case papers
    a. General approach
        - decide about a venue for presenting, collect a team of co-authors, have designated calls, report at regular telcos
- LREC, deadline Oct 13
    a. follow up on TEI+RDFa (w. Gilles), see notes from 2023-05-11
        - FK,GS,CC,MI, based on Graz abstract
        - **DONE@FK**: coordination email meeting
    b. **Max + Zagreb team:** follow-up on datathon, add frequencies and examples
    c. **[status?]** Arabic corpora as a use case for FrAC+Morph
        - Proposed by FK
        - Potential contributors: FK, CC, MI
        - Potential venues: very early stage (LREC-COLING 2024?)
        - Unlikely to happen until LREC, need more native input
    d. **[status?]** Middle Persian@Cologne to LREC?
        - MI: maybe using FrAC (and maybe Morph) for the Middle Persian
        - Potential contributors: MI, CC, FK + people from MPCD
    e. Query proposal
        - Contributors: CC, RS, ?MI; GS (for checking modelling)
        - **TODO@RS**: sample data by end of next week
        - **TODO@CC**: start overleaf, put link here
            - **DONE 2023-09-29:** https://www.overleaf.com/9259751545ptffmxxswhpz
    f. Ciprian: writing something to use morph along with FrAC
        => MWE chapter
- **Consensus**: submit overall description/summary try a journal publication
    a. Last COLING already overview on FrAC, no novelty here except for queries

Queries/Patterns

- Corpus queries@LDK: presentes as an extension under discussion, not as a result, yet

- CC: minor renaming for LDK presentation, see slides there
- **DONE@Ranka**: check whether SketchEngine provides free API(!) access to brown corpus for collocation analysis (to illustrate query extension to FrAC)

  https://www.sketchengine.eu/brown-corpus/

  Corpus freely browseable, but not via API, needs API key

  cat >bl.wl << EOF and the of in on at a an to that is EOF curl -F "wlsort=frq" -F "wlattr=word" -F "format=json" -H "Content-Type: multipart/form-data" -F "wlblacklist=@bl.wl" --user "USERNAME:APIKEY" "https://api.sketchengine.eu/bonito/run.cgi/wordlist?corpname=preloaded/bnc2_tt21" > result.json
    - Ranka got an API key, also asking for API key
- Detailed discussion in LREC paper prloposal, submission basis for discussion after Oct 15

Consolidation

- Check vs:status
- frac:Observable: approved
- frac:observedIn
  - Gilles: cf. DataCube vocabulary for observations => “I belong to the datacube that takes this kind of dimensions”, e.g., lexical entries x corpora
    - NB: also related to frac:unit, frac:total
    - DataCube: set of observations. Several data cubes for the same corpus
  - **TODO@Gilles**: propose a modelling of some FrAC use cases for comparison

Attestations

- CC: Diamant example is incomplete and not reconstructable
  - Needs to be replaced
- **TO-BE-DONE@FK**: look into example after https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#gloss-property
  - Dug out literature referecnes for old english examples, see comments *below*

AoB

Next call: ???

Agenda 2023-07-06

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/tru-zxuj-cyv
- Second link (if needed):

Time: 14:00–15:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Publications

- Updates on the MWE chapter?
    a. On it
- Use case papers
    a. General approach
        - decide about a venue for presenting, collect a team of co-authors, have designated calls, report at regular telcos
    b. **[pushed back to July]:** follow up on TEI+RDFa (w. Gilles), see notes from 2023-05-11
        - FK,GS,CC,MI, based on Graz abstract
        - TODO@FK: coordinate meeting in Sep
    c. **[status?]** Arabic corpora as a use case for FrAC+Morph
        - Proposed by FK
        - Potential contributors: FK, CC, MI
        - Potential venues: very early stage (LREC-COLING 2024?)

        - Unlikely to happen until LREC, need mnore native input
    d. **[status?]** Middle Persian@Cologne to LREC?
        - MI: maybe using FrAC (and maybe Morph) for the Middle Persian
        - Potential contributors: MI, CC, FK + people from MPCD
- To which paper does that belong?
    a. **Consensus**: submit to LREC-COLING (as a poster), then (or in parallel) try a journal publication
    b. When is LREC deadline?
        - Oct 13

Consolidation

No updates on the diagram / model, no updates to/at github

Attestations

- Check vs:status
- CC: Diamant example is incomplete and not reconstructable
    - Needs to be replaced
- **TODO@FK**: look into example after https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#gloss-property
    - Dug out literature referecnes for old english examples, see comments *below*

Frac:total / corpus class and property

Suggestion for corpus:

- Replace frac:corpus with frac:observedIn
- Abandon frac:Corpus, instead, require that the object of frac:observedIn should be assigned a DCMIType, give Dataset, Collection and Text as examples
- Proböem_ frac:total doesn't have a name anymore, w ecould just say that range is any member of DCMIType
- Frac:locus would be unaffected

Suggestuon for total:

- Point to frequency, instead of number
- Create new property unit for frequency objects

Discussed and merged

- TODO@CC: diagram

Other issues

- TODO@Ciprian: standardize cscore subproperties for lowercase camelVCase (DONE)
- TODO@Ranka: check whether SketchEngine provides free API(!) access to brown corpus for collocation analysis (to illkustrate query extensioin to FrAC)
  https://www.sketchengine.eu/brown-corpus/
- Corpus queries@LDK: present this as an extension under discussion, not as a result, yet
- FK: W3C Day session coordination meetings needed (bilateral, only), won't be available in August

AoB

Next call: 2023-07-20

Agenda 2023-06-22

Telco link (please check here for updated link):

- **One-time** link:https://meet.google.com/zcp-dkjq-rcd
- Second link (if needed): https://meet.google.com/jzn-uxzz-hzf

Time:	14:00–15:00 CET
Draft:	Github
Diagram	https://github.com/ontolex/frequency-attestation-corpus-information note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Aposto

Regrets:

Katerina Gkirtzou
Max Ionov

Publications

- Updates on the MWE chapter?
    a. CC: Yes, response totally overlooked, deadline July 8
- Use case papers
    a. General approach
        - decide about a venue for presenting, collect a team of co-authors, have designated calls, report at regular telcos
    b. **[pushed back to July]:** follow up on TEI+RDFa (w. Gilles), see notes from 2023-05-11
    c. **[status?]** Arabic corpora as a use case for FrAC+Morph
        - Proposed by FK
        - Potential contributors: FK, CC, MI
        - Potential venues: very early stage (LREC-COLING 2024?)
    d. **[status?]** Middle Persian@Cologne to LREC?
        - MI: maybe using FrAC (and maybe Morph) for the Middle Persian
        - Potential contributors: MI, CC, FK + people from MPCD
- To which paper does that belong?
    a. **Consensus**: submit to LREC-COLING (as a poster), then (or in parallel) try a journal publication

Consolidation

No updates on the diagram / model, no updates to/at github

Attestations

- Check vs:status
- CC: Diamant example is incomplete and not reconstructable
    - Needs to be replaced
- **TODO@FK**: look into example after https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#gloss-property

Corpus queries — postponed (CC, RS, MI)

More work and discussions needed before we can talk about this. At the next meeting we will have some ideas/results

- **[postponed until MI is back]**

frac:total

Open issue: https://github.com/ontolex/frequency-attestation-corpus-information/issues/5
Domain declaration of frac:total has been the motivation to introduce corpus class

- **Consensus**: we need to make explicit what we count, we need a best practice how to encode that => *need total, but also units* (tokens, lexemes)
- **Consensus**: we count occurrences (tokens / sentences in a corpus)
    - **TODO@CC**: make that clear in the wording

The Ilse of Man is part of the United Kingdom.

- total=length([the, isle, of, man, is, part, of, the, U, K, .])
- total=length([isle of man, united kingdom])

CC: Can we use dc:Dataset?
RS: NIF has a layer for named entities
KG: for Solution 1 each corpus can belong to a dc:Collection?
RS: In this solution a corpus is a layer? Yes, weirdly
FK: reify unitTotal? and parameterise it for kind of unit? CC: it's Solution 3

Totals per units:

- **Solution 1:** One corpus object per possible type of total

  This means that a corpus is a collection of something countable, and the subclass (or the description) says what this countable thing is
- **Solution 2:** specialized properties (to be discussed with Cyprian and Ranka)
    - Last time@Ranka: lemma frequency is at lexical entry, form frequency is at form, doesn't affect frac:total
    - ?Ciprian: what was the open issue?
- **Solution 3**
    - Christian: reified frequency with units could be frac:Frequency, e.g., number of sentences, number of tokens, i.e.
        - Frac:total points to a frac:Frequency object
        - frac:Frequency can contain frac:unit
        - Relative frequencies can be calculated if units match, we can multiple frequency types, e.g., document frequency vs. token frequency
    - Note: counting is a property of the dictionary (corpus query) rather than a property of the corpus, but if corpus not available for querying, needs to be explicit
- **Solution 4**: link with different MetaShare objects (=> Size (of a resource), SizeUnit, property value, unit)

    - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/Size
    - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/SizeUnit
    - frac:Corpus -ms:distribution-> ms:DatasetDistribution
    - ms:DatasetDistribution -ms:size-> ms:Size
    - ms:Size -ms:size-> Literal
    - ms:Size -ms:sizeUnit-> ms:SizeUnit
        - ms:SizeUnit: word, token, entry; **TBC@KG**: add lexeme
    - We could introduce Lexeme as a unit and just use that
- **Solution 5**: DataCube?
    - **TO-BE-DONE**@Gilles: sample

TODO@??? After the next call: model different segmentations (using Solution 3) from

- https://korpling.german.hu-berlin.de/annis3/#_q=dG9rCg&_c=cGNjMg&cl=5&cr=5&s=0&l=10

TODO@RS: MWE example data

- As corpus query

Examples

- Google NGrams
    - N-gram / year / occurrences (token frequency) / document frequency (no totals, we can make up totals)
- **TODO@all: more examples**

Corpus class and property

- https://github.com/ontolex/frequency-attestation-corpus-information/issues/6
- Corpus class
    - something that covers both NLP corpora, other data collections but also invidual texts and things described by bibliographical entries
    - Collection of countable items (=> frac:total)
    - **Tbc**: are we happy with the naming?
- **Tbc**: is it sufficiently clear from the guidelines that this isn't obligatory?
    - Can be used alongside with (or, to some extent, replaced by) frac:locus
    - Frac:locus is the solution for the DH paper
- Possible modelling strategies (**TODO**: vote)
    - Keep frac:Corpus (and use alongside with frac:locus property, which might point to a bibref, for example, or to a specific passage in a book or corpus)
    - Use dct:DataSet instead (can also be a single text, i.e., set with one element)
    - Use (member of) dct:DCMIType?

        - Cf. https://www.dublincore.org/schemas/rdfs/, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
        - Actual object could then be any of Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text, cf. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/dublin_core_type.nt
        - Disadvantage: we cannot put a proper class name into the diagram, but only "member of dct:DCMIType": frac:corpus rdfs:range [http://purl.org/dc/dcam/memberOf dct:DCMIType]
            - In the ontology, this cannot be formalized in RDFS, but requires OWL2

Note on corpus property: when we use dct:DCMIType or dct:DataSet, instead, we might want to reconsider to use dct:source, as these are supposed to work together

Priorities

- Near-complete text (minus queries) until LDK meting
- Prios: corpus > total > queries
- As soon as we know how to handle total (and corpus), assign people for rewriting individual sections

Suggestion for corpus:

- Replace frac:corpus with frac:observedIn
- Abandon frac:Corpus, instead, require that the object of frac:oibservedIn should be assigned a DCMIType, give Dataset, Collection and Text as examples
- Proböem_ frac:total doesn't have a name anymore, w ecould just say that range is any member of DCMIType
- Frac:locus would be unaffected

Suggestuon for total:

- Point to frequency, instead of number
- Create new property unit for frequency objects

TODO@CC: provide a revision of the text with these changes
TODO@FK: coordination call MWE paper
TODO@CC: ask for 2 week extension

AoB

Next call: 25.05

Agenda 2023-05-11

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/ygk-sfmg-azi

Time: 14:00–15:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Aposto

Publications

- Updates on the MWE chapter?
    a. Not yet, **TODO@CC** ask for an update
- Fahad: follow up on TEI+RDFa (w. Gilles) ← no updates yet, **push back to July**
    a. Potential contributors: GS, FK, CC [could integrate older CC+MI work from DH2018]
    b. Potential venues: Journal?
        CC and MI had an invitation to the International Journal of Digital Humanities from a 2019 paper with a related topic (“Linking the TEI: Approaches, Limitations,
        Use Cases”, where we compare an RDFa modelling with its alternatives), we could just get back to them with a proposal that wraps both pieces of work
        International Journal of Digital Humanities, see email quoted 2023-04-13
- Fahad: Arabic corpora as a use case for FrAC+Morph ← status?

    a. Potential contributors: FK, CC, MI
    b. Potential venues: very early stage (LREC-COLING 2024?)
- Middle Persian@Cologne to LREC?
    a. MI: maybe using FrAC (and maybe Morph) for the Middle Persian
    b. Potential contributors: MI, CC, FK + people from MPCD
- CC: decide about a venue for presenting, collect a team of co-authors, and have designated calls ← for which of these? Status?
- **Consensus**: submit to LREC-COLING (as a poster), then (or in parallel) try a journal publication

Consolidation

Any updates on the diagram / model?

Attestations

Issue closed: https://github.com/ontolex/frequency-attestation-corpus-information/issues/15
Any updates on TODOs? Anything to add?
TODO@FK: look into example after
https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#gloss-property

Corpus queries — postponed (CC, RS, MI)

More work and discussions needed before we can talk about this. At the next meeting we will have some ideas/results

frac:total

Open issue: https://github.com/ontolex/frequency-attestation-corpus-information/issues/5
Domain declaration of frac:total has been the motivation to introduce corpus class

- **Consensus**: we need to make explicit what we count, we need a best practice how to encode that => *need total, but also units* (tokens, lexemes)
- **Consensus**: we count occurrences (tokens / sentences in a corpus)
    - **TODO@CC**: make that clear in the wording

The Ilse of Man is part of the United Kingdom.

- total=length([the, isle, of, man, is, part, of, the, U, K, .])
- total=length([isle of man, united kingdom])

CC: Can we use dc:Dataset?
RS: NIF has a layer for named entities
KG: for Solution 1 each corpus can belong to a dc:Collection?

RS: In this solution a corpus is a layer? Yes, weirdly
FK: reify unitTotal? and parameterise it for kind of unit? CC: it's Solution 3

Totals per units:

- **Solution 1:** One corpus object per possible type of total
  This means that a corpus is a collection of something countable, and the subclass (or the description) says what this countable thing is
- **Solution 2:** specialized properties (to be discussed with Cyprian and Ranka)
  - Last time@Ranka: lemma frequency is at lexical entry, form frequency is at form, doesn't affect frac:total
  - ?Ciprian: what was the open issue?
- **Solution 3**
  - Christian: reified frequency with units could be frac:Frequency, e.g., number of sentences, number of tokens, i.e.
    - Frac:total points to a frac:Frequency object
    - frac:Frequency can contain frac:unit
    - Relative frequencies can be calculated if units match, we can multiple frequency types, e.g., document frequency vs. token frequency
  - Note: counting is a property of the dictionary (corpus query) rather than a property of the corpus, but if corpus not available for querying, needs to be explicit
- **Solution 4**: link with different MetaShare objects (=> Size (of a resource), SizeUnit, property value, unit)
  - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/Size
  - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/SizeUnit
  - frac:Corpus -ms:distribution-> ms:DatasetDistribution
  - ms:DatasetDistribution -ms:size-> ms:Size
  - ms:Size -ms:size-> Literal
  - ms:Size -ms:sizeUnit-> ms:SizeUnit
    - ms:SizeUnit: word, token, entry; **TBC@KG**: add lexeme
  - We could introduce Lexeme as a unit and just use that
- **Solution 5**: DataCube?
  - **TO-BE-DONE**@Gilles: sample

TODO@??? After the next call: model different segmentations (using Solution 3) from

- https://korpling.german.hu-berlin.de/annis3/#_q=dG9rCg&_c=cGNjMg&cl=5&cr=5&s=0&l=10

TODO@RS: MWE example data

- As corpus query

Examples

- Google NGrams
    - N-gram / year / occurrences (token frequency) / document frequency (no totals, we can make up totals)
- **TODO@all: more examples**

Corpus class and property

- https://github.com/ontolex/frequency-attestation-corpus-information/issues/6
- Corpus class
    - something that covers both NLP corpora, other data collections but also invidual texts and things described by bibliographical entries
    - Collection of countable items (=> frac:total)
    - **Tbc**: are we happy with the naming?
- **Tbc**: is it sufficiently clear from the guidelines that this isn't obligatory?
    - Can be used alongside with (or, to some extent, replaced by) frac:locus
    - Frac:locus is the solution for the DH paper
- Possible modelling strategies (**TODO**: vote)
    - Keep frac:Corpus (and use alongside with frac:locus property, which might point to a bibref, for example, or to a specific passage in a book or corpus)
    - Use dct:DataSet instead (can also be a single text, i.e., set with one element)
    - Use (member of) dct:DCMIType?
        - Cf. https://www.dublincore.org/schemas/rdfs/, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
        - Actual object could then be any of Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text, cf. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/dublin_core_type.nt
        - Disadvantage: we cannot put a proper class name into the diagram, but only "member of dct:DCMIType": frac:corpus rdfs:range [http://purl.org/dc/dcam/memberOf dct:DCMIType]
            - In the ontology, this cannot be formalized in RDFS, but requires OWL2

Note on corpus property: when we use dct:DCMIType or dct:DataSet, instead, we might want to reconsider to use dct:source, as these are supposed to work together

AoB

Next call: 25.05

Agenda 2023-04-27

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/jez-xvvu-gbe
- One-time for **over-time** discussion (as of 14:45): https://meet.google.com/bbh-pdin-cjy

Time:	14:00–15:00 CET
Draft:	Github
Diagram	https://github.com/ontolex/frequency-attestation-corpus-information note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FK - Fahad Khan (excused)
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică (excused)
KG - Katerina Gkirtzou
EA - Elena Apostol

Publications

- updates?
    - CT: ESWC? (deadline in Dec)
    - MWE? (still waiting)
    - Fahad: follow up on TEI+RDFa (w. Gilles)
        - Potential contributors: GS, FK, CC [could integrate older CC+MI work from DH2018]
        - Potential venues: Journal?
          CC and MI had an invitation to the International Journal of Digital Humanities from a 2019 paper with a related topic ("Linking the TEI: Approaches, Limitations,
          Use Cases", where we compare an RDFa modelling with its alternatives), we could just get back to them with a proposal that wraps both pieces of work

        - *International Journal of Digital Humanities*, see email quoted 2023-04-13
- Fahad: Arabic corpora as a use case for FrAC+Morph
    - Potential contributors: FK, CC, MI
    - Potential venues: very early stage (LREC-COLING 2024?)
- CC: decide about a venue for presenting, collect a team of co-authors, and have designated calls

Consolidation

procedure

- Open issues/addenda:
    - Corpus queries
        - POS-sensitive collocation metrics
    - frac:total (frequencies for lemmas and tokens)
    - Frac:corpus
    - **No more use cases**
- Vis: render https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl with WebVOWL (https://service.tib.eu/webvowl/#iri=https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl)

Looking through the draft of the FrAC guidelines...we should try and clean them up

- Spec and OWL now both generated from/contained in Markdown
- CC: currently, this is done in the ontology, with a special property (vs:term_status) indicating terms that haven't confirmed to be stable
    - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
- **Todo**: check vs:term_status flags in the markdown

Attestations

Open issues

- To be implemented: https://github.com/ontolex/frequency-attestation-corpus-information/issues/15
    - **DONE@CC:** https://github.com/ontolex/frequency-attestation-corpus-information/pull/16
        - MI: Drop syntactic argument
        - MI: slashing frac:gloss? MI: no!
        - merged

    - OE example needs to be compared with the original source from which it quotes
- **TODO**@all: think about new/better/corrected examples
    - RS: Serbian
    - MI: Middle Persian
- **TODO@FK**:
    - Review/edit/annotate attestation section, pull request, if ready, we discuss the merge
- **TODO@MI:** examples for locus and corpus
    - Was TO-BE-DONE@FK
- **Open Question**: GS: gloss as a structured object? cf. gloss in DBnary
    - **TO-BE-DONE**@GS: Example next time [postponed]
- Ranka question for aligned corpus
    - Is it vartrans:Translation + vartrans:source+ vartrans:target appropriate for sentences? I have a TMX corpus and I want to convert to NIF.
    - Attestations for bilingual dictionary, from aligned corpus, can you suggest a respectable example ?
    - Each bilingual entry should have a bilingual attestation
        - Observable: vartrans:Translation (or ontolex:LexicalConcept)
    - :111930-sr-sense-111930-de-sense-trans a vartrans:Translation;
    - vartrans:source :111930-sr-sense;
    - vartrans:target :111930-de-sense.
    - :111930-sr-sense-111930-de-sense-trans a frac:Observable; frac:attestation _:UPCOMING.
    - **TODO@CC**: propose something for _:UPCOMING

Postponed

Frac:total

Domain declaration of frac:total has been the motivation to introduce corpus class

- https://github.com/ontolex/frequency-attestation-corpus-information/issues/5
- **Consensus**: we need to make explicit what we count, even if different counts for different corpus objects, we need a best practice how to encode that => *need total, but also units* (tokens, lexemes)
- **Consensus**: we count occurrences
    - **TO-BE-DONE@CC**: make that clear in the wording

Totals per units:

- **Solution 1:** One corpus object per possible type of total
  This means that a corpus is a collection of something countable, and the subclass (or the description) says what this countable thing is

- **Solution 2:** specialized properties (to be discussed with Cyprian and Ranka)
    - Last time@Ranka: lemma frequency is at lexical entry, form frequency is at form, doesn't affect frac:total
    - ?Ciprian: what was the open issue?
- **Solution 3**
    - Christian: reified frequency with units could be frac:Frequency, e.g., number of sentences, number of tokens, i.e.
        - Frac:total points to a frac:Frequency object
        - frac:Frequency can contain frac:unit
        - Relative frequencies can be calculated if units match, we can multiple frequency types, e.g., document frequency vs. token frequency
    - Note: counting is a property of the dictionary (corpus query) rather than a property of the corpus, but if corpus not available for querying, needs to be explicit
- **Solution 4**: link with different MetaShare objects (=> Size (of a resource), SizeUnit, property value, unit)
    - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/Size
    - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/SizeUnit
    - frac:Corpus -ms:distribution-> ms:DatasetDistribution
    - ms:DatasetDistribution -ms:size-> ms:Size
    - ms:Size -ms:size-> Literal
    - ms:Size -ms:sizeUnit-> ms:SizeUnit
        - ms:SizeUnit: word, token, entry; **TBC@KG**: add lexeme
    - We could introduce Lexeme as a unit and just use that
- **Solution 5**: DataCube?
    - **TO-BE-DONE**@Gilles: sample

Examples

- Google NGrams
    - N-gram / year / occurrences (token frequency) / document frequency (no totals, we can make up totals)
- **TODO@all: more examples**

Corpus class and property [postponed until after clarification of total]

- https://github.com/ontolex/frequency-attestation-corpus-information/issues/6
- Corpus class
    - something that covers both NLP corpora, other data collections but also invidual texts and things described by bibliographical entries

    - Collection of countable items (=> frac:total)
    - **Tbc**: are we happy with the naming?
- **Tbc**: is it sufficiently clear from the guidelines that this isn't obligatory?
    - Can be used alongside with (or, to some extent, replaced by) frac:locus
    - Frac:locus is the solution for the DH paper
- Possible modelling strategies (**TODO**: vote)
    - Keep frac:Corpus (and use alongside with frac:locus property, which might point to a bibref, for example, or to a specific passage in a book or corpus)
    - Use dct:DataSet instead (can also be a single text, i.e., set with one element)
    - Use (member of) dct:DCMIType?
        - Cf. https://www.dublincore.org/schemas/rdfs/, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
        - Actual object could then be any of Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text, cf. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/dublin_core_type.nt
        - Disadvantage: we cannot put a proper class name into the diagram, but only "member of dct:DCMIType": frac:corpus rdfs:range [http://purl.org/dc/dcam/memberOf dct:DCMIType]
            - In the ontology, this cannot be formalized in RDFS, but requires OWL2

Note on corpus property: when we use dct:DCMIType or dct:DataSet, instead, we might want to reconsider to use dct:source, as these are supposed to work together

Corpus queries

Last time: Separate call with Max and Ranka, results to be presented at next call

https://www.korpus.cz/kontext/view?viewmode=kwic&pagesize=40&attrs=word&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.aref&q=~5MqQkiMmMA8E

- **Status**
    - feedback integrated into https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md
        - frac:variables now rdf:Seq (not a json array)
    - **DONE@MI:** example
        - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/samples/queries/query-example.ttl
        - Could be nicely visualised via https://sketch.zazuko.com/
    - **DONE@CC**: modify example according to the proposal

        - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/samples/queries/query-example.2022-10-03.CC.ttl
- **Suggestions**
    - Add queryLink to QueryResult (string literal) [not discussed yet]
        - MI (old): A link to the corpus management interface to access the same results. Should be optional. Not discussed yet
    - don’t formalize query semantics [discussed last time]
    - rename QueryResult to Query, that might be necessary because it seems to have been misunderstood [not discussed]
- **TODO@CC**: update document according to example
- **TODO@CC**: prepare meeting with Ranka and Max, results to be presented mid-May

POS-aware MWE metrics

- Last time: problem confirmed (examples from Ciprian and Ranka)
    - https://personalpages.manchester.ac.uk/staff/sophia.ananiadou/ijodl2000.pdf suggest filters of regular expressions of POSes:
        - 1. Noun+ Noun,
        - 2. (Adj jN oun)+ N oun,
        - 3. ((Adj jN oun)+ j((Adj jN oun) (N ounP r ep)?)(Adj jN oun))N oun
    - RS: SketchEngine: predefined and custom filters
- Possible solution; constraints on collocation object e,.g., a pattern represented like a corpus query)
- postponed after discussing corpus queries
    - we need some mechanism to express patterns/filters

?wrapper around SketchEngine API

Agenda 2023-04-12

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/tma-jead-kpm
- One-time for **over-time** discussion (as of 14:45): https://meet.google.com/hek-xasw-vaw

Time: 14:00–15:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Consolidation [procedure, not discussion]

- In order to wrap up, what are the open issues with the current FrAC model (as in diagram)
- Addenda:
  - Corpus queries
  - frac:total (frequencies for lemmas and tokens)
  - Frac:corpus
  - No more use cases
- Vis: render https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl with WebVOWL

(https://service.tib.eu/webvowl/#iri=https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl)

Looking through the draft of the FrAC guidelines...we should try and clean them up

- Spec and OWL now both generated from/contained in Markdown
- CC: currently, this is done in the ontology, with a special property (vs:term_status) indicating terms that haven't confirmed to be stable
  - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
- **Todo**: check vs:term_status flags in the markdown

Attestations

- To be implemented (not discussed)
  - Open issues with attestation section:
    - new/better/corrected example? (Serbian?)
    - **TODO**@all: think about examples
    - **TO-BE-DONE@FK**:
      - Review/edit/annotate attestation section, pull request, if ready, we discuss the merge
      - examples for locus and corpus
    - **TODO@CC**: mention decomp:Component as possible observable
      - Must be made clear that the list is not exhaustive
      - GS: language phenomena?
      - To be discussed *at Morph*: why is Morph observable but Rule not?
  - Frac:gloss vs. rdf:value
    - consensus: rdf:value is the actual text, frac:gloss is how it is represented in the resource (or derived from that)
    - GS: it was unclear from the description in the guidelines what the rdf:value would be
      - **TODO@CC**: add definition of rdf:value in Attestation class
      - **TO-BE-DONE@FK**: provide example for frac:gloss in guidelines
        - **TO-BE-DONE@CC**: provide example for rdf:value in guidelines
      - **FK**: Upcoming: More examples from another dictionary (Portuguese, not always nicely cited; *whole dict*: here are scans from the first edition https://mordigital.fcsh.unl.pt/dicionario_1789/a/
- **Open Question**: GS: gloss as a structured object? cf. gloss in DBnary
  - **TODO**@GS: Example next time [postponed]

POS-aware MWE metrics?

- Raised as problem during MWE call

- **TO-BE-DONE**: first need to confirm problem!
    - Idea: wrapper around SketchEngine API
        - **TO-BE-DONE@MI:** tackle a first look on skletchengine API
            - focus on modelling, not writing a wrapper!
    - Discussion so far:
        - Same metric with different POSes?
        - Open question: which metrics is this about? **=> Ciprian**
            - **https://personalpages.manchester.ac.uk/staff/sophia.ananiadou/ijodl2000.pdf suggest filters of regular expressions of POSes:**
                - **1. Noun+ Noun,**
                - **2. (Adj jN oun)+ N oun,**
                - **3. ((Adj jN oun)+ j((Adj jN oun) (N ounP r ep)?)(Adj jN oun))N oun**
            - RS: Function NC and CValue are not say, SketchEngine
            - BK: filters are language-specific
            - RS: for terminology extraction, we reimplemented that, this is for terminology extraction; in sketch engine, you can define patterns, but not related with measures
            - RS: many papers with alternative patterns
            - CT: used in terminology extraction
            - CC: join this discussion with the discussion of corpus queries
        - Do we need to capture full word sketches? (maybe not?)
            - RS: SketchEngine: predefined and custom filters
        - **CC:**
            - Possible solution; constraints on collocation object e,.g., a pattern represented like a corpus query
        - **KG:** designated sub-properties for contextual patterns

Corpus queries

https://www.korpus.cz/kontext/view?viewmode=kwic&pagesize=40&attrs=word&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.aref&q=~5MqQkiMmMA8E

- **Status**
    - feedback integrated into https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md
        - frac:variables now rdf:Seq (not a json array)
    - **DONE@MI:** example
        - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/samples/queries/query-example.ttl

        - Could be nicely visualised via https://sketch.zazuko.com/
    - **DONE@CC**: modify example according to the proposal
        - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/samples/queries/query-example.2022-10-03.CC.ttl
- **TBC:** Add queryLink to QueryResult (string literal)
    - MI: A link to the corpus management interface to access the same results. Should be optional. Not discussed yet
- **TBC:** don't formalize query semantics
- **TBC**: rename QueryResult to Query, that might be necessary because it seems to have been misunderstood
- **TODO@CC**: update document according to example
- **TODO@CC**: prepare meeting with Ranka and Max, results to be presented mid-May

Publications

- CT: ESWC? (deadline in Dec)
- MWE? (still waiting)

- Fahad: follow up on TEI+RDFa (w. Gilles)
    - Potential contributors: GS, FK, CC [could integrate older CC+MI work from DH2018]
    - Potential venues: Journal?
      CC and MI had an invitation to the International Journal of Digital Humanities from a 2019 paper with a related topic ("Linking the TEI: Approaches, Limitations, Use Cases", where we compare an RDFa modelling with its alternatives), we could just get back to them with a proposal that wraps both pieces of work

> From: **Thorsten Ries** <T.Ries@sussex.ac.uk>
> Date: Sat, Jul 20, 2019, 13:14
> Subject: Invitation to contribute to International Journal of Digital Humanities
> To: <chiarcos@informatik.uni-frankfurt.de>, <max.ionov@gmail.com>
>
> Dear Christian Chiarcos and Max Ionov,
>
> I trust this email finds you well. I hope you remember we talked briefly after your fascinating paper "Linking the TEI: Approaches, Limitations, Use Cases". As mentioned, I would like to cordially invite you, also on behalf of the editor-in-chief Gábor Palkó, to submit the write-up of your paper to the peer-reviewed International Journal of Digital Humanities. https://link.springer.com/journal/42803.

I am also Chair of the AgE-DH commission of the Arbeitsgemeinschaft germanistische Edition and I totally share your view that XML is losing ground and support, we need to think about different models (RDFa, JSON-LD, etc). When you mentioned that in another presentation you had a
decision node "Do I need TEI" I was delighted:
Thinking about longterm sustainability of scholarly editions, I have been eyeing "minimal computing" editions based on Hugo, Jekyll etc for a while - HTML 5, maybe enriched with some JS, does pretty much everything
that the scholarly editor wants and it is the most sustainable format there is together with PDF/A ... so please leave the "Do I need TEI" node in if you decide to publish with us!

There is currently no deadline set for the next issue of IJDH. If you are willing to submit to the journal, please let us know so we can coordinate an optimal timeline for submission that ensures you have time to do your write-up, get peer-reviewed and see it published quickly afterwards.

The publisher does not charge APC for contributions to International Journal of Digital Humanities. After an embargo period of 12 months, authors may publish preprint versions on their institutional repositories and personal websites.

With best wishes,

Thorsten Ries

--
Thorsten Ries
Arts A A137, University of Sussex, HAHP / SHL
Brighton BN1 9QN
United Kingdom

Marie Sklodowska-Curie Fellow
School of History, Art History and
Philosophy (HAHP),
Sussex Humanities Lab (SHL)

Twitter: @riesthorsten
Phone (campus): +44-1273873188
Mobile: +44-7852433166 (also: Signal messenger)

https://thorsten-ries.online

- Fahad: Arabic corpora as a use case for FrAC+Morph
    - Potential contributors: FK, CC, MI
    - Potential venues: very early stage (LREC-COLING 2024?)
- CC: decide about a venue for presenting, collect a team of co-authors, and have designated calls

POSTPONED

- Frac:total
- Corpus class and property (postponed after clarification of total)

Agenda 2023-03-30

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/imj-kemo-ajy
- One-time for **over-time** discussion (as of 14:45): https://meet.google.com/rjc-qvyv-ozf

Time: 14:00–15:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Publication & Dissemination

MWE chapter

- Draft under: https://www.overleaf.com/8285444258rpfnbwgwbrdp
- Submitted, no feedback yet
    - Substantial last-minute cuts and restructuring, please double-check to make sure you're ok with the revisions, we can restore text for revision/final submission (after acceptance)

UniDive

- Working group on the lexicon-corpus interface,

BPMLOD

https://www.w3.org/community/bpmlod/wiki/Guidelines_and_best_practices_for_LLOD

LDK

- Workshop planning
- TEI+RDFa@LDK ? (from DH)
  - Probably too late
  - GS: to be coordinated with Fahad

Orga

- In order to wrap up, what are the open issues with the current FrAC model (as in diagram)
- Addenda:
  - Corpus queries
  - frac:total (frequencies for lemmas and tokens)
  - frac:corpus
- Cf. consolidation section below

Serbian NIF corpora

- Ranka: Published NIF corpora (Serbian) via Fuseki (1000 novels)
  [On-ELTeC-TEI2NIF-LLOD-report - Google документи]
  Nice illustrator for all FrAC capabilities
  => possible to create a use case using SPARQL LOAD?
- LDK paper (follow-up call with CC today 15:15)
- To be discussed today
  - Token vs. lemma frequency (=> frac:total)
  - Grammatical feature values not in lexinfo (maybe better Serbian MTE ontology => http://nl.ijs.si/ME/owl/)
  - Dictionaries available in relDB
  - NIF data probably compiled in about 3 weeks, then CoNLL transformation
    - Good topic for an LD4LT call in about 4 weeks
    - Also ties with UniDive
  - Corpora in RDF
    - To connect Serbian dictionaries with Serbian NIF corpora
    - Export as RDF
  - Synergies with Katherinas work possible

Consolidation

- Vis: render https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl with WebVOWL (https://service.tib.eu/webvowl/#iri=https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl)

Looking through the draft of the FrAC guidelines...we should try and clean them up at some stage.

- Spec and OWL now both generated from/cntained in Markdown
- CC: currently, this is done in the ontology, with a special property (vs:term_status) indicating terms that haven't confirmed to be stable
    - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
- **Todo**: check vs:term_status flags in the markdown
- **Make this a priority of the next call (start with that topic)**
- No more use cases

Attestations

- To be implemented (not discussed)
    - Open issues with attestation section:
        - new/better/corrected example? (Serbian?)
        - **TODO@**all: think about examples
        - **TO-BE-DONE@FK**:
            - Review/edit/annotate attestation section, pull request, if ready, we discuss the merge
            - examples for locus and corpus
        - **TODO@CC**: mention decomp:Component as possible observable
            - Must be made clear that the list is not exhaustive
            - GS: language phenomena?
            - To be discussed *at Morph*: why is Morph observable but Rule not?
    - Frac:gloss vs. rdf:value
        - consensus: rdf:value is the actual text, frac:gloss is how it is represented in the resource (or derived from that)
        - GS: it was unclear from the description in the guidelines what the rdf:value would be
            - **TODO@CC**: add definition of rdf:value in Attestation class

        - **TO-BE-DONE@FK**: provide example for frac:gloss in guidelines
            - **TO-BE-DONE@CC**: provide example for rdf:value in guidelines
        - **FK**: Upcoming: More examples from another dictionary (Portuguese, not always nicely cited; *whole dict*)
- **Open Question**: GS: gloss as a structured object? cf. gloss in DBnary
    - **TODO**@GS: Example next time

POS-aware MWE metrics? [postponed, last in line]

- Raised as problem during MWE call
- **TO-BE-DONE**: first need to confirm problem!
    - Idea: wrapper around SketchEngine API
        - **TO-BE-DONE@MI:** tackle a first look on skletchengine API
            - focus on modelling, not writing a wrapper!
    - Discussion so far:
        - Same metric with different POSes?
        - Open question: which metrics is this about? **=> Ciprian**
        - Do we need to capture full word sketches? (maybe not?)
        - **CC:**
            - Possible solution; constraints on collocation object e,.g., a patterm represented like a corpus query
        - **KG:** designated sub-properties for contextual patterns

frac:total

- Domain declaration of frac:total has been the motivation to introduce corpus class
- **previously**: Do we want introduce/keep our own property/ies? => Vote => Yes, we need our own property/properties
    - TODO: discuss the naming

- **Note** that we don't say what kind of total that is: tokens, lexemes? **Earlier agreement** : we need to make explicit what we count, even if different counts for different corpus objects, we need a best practice how to encode that => *need total, but <u>also units</u>*
- **Solution 1:** One corpus object per possible type of total

    This means that a corpus is a collection of something countable, and the subclass (or the description) says what this countable thing is
- **Solution 2:** specialized properties (to be discussed with Cyprian and Ranka)
    - Ranka: token freq, lemma freq (number of occurrences in corpus)
        - Cf. German "Wörter" ~ count over a <u>list</u> of forms ~ SPARQL COUNT
        - Ex. "a b c d e f f d e g h i" => 12 tokens
    - Besim/Christian: ?type frequency (number of *different* forms in a corpus)

        - Cf. German: “Worte” ~ count over a set of forms ~ SPARQL COUNT DISTINCT
        - Ex. “a b c d e f f d e g h i” => 9 types
    - MI: Problem: if we want to give sense frequencies, type frequencies are possible only relative to a lexical resource
        - MI: we cannot do type frequency because the data provider operates independently from our resource (i.e., they don’t know our types), so we count occurrences
        - **TODO**@CC: make clear that we count occurrences
    - Ranka

        :SrFudKo_token_freq rdfs:subClassOf frac:Frequency, :SrFudKo, [a owl:Restriction; owl:onProperty dct:description; owl:hasValue "token frequency"].
        :SrFudKo_lemma_freq rdfs:subClassOf frac:Frequency, :SrFudKo, [owl:Restriction; owl:onProperty dct:description; owl:hasValue "lemma frequency"].
        Lemma frequency: total of all inflected forms
        “Mouse mice house”
        freq(mouse)= 1
        lfreq(mouse)=2
        Suggestion:
        lemma freq = frequency attached to lexical entry
        freq = frequency attached to form
        No revision necessary for single word units, but for multi-word expression we might need

- **Solution 2.a**
    - Christian: reified frequency with units could be frac:Frequency, e.g., number of sentences, number of tokens, i.e.
        - Frac:total points to a frac:Frequency object
        - frac:Frequency can contain frac:unit
        - Relative frequencies can be calculated if units match, we can multiple frequency types, e.g., document frequency vs. token frequency
    - Gilles: isn’t counting a property of the dictionary (corpus query) rather than a property of the corpus
        - MI: yes, if corpus not available for querying, needs to be explicit
- **Solution 3**: link with different MetaShare objects (=> Size (of a resource), SizeUnit, property value, unit)
    - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/Size
    - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/SizeUnit

    - frac:Corpus -ms:distribution-> ms:DatasetDistribution
    - ms:DatasetDistribution -ms:size-> ms:Size
    - ms:Size -ms:size-> Literal
    - ms:Size -ms:sizeUnit-> ms:SizeUnit
        - ms:SizeUnit: word, token, entry; **TBC@KG**: add lexeme
    - We could introduce Lexeme as a unit and just use that
- **Solution 4**: DataCube?
    - **TO-BE-DONE for January**@Gilles: sample

TODO@all: example, PLEASE ADD BELOW

Google NGrams:
N-gram / year / occurrences (token frequency) / document frequency
(no totals, we can make up totals)

Postponed

- Corpus class and property [postponed until after clarification of total]
- Corpus queries

Agenda 2023-02-16

Telco link (please check here for updated link):

- **One-time** link:https://meet.google.com/wes-zaqt-ybf
- One-time for **over-time** discussion (as of 14:45): https://meet.google.com/izh-ppah-sbf

Time:	14:00–15:00 CET
Draft:	Github
Diagram	https://github.com/ontolex/frequency-attestation-corpus-information note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković - Belgrade
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică

KG - Katerina Gkirtzou
EA - Elena Apostol

BPMLOD

Neurosymbolic: FrAC?
Andon is the person who propose guidelines on this
To be contacted, feedback at next call

Publication & Dissemination

eLex

- Max in charge of eLex submission
    - On morph+frac
    - Deadline Feb 6th – submitted? link?

UniDive

- Old minutes:
    - Deadline at the day of the last telco
    - CC: trying to make up sth (just general ontolex overview)
        - Co-authors:
            - Fahad
            - Add yourself here Ciprian
- After call
    - CC reworked COLING submission, added COLING co-authors and Fahad
        - Tbc: add Katerina? [at the moment, not possible]
- Notification to come *today*, participation must be confirmed *tomorrow*

MWE chapter

- Draft under: https://www.overleaf.com/8285444258rpfnbwgwbrdp
- Deadline March 1
- CC:
    - Globally restructured: use cases in one section, querying and web services in outlook
    - All text updated, copied and pasted text fully rephrased (except for some globalex quotes … with is is ok because this is a follow-up paper and we can say so)
- On last-time gaps:
    - \subsection{Multi-Word Expressions as Phrases [unassigned]} [Decomp]
        - Renamed to Structure of MWEs

    - Christian+Fahad, done
  - \subsection{The morphological dimension: Compounds as Multi-Word Expressions [unassigned]}
    - Christian+Fahad, sample data from german
  - \section{Use Case: Modelling a Collocation Dictionary (OZDIC)}
    - Last time: SKIP To be written into coherent text [use ODCS, instead]
    - Since then: OZDIC is shorter than ODCS, fully revised by CC, drop ODCS, instead
  - Use Case: N-Grams (Google Books)
    - Old todo: write into coherent text (done)
  - Use Case: Enriching a Collocation Dictionary with Collocation Scores
    - Last time:
      - SKIP? (To be updated)
      - [1 page]: merge / append to ODCS
    - Fully rewritten to match OZDIC
  - Use Case: Designing an API for Collocation Analysis on the Web
    - Last time: Ranka: did scripts, could elaborate that
    - Christian: moved into discussion
  - Redundancy in TEI/LMF discussion (modelling and discussion)
    - **TODO@FK**: cross-reference and check for redundancy
      - Move from discussion here?
      - TODO@Morph call: status
  - Collocations scores
    - **TODO@KG**: (some) cscores into appendix
  - Overall
    - CC+MI: define page lenmgths per sections/(subsectiomns [TODAY]
      - maybe add the page lengths in the overleaf in the title
    - Katerina: balance level of detail of modelling section (\section{The OntoLex Vocabulary})
    - Christian: Cutting (at the very end)
  - Content more or less ok, all coherent, but WAY too long
  - What shall we do about web services?
    - skip
- Confirm status:
  - Who knows what to do, status
    - all : proof-read, correct; if cutting: make a note with a suggestion (unless it'S YOUR SECTION)
    - All: cut (at your sections)
    - FK: merging TEI/LMF sections, check OZDIC section
    - BK: MWE typology, rephrase literal quotations
    - MI page lengths, read/comment/improve overall, incl. Cutting *suggestions*

        - CC: page lengths, maybe proof-reading
            - TODO: We need to mention Lexicog in the beginning, because we use it for some modeling. TODO: add this to the structure
        - KG: more cscores into appendix (possibly: cuts in that section), balancing/cutting in modelling section
- Coordination
    - CC: limited capacity for coordination. Could do a tour-de-force writing/cutting, only
- Page limit? (15-25 pages max excl. refs)
    - 43 pages now
    - Must be <=25 excl. refs

Postponed

Serbian NIF corpora

- Ranka:
    - Published NIF corpora (Serbian) via Fuseki (1000 novels)
        - Question: CLARIN hosrting for 9 languages confirmed?
        - Last time Suggestion: publish uncompressed (rather than zip)
            - => possible to create a use case using SPARQL LOAD
            - Next time: (some) data => Then: try it out ;)
    - Next: FrAC
        - To connect Serbian dictionaries with Servbian NIF corpora
        - Dictionaries available in relDB
        - Export as RDF
        - No questions yet
        - Ranka showed original data
            - Incl. corpus linking (NoSketchEngine), dynamic frequency
        - Synergies with Katherinas work possible
        - Discuss as top priority (after publications) next time
            - Nice illustrator for all FrAC capabilities

Other open issues

- Attestations
- frac:total
- Query extension

AoB

Agenda 2023-02-02

Telco link (please check here for updated link):

- **One-time** link: https://meet.google.com/znj-hpiz-pwx
- One-time for **over-time** discussion (as of 14:45): https://meet.google.com/ete-incw-xww

Time:	14:00–15:00 CET
Draft:	Github
Diagram	https://github.com/ontolex/frequency-attestation-corpus-information note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos,U Augsburg
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică (excused)
KG - Katerina Gkirtzou (excused)
EA - Elena Apostol

Publication & Dissemination

- eLex
    - Old minutes:
        - Anything?
        - Separate from Morph?
        - No separate abstract, no capacity. Maybe mention FrAC in the morph talk
    - Max in charge of eLex submission
        - On morph+frac
        - Deadline Feb 6th
        - Max working on it

=> TODO@CC: ask Max for overleaf link etc.

UniDive

- Deadline *today*
    - CC: trying to make up sth (just general ontolex overview)
        - Co-authors:
            - Fahad
            - Add yourself here

MWE chapter

- Draft under: https://www.overleaf.com/8285444258rpfnbwgwbrdp
- Who have already contributed, status
- Who knows what to do, status
- Who doesn't know what to do
- Prognosis
- We need to mention Lexicog in the beginning, because we use it for some modeling. TODO: add this to the structure
- After call comments from CC
    - Will ask for an extension
    - No capacity for coordination. MI? Depends on the extension
    - Parts of the model where there is no final consensus should go as is in the latest discussions, can be updated in further versions of the paper before final publication
- Page limit? (15-25 pages max excl. refs)
    - 49 pages now
    - Tbc.
- Needs much more coordination,cross-dependencies
    - Needs more balance
- Gaps (assume we get a month)
    - \subsection{Multi-Word Expressions as Phrases [unassigned]} [Decomp]
        - Ranka? Simile detection crosslingual, some may be mwes
    - \subsection{The morphological dimension: Compounds as Multi-Word Expressions [unassigned]}
        - Plan to re-use Latin failed
        - Ranka: sample data from Serbian (not writing)
    - \section{Use Case: Modelling a Collocation Dictionary (OZDIC)}
        - SKIP To be written into coherent text [use ODCS, instead]
    - Use Case: N-Grams (Google Books)
        - Katherina: write into coherent text
    - Use Case: Enriching a Collocation Dictionary with Collocation Scores

  - SKIP? (To be updated)
  - [1 page]: merge / append to ODCS
- Use Case: Designing an API for Collocation Analysis on the Web
  - Ranka: did scripts, could elaborate that
- Gaps in discussion
  - Christian
- Overall
  - Katherina: balance level of detail of modelling section (\section{The OntoLex Vocabulary})
  - Christian: Cutting (at the very end)
  - Balancing page lengths => CC+MI?

Serbian

- Ranka:
  - Published NIF corpora (Serbian) via Fuseki (1000 novels)
    - Question: who could host 9 languages?
      - CLARIN?
      - Francesca Frontini: no end point, yet, but working on it
      - Fahad: they should be able to host, setting up data center, but might take some time
    - Published as zip files
    - Suggestion: uncompressed
      - => possible to create a use case using SPARQL LOAD
      - Next time: (some) data => Then: try it out ;)
  - Next: FrAC
    - To connect dictionaries with corpora
    - Dictionaries available in relDB
    - Export as RDF
    - No questions yet
    - Ranka showed original data
      - Incl. corpus linking (NoSketchEngine), dynamic frequency
    - Synergies with Katherinas work possible
    - Discuss as top priority (after publications) next time
      - Nice illustrator for all FrAC capabilities

postponed

- Attestations

- frac:total
- Queries

AoB

- Time slot fixed?
- Next call: 16.02.23, 2pm CET

Agenda 2023-01-19

Telco link (please check here for updated link): https://meet.google.com/wgw-otnz-woy

Time: 14:00–15:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, U Cologne (excused)
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică (excused)
KG - Katerina Gkirtzou (excused)
EA - Elena Apostol

Publication & Dissemination

- eLex
  - Anything?
  - Separate from Morph?
  - No separate abstract, no capacity. Maybe mention FrAC in the morph talk
- MWE chapter: deadline extension request? A separate call seems to be necessary
  - MI: Asked CC if he could do the coordination soon

MWE chapter

- Draft under: https://www.overleaf.com/8285444258rpfnbwgwbrdp
- Who have already contributed, status
- Who knows what to do, status
- Who doesn’t know what to do
- Prognosis
- We need to mention Lexicog in the beginning, because we use it for some modeling. TODO: add this to the structure
- After call comments from CC

    - Will ask for an extension
    - No capacity for coordination. MI? Depends on the extension
    - Parts of the model where there is no final consensus should go as is in the latest discussions, can be updated in further versions of the paper before final publication

Attestations

-

frac:total

-

AoB

- Time slot fixed?
- Next call: 02.02.23, 2pm, right?

Agenda 2022-12-08

Telco link (please check here for updated link): https://meet.google.com/wgw-otnz-woy

Time: 14:00–15:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică (excused)
KG - Katerina Gkirtzou
EA - Elena Apostol

Publication & Dissemination

- eLex: https://elex.link/elex2023/call-for-papers/
  - on TEI+RDFa?
    - Partial derivation via XSLT/GRDDL?
    - Maybe just provide an XSLT script, unless there is any additional information in the RDF representation
  - Jan 31st: abstract (500 words)
  - Full paper: approx mid 2023, cf. 2019: abstract was due in February, final paper deadline: June
- LDK workshops:
  - To be submitted by Dec 19
  - Two meetings for a W3C community groups for language technology day, feat the 4th OntoLex face-to-face meeting
    - https://docs.google.com/document/d/1OYupkE_vnZisQCVarYoqfIgnLQTD727kez7IRNAbNhE/edit?usp=sharing
  - Models should be ready by then (Sep 2022)
- MWE chapter
  - No calls since, status tbc.

    - Slack channel: #frac-mwe-chapter
    - Status: discussion to be continued via Slack (exclusively)

Temp: Doodle-Poll

- Majority vote
- Ciprian cannot make it to this slot until end of semester
- **TODO:** Yet another Doodle

Attestations

- TEI+RDFa@Gilles: Updates?
    - Last time: Fahad submitted corrected version to Github, but not reviewed/reannotated yet for RDFa attributes
    - GS forked and added RDFa to the TEI data (older version, still to be adapted to new TEI version): https://github.com/serasset/attestationExample/tree/rdfa
- Open issues with attestation section:
    - new/better/corrected example? (Serbian?)
        - Last time: Ranka: First prep NIF corpora, data has been updated
            - Waiting for some input from Max, Christian (and all ;)
            - Corpus with 9 languages, published as documents, only, right now
            - On-ELTeC-TEI2NIF-LLOD-report - Google документи : description and links to data
            - **TO-BE-DONE@all**: peek into the data to spot improvements ;)
                - Max: roughly ok
            - Chat:
                - nif:EntityOccurrence doesn't seem to be in NIF 2.0: https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html#classes
                - CC, FK: metadata looks ok
            - Follow-up comments (offline)
                - CC: you might save a triple per word by dropping nif:OffsetBasedString. All nif:RFC5147Strings are (I found that not all EntityOccurrences in http://llod.jerteh.rs/ELTEC/srp/NIF/SRP18780_BrankoM_PredZoru-L2.ttl have that)
                    - RS: tnx, yes, I found mistake in my code.
                - CC: nif:posTag is ok as is, but according to the NIF 2.0 doc, its usage is discouraged [you don't need to fix, but can be future imnprovement]
                - CC: I suggest you create additional links between entity occurrences and words. Right now, they don't seem to be connected, this could be either
                    - <http://llod.jerteh.rs/ELTEC/srp/NIF/SRP18780_BrankoM_PredZoru-L2.txt#char=13948,13954> nif:superString

          <http://llod.jerteh.rs/ELTEC/srp/NIF/SRP18780_BrankoM_PredZoru-L2.txt#char=13948,13962>

          or

          - <http://llod.jerteh.rs/ELTEC/srp/NIF/SRP18780_BrankoM_PredZoru-L2.txt#char=13948,13962> nif:subString <http://llod.jerteh.rs/ELTEC/srp/NIF/SRP18780_BrankoM_PredZoru-L2.txt#char=13948,13954>
        - CC: other than that, it looks ok
        - CC: minor suggestion: it depends a bit on your use case, but NIF has a lot of inverse properties. To save space, you could document which direction you used and provide only these. That’s what I did for CoNLL-RDF, for example. Just make sure your users know that they have to look for nif:sentence instead of nif:word (or the like)
  - **TO-BE-DONE**: Check readability and consistency
    - Edit and pull requests: MI?, CC?, GS?
      - FK: some time early Dec?
      - **TODO@CC**: remove outdated diagrams
  - **TODO@FK**: after this meeting
    - Review/edit/annotate attestation section, pull request, if ready, we discuss the merge
    - examples for locus and corpus
    - example from LiLa dicts?
      - **TODO@FK: talk to LiLa** (Lewis/Short)
    - Too late for DH reviews => lower priority
- Frac:gloss vs. rdf:value
  - consensus: rdf:value is the actual text, frac:gloss is how it is represented in the resource (or derived from that)
  - GS: it was unclear from the description in the guidelines what the rdf:value would be
    - **TO-BE-DONE@CC**: add definition of rdf:value in Attestation class
    - **TO-BE-DONE@FK**: provide example for frac:gloss in guidelines
      - **TO-BE-DONE@CC**: provide example for rdf:value in guidelines
    - **FK**: Upcoming: More examples from another dictionary (Portuguese, not always nicely cited; *whole dict*)

Frac:total

- Domain declaration of frac:total has been the motivation to introduce corpus class
- **Last time**: Do we want introduce/keep our own property/ies?
  - Last call: **Vote** confirmed that we need our own property/properties
  - TODO: discuss the naming

- Note that we don’t say what kind of total that is: tokens, lexemes?

    => One corpus object per type of total possible
    This means that a corpus is a collection of something countable, and the subclass (or the description) says what this countable thing is

    - **TBC**: are we ok with that?
    - Ranka (last time): Lexicographers need total of lexemes
    - Katerina: we need to make explicit what we count, even if different counts for different corpus objects, we need a best practice how to encode that
        - Metashare Size (of a resource), SizeUnit, property value, unit
            - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/Size
            - http://metashare.ilsp.gr/ontologies/meta-share/meta-share-ontology.owl.v2.0.0.prelease-beta/documentation/index-en.html#/SizeUnit
            - frac:Corpus -ms:distribution-> ms:DatasetDistribution
            - ms:DatasetDistribution -ms:size-> ms:Size
            - ms:Size -ms:size-> Literal
            - ms:Size -ms:sizeUnit-> ms:SizeUnit
                - ms:SizeUnit: word, token, entry; **TBC@KG**: add lexeme
            - We could introduce Lexeme as a unit and just use that
        - DataCube?
            - **TODO for January**@Gilles: sample
            - KG: need total, but also units
        - Specialized properties
            - Tokens, types (ask Cyprian?)
            - **TODO**@discuss next time

Corpus class and property [postpone that after clarification of total]

- https://github.com/ontolex/frequency-attestation-corpus-information/issues/6
- Corpus class
    - something that covers both NLP corpora, other data collections but also invidual texts and things described by bibliographical entries
    - Collection of countable items (=> frac:total)
    - **Tbc**: are we happy with the naming?
- **Tbc**: is it sufficiently clear from the guidelines that this isn’t obligatory?
    - Can be used alongside with (or, to some extent, replaced by) frac:locus
    - Frac:locus is the solution for the DH paper
- Possible modelling strategies (**TODO**: vote)
    - Keep frac:Corpus (and use alongside with frac:locus property, which might point to a bibref, for example, or to a specific passage in a book or corpus)
    - Use dct:DataSet instead (can also be a single text, i.e., set with one element)

    - Use (member of) dct:DCMIType?
        - Cf. https://www.dublincore.org/schemas/rdfs/, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
        - Actual object could then be any of Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text, cf. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/dublin_core_type.nt
        - Disadvantage: we cannot put a proper class name into the diagram, but only "member of dct:DCMIType": frac:corpus rdfs:range [http://purl.org/dc/dcam/memberOf dct:DCMIType]
            - In the ontology, this cannot be formalized in RDFS, but requires OWL2
    - Note on corpus property: when we use dct:DCMIType or dct:DataSet, instead, we might want to reconsider to use dct:source, as these are supposed to work together

POS-aware MWE metrics? [postponed]

- Raised as problem during MWE call
- Last telco: first need to confirm problem!
- (probably) no time to discuss in depth. Any volunteer to present a problem description at next call?
    - Idea: wrapper around SketchEngine API
        - **TODO@MI:** tackle a first look on skletchengine API
            - focus on modelling, not writing a wrapper!
    - Discussion so far:
        - Same metric with different POSes?
        - Open question: which metrics is this about? **=> Ciprian**
        - Do we need to capture full word sketches? (maybe not?)
        - **CC:**
            - Possible solution; constraints on collocation object e,.g., a patterm represented like a corpus query
        - **KG:** designated sub-properties for contextual patterns

Consolidation

- Vis: render https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl with WebVOWL (https://service.tib.eu/webvowl/#iri=https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl)

Looking through the draft of the FrAC guidelines...we should try and clean them up at some stage.

- Spec and OWL now both generated from/cntained in Markdown
- CC: currently, this is done in the ontology, with a special property (vs:term_status) indicating terms that haven't confirmed to be stable
    - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
- **Todo**: check vs:term_status flags in the markdown
- **Make this a priority of the next call (start with that topic)**
- No more use cases

Corpus queries [postponed]

https://www.korpus.cz/kontext/view?viewmode=kwic&pagesize=40&attrs=word&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.aref&q=~5MqQkiMmMA8E

- **Status**
    - feedback integrated into https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md
        - frac:variables now rdf:Seq (not a json array)
    - **DONE@CC**: modify an example according to the proposal
        - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/samples/queries/query-example.2022-10-03.CC.ttl
- **TBC:** Add queryLink to QueryResult (string literal)
    - MI: A link to the corpus management interface to access the same results. Should be optional. Not discussed yet
- **TBC:** don't formalize query semantics
- **TBC**: rename QueryResult to Query, that might be necessary because it seems to have been misunderstood

AOB

- Next call
- **TO-BE-DONE@MI**: archive old channels, create a new channel

Minutes 2022-11-24

Telco link (please check here for updated link): https://meet.google.com/wgw-otnz-woy

Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg - excused
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Max Ionov, U Cologne
TD - Thierry Declerck (excused)
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Publication & Dissemination

- Any specific feedback from COLING?
  - Besim: two interested colleagues from Korea, but no direct follow-up
- MWE chapter
  - No calls since last frac telco, status tbc.
  - Slack channel: #frac-mwe-chapter
  - Status: looking for a possible meeting slot to coordinate writing
    - https://doodle.com/meeting/participate/id/dNLy4opa
- eLex: https://elex.link/elex2023/call-for-papers/
  - on TEI+RDFa?
    - Partial derivation via XSLT/GRDDL?
  - Jan 31st: abstract (500 words)
  - Full paper: approx mid 2023, cf. 2019: abstract was due in February, final paper deadline: June
- LDK:
  - Fahad is program chair, shouldn't submit (?)
- LDK workshops:

    - To be submitted by Dec 19
    - **DONE@CC**: msg to Ontolex mailing list about a workshop: https://doodle.com/meeting/participate/id/er8nAW2b
    - Models should be ready by then (Sep 2022)

Temp: Doodle-Poll

Please set your prefs
https://doodle.com/meeting/participate/id/ep8I0Epe

Attestations

- TEI+RDFa@Gilles: Updates?
    - Fahad submitted corrected version to Github, but not reviewed/reannotated yet for RDFa attributes
- Open issues with attestation section:
    - new/better/corrected example? (Serbian?)
        - Ranka: First prep NIF corpora, data has been updated
            - Waiting for some input from Max, Christian (and all ;)
            - Corpus with 9 languages, published as documents, only, right now
            - On-ELTeC-TEI2NIF-LLOD-report - Google документи : description and links to data
            - **TODO@all**: peek into the data to spot improvements ;)
            - Chat:
                - nif:EntityOccurrence doesn't seem to be in NIF 2.0: https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html#classes
                - CC, FK: metadata looks ok
    - **TO-BE-DONE**: Check readability and consistency
        - Edit and pull requests: MI?, CC?, GS?
            - FK: some time early Dec?
            - **TODO@CC**: remove outdated diagrams
    - **TODO@FK**: *after* this meeting
        - Review/edit/annotate attestation section, pull request, if ready, we discuss the merge
        - examples for locus and corpus
        - example from LiLa dicts?
            - **TODO@FK: talk to LiLa** (Lewis/Short)
- Frac:gloss vs. rdf:value
    - consensus: rdf:value is the actual text, frac:gloss is how it is represented in the resource (or derived from that)
        - GS: it was unclear from the description in the guidelines what the rdf:value would be

        - **TO-BE-DONE@CC**: add definition of rdf:value in Attestation class
        - **TO-BE-DONE@FK**: provide example for frac:gloss in guidelines
            - **TO-BE-DONE@CC**: provide example for rdf:value in guidelines
        - **FK**: Upcoming: More examples from another dictionary (Portuguese, not always nicely cited; *whole dict*)

Frac:total

- Domain declaration of frac:total has been the motivation to introduce corpus class
- **Tbc**: Do we want introduce/keep our own property/ies?
    - FK: recommend to dct:extent for file size
    - FK: re-use sth from Metashare?
        - GS: not immediately found
    - GS: frequency must be related to something. Unit of counting can be difficult, e.g., token count for MWEs. different tokenizers can give different numbers. Could be solved by defining different corpus elements, even if over the same data.
    - CC: suggestion: extend guidelines with “A frac:Corpus should provide a rdfs:description to specify what unit is meant by frac:total, e.g., what tokenizer is being used.”
    - MI: then, the definition should state very clearly, what we mean by frac:Corpus — it’s not so much a corpus, but a corpus + its configuration
        - CC,GS: +2
    - CC: shall we rename corpus to corpus configuration? Issue of naming corpus is next on agenda
    - GS: relative frequency is a minimal use, so we need an explicit property
    - CC/GS: total must not be obligatory (other use cases, e.g. attestation), and incomplete data
    - **Vote:** confirm that we need our own property/properties:
        - 5 yes
        - 0 no

- Note that we don’t say what kind of total that is: tokens, lexemes?

    => One corpus object per type of total possible

    This means that a corpus is a collection of something countable, and the subclass (or the description) says what this countable thing is
    - **TBC**: are we ok with that?
    - Ranka: Lexicographers need total of lexemes
    - To be discussed next call => Doodle for time slot

POSTPONED

- Corpus class and property
- POS-aware MWE metrics?
- Consolidation
- corpus queries

Minutes 2022-11-10

Telco link (please check here for updated link): https://meet.google.com/wgw-otnz-woy

Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg - excused
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Max Ionov, U Cologne
TD - Thierry Declerck (excused)
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Publication & Dissemination

- DH submission TEI-OntoLex DH submission
    - Submitted, notification March (!)
    - Includes links to github, might affect judgment
    - Any specific feedback from COLING? [whenever Besim is back]
- MWE chapter
    - No calls since last frac telco, status tbc.
    - Slack channel: #frac-mwe-chapter
    - Status: to be pushed now (was postponed until DH is submitted)

Temp: Doodle-Pool

https://doodle.com/meeting/participate/id/ep8l0Epe

BK:

Mo: 9-11
Th: 14-16
Fr: 10-12
Fr 14-16
Fr 16-18

Tentative preference for Th 14-15

CC: not before 10 or after 16

TODO@CC: doodle later today, all of Besims options minus those in conflict with mine

Attestations

- Gilles: looking in to producing RDFa version of the TEI => Attestation
  - Two files, TEI and RDF
  - Gilles working on merging both
    - Much redundancy
    - RDFa needed mostly for providing a normalized interpretation of properties and values (URIs instead of values)
    - Is there information that is not representable in TEI?
      - Parts are derivable via XSLT (just explicitation)
    - Different mode4lling strategies possible, nice topic for a paper
    - eLex: https://elex.link/elex2023/call-for-papers/
      - Jan 31st: abstract (500 words)
      - Full paper: approx mid 2023
      - https://elex.link/elex2023/call-for-papers/
      - Cf. 2019: abstract was due in February, final paper deadline: June
    - LDK:
      - Fahad is program chair, shouldn't submit (?)
- LDK workshops:
  - To be submitted by Dec 19
  - **TODO@CC**: msg to Ontolex mailing list about a workshop
  - Models should be ready by then (Sep 2022)
- Discussed in the context of DH submission
- Frac:locus
  - **Confirmed**: We had an agreement to keep both frac:locus and frac:corpus

- **DONE@CC**: add frac:locus to TTL file

- Frac:gloss vs. rdf:value
    - Suggestion: rdf:value is the actual text, frac:gloss is how it is represented in the resource (or derived from that)
        - GS: it was unclear from the description in the guidelines what the rdf:value would be
            - **TODO@CC**: add definition of rdf:value in Attestation class
            - **TODO@FK**: provide example for frac:gloss in guidelines
                - **TODO@CC**: provide example for rdf:value in guidelines
            - **TODO@FK**: update DH example (not next two weeks)
        - approved
- Open issues with attestation section:
    - new/better/corrected example
        - Possibly Serbian example from Ranka
    - Check readability and consistency
        - Edit and pull requests: MI?, CC?, GS?
    - **TODO@FK**: *after* next meeting
        - Review/edit/annotate attestation section, pull request, today, we discuss the merge
        - examples for locus and corpus
        - example from LiLa dicts?
            - **TODO@FK: talk to LiLa** (Lewis/Short)

POSTPONED

- Frac:total
- Frac:corpus class and property
- POW-aware MWE metrics?
- Consolidation
- Corpus queries

AOB

- Next call
- **TO-BE-DONE@MI**: archive old channels, create a new channel

Minutes 2022-10-27

Telco link (please use this one): https://meet.google.com/wgw-otnz-woy

Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg - excused
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Publication & Dissemination

DH Submission

Status:

- Submission status? (@FK)
- Deadline 4.11.
- TEI-OntoLex DH submission
- Text: Almost done, feedback needed, too long
- Priority: clean up guidelines (second week of November)
    - **TODO@all!**
- Issues with modelling?

COLING

- Successfully presented by Besim (no details known yet)

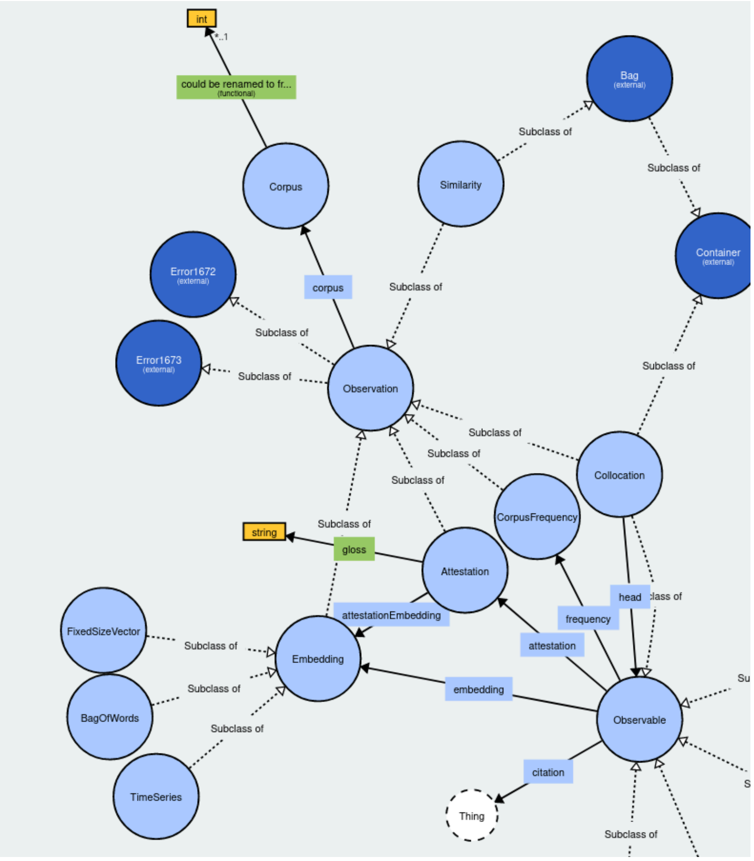
Embeddings on SW mailing list

- https://lists.w3.org/Archives/Public/semantic-web/2022Oct/0059.html
- Potential application in ONNX ?

Chapter proposal: MWE in lexical resources (with Morph)

- Slack channel: #frac-mwe-chapter
- Designated discussion last Monday (see minutes in this document, below)
- MI => sketchenine
    - Idea: wrapper around SketchEngine API
    - Are POS-aware MWE metrics a problem?
        - Same metric with different POSes?
        - We might need to track the extra information about metric
        - Open question: which metrics is this about? **=> Ciprian**
        - Do we need to capture full word sketches? (maybe not?)
        - **CC:**
            - Possible solution; constraints on collocation object e,.g., a üpatterm represented like a corpus query
        - **KG:** designated sub-properties for contextual patterns
        - **MI**: first need to confirm problem!
            - CC: +1 , KG +1
        - **TODO@MI**: tackle a first look on skletchengine API
            - NB: focus on modelling, not writing a wrapper

Consolidation



(vis of https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl with WebVOWL (https://service.tib.eu/webvowl/#iri=https://github.com/ontolex/frequency-attestation-corpus-information/raw/master/owl/frac.ttl)

Looking through the draft of the FrAC guidelines...we should try and clean them up at some stage.

- CC: currently, this is done in the ontology, with a special property (vs:term_status) indicating terms that haven't confirmed to be stable
    - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
- Is there a good technology to develop spec and OWL in a single document?
    - **DONE@CC**: now, Turtle is directly integrated into Markdown (everything enclosed by ` ``` ` is extracted into Turtle file). Has been confirmed to match the original turtle file (excerpt for its omissions) and that it is valid and loadable OWL2/DL. Some restructuring of guidelines in the process.

- Open issues:
    - Corpus class: naming is poor (MI: or maybe not, just shouldn't be obligatory)
        - [it isn't obligatory, but recommended "SHOULD", and frac:locus can be used in addition or as replacement, if we don't want to point to a corpus]
        - We need (and introduced) the class for frac:total
        - Issue not relevant for DH paper (we use frac:locus there)
        - Possible modelling strategies:
            - Keep frac:Corpus (and use alongside with frac:locus property, which might point to a bibref, for example, or to a specific passage in a book or corpus)
            - Use dct:DataSet instead (can also be a single text, i.e., set with one element)
            - Use (member of) dct:DCMIType?
                - Cf. https://www.dublincore.org/schemas/rdfs/, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
                - Actual object could then be any of Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text
                - Disadvantage: we cannot put a proper class name into the diagram, but only "member of dct:DCMIType"
    - Corpus property
        - Naming: same issue
        - Obligatoriness: isn't
        - Former issue (merge with frac:locus) resolved: we re-introduced frac:locus
        - Any issues with that?
        - [CC, post-call: when we use dct:DCMIType or dct:DataSet, instead, we might want to reconsider to use dct:source, as these are supposed to work together]
    - Total property
        - FK: don't be too generic, not dct:extent
        - Possible issue: we don't say what kind of total that is: tokens, lexemes?
            - KG: One corpus object per type of total possible
    - **Open issues with attestation section:**
        - new/better/corrected example
        - Check readability and consistency
    - **TODO@FK**:
        - Review/edit/annotate attestation section, pull request, next week, we discuss the merge

Postponed

- Corpus queries
- Attestations (except for discussion in paper)

Corpus queries [postponed]

https://www.korpus.cz/kontext/view?viewmode=kwic&pagesize=40&attrs=word&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.aref&q=~5MqQkiMmMA8E

- **Status**
  - feedback integrated into https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md
    - frac:variables now rdf:Seq (not a json array)
  - **TO-BE-DONE@CC**: modify an example according to the proposal
    - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/samples/queries/query-example.2022-10-03.CC.ttl
- **TBC:** Add queryLink to QueryResult (string literal)
  - MI: A link to the corpus management interface to access the same results. Should be optional. Not discussed yet
- **TBC:** don’t formalize query semantics
  - FK: it’s better to have representation for query semantics, not just KWIC representation
  - Hard to achieve / beyond scope?
- **TBC**: rename QueryResult to Query, that might be necessary because it seems to have been misunderstood

Attestations [postponed]

- Frac:gloss vs. rdf:value
  - Suggestion: rdf:value is the actual text, frac:gloss is how it is represented in the resource (or derived from that)
    - **TODO@today**: confirm
- Frac:locus
  - Confirmed: We had an agreement to keep both frac:locus and frac:corpus
    - **DONE@CC**: add frac:locus to TTL file
- “Corpus” as something that covers both NLP corpora, other data collections but also invidual texts and things described by bibliographical entries
  - Agreement that this is a problematic term, issue https://github.com/ontolex/frequency-attestation-corpus-information/issues/6

- CC: instead of frac:Corpus, use (member of) dct:DCMIType?
    - Cf. https://www.dublincore.org/schemas/rdfs/, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
    - Actual object could then be any of Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text
        - Cf. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/dublin_core_type.nt
    - After last call: frac:corpus rdfs:range [http://purl.org/dc/dcam/memberOf dct:DCMIType]
        - In the ontology, this cannot be formalized in RDFS, but requires OWL2
- To be discussed: what could be a better name for frac:corpus?
    - CC, After last call: using dct:DCMIType strongly encourages reconsidering dc:source. The type of resource can (and should) be made explicit by its DCMIType.
- **TO-BE-DONE@FK:** example for attestation
    - examples for locus and corpus
    - example from LiLa dicts?
        - **TODO@FK: talk to LiLa**
    - Cf. Perseus+Lewis-Short
- **TO-BE-DONE@FK:** Cf. TEI, chap. 12: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html (cited quotation), chap. 9 (attestation, in critical apparatus only)

AOB

- Next call
- **TO-BE-DONE@MI**: archive old channels, create a new channel

Minutes 2022-10-24: MWE Chapter

Telco link (please use this one):
Time:
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, U Cologne

FK - Fahad Khan
MI - Max Ionov, U Cologne
JM - John McCrae
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol
EB - Elena B.

Chapter proposal: MWE in lexical resources (with Morph)

- NEW Draft for chapter: https://www.overleaf.com/3446689781vctdrjcdwspj (chapters/01.tex)
  - Original draft: https://www.overleaf.com/8285444258rpfnbwgwbrdp *(draw text from there, but not structure)*
- Initial abstract: https://docs.google.com/document/d/1HLTsYScMiZE4Nb6b7mozN-SoKAIWtisdiKLa0OoXPuU/edit?usp=sharing
  - Ciprian: Extraction of domain-specific MWEs, special scores (cvalue)
    - Falls in (d), but additional POS information to extract them
      - Maybe that doesn't fit FrAC
      - https://link.springer.com/article/10.1007/s007999900023
      - Also cf. SketchEngine
    - Other metrics
    - Suggestion: as problem or solution into paper, but discussion in FrAC telcos, not here
- Slack channel: #frac-mwe-chapter

1. Introduction and Background
   - Types and relevance of multi-word expressions, specific challenges **[to be assigned; BK]**
   - Conventional (pre-RDF) modelling strategies for MWEe and machine-readable dictionaries [~ related research] **[FK; BK]**
   - Motivating Linguistic Linked Open Data for lexical resources **[text reuse]**
   - Brief history of OntoLex **[text reuse]**
   - High-level overview over OntoLex (2016) modules, OntoLex-lexicog (2019), OntoLex-Morph (under development) and OntoLex-FrAC (under development) **[CC; …]**
2. Modelling Choices
   - We describe different possibilities for modelling
     (a) lexicalized multi-word expressions (=> OntoLex-Lemon, i.e., *ontolex:MultiWordExpression*), **[CC, FK…]**
     (b) phrasal/free multi-word expressions (=> OntoLex-decomp), **[CC, EA, BK, …]**

(c) morphological compounds (=> OntoLex-Morph) and **[CC; BK, MP …]**
(d) collocations and collocation scores (=> OntoLex-FrAC). **[CC, CT, EA, FK, KG, BK, …]**

- This section introduces the vocabularies, their usage and combination is addressed in the following use case sections. This section could be considered a literature survey.

3. Use Case: Modelling MWEs in dictionaries
   - Introducing OZDIC, a collocation dictionary designed as a learning tool for IELTS, TOEFL and PTE writing tests. Modelling collocations with OntoLex-Lemon (ontolex:MultiWordExpression) **[CC, FK, BK, …]**
4. Use Case: Modelling MWEs in lexical databases **[..., , … CC]**
   - Resource to be confirmed **[EB?; MP?]**
     - Soomething on compounding in Romance languages?
     - *Discarded idea:* A database of morphological compounds: Discussing the relation between morphological compounding and multi-word expressions, examples for compounding from German (GermaNet) and Latin (LiLa). Modelling with OntoLex-Morph and/or OntoLex-Decomp.
       - Matteo: maybe LiLa doesn't fit super well, few spaces in Latin compounds (CC: same for German)
     - *Not yet discarded idea:* Databases of automatically extracted MWE candidates: Google Book n-grams and Leipzig Wortschatz Portal. Modelling with OntoLex-FrAC, comparison with OntoLex-Decomp. [maybe not, but experiment exists]
5. Use Case: Enriching dictionaries with collocation scores **[EA; CT, KG]**
   - Using the Oxford Collocations Dictionary for Students of English (OCDS). Modelling with OntoLex-Lemon, OntoLex-Lexicog, optionally OntoLex-Decomp. Enrichments with OntoLex-FrAC.
   - Discussion on metrics? [if too long => discussion section]
6. Use Case: Querying **[CC; FK?, BK?, …]**
   - Using OCDS, OZDIC and GermaNet compounds. Illustrate that LLOD/OntoLex technology allows to seamlessly integrate information from different sources (here: OntoLex-Lemon, OntoLex-Morph, OntoLex-Decomp + OntoLex-FrAC).
     - We have that for morph and frac example, not the others
7. UseCase: Designing an API for collocation analysis **[MI; BK, …]**
   - Sketching an OntoLex wrapper around SketchEngine (Kilgariff et al. 2014; mostly using OntoLex-FrAC)
     - POS-sensitive scores? **[+ CT]**
     - TF/IDF **[EA]**
8. … (more use cases needed?; **we have been cut off here**)
9. Insights and Applications **[to be discussed]**
   - Recommend modelling choices (based on the use cases)
   - Describing down-stream applications
10. Discussion and Outlook **[to be discussed]**
    - Describe achievements: Easy information integration, linking and reusability (comparison with non-RDF technologies)

- Describe limitations: We only cover the *structural* of multi-word expressions and technical aspects of *collocation analysis*, but not the semantic side (e.g., semantic constraints on the combinations of words).
- Outlook: We now can integrate all kinds of MWEs and information about MWEs from lexical resources, but there are virtually no tools that *natively consume* such data. Current developments in two directions: (a) Recent research on integrating LLOD technology with workflows based on legacy formats (Fäth et al. 2020), so that OntoLex data can be transformed to the required input. (b) Tools with native OntoLex support are emerging, as well, e.g., in VocBench (Fiorelli et al., 2020) and LexO (Bellandi 2021) – but none of these are specifically designed for MWEs or collocations.

Nexus chat:

- Next call in two weeks
  - Then confirm overall structure and update responsibilities
    - Also check match and assignments from old overleaf
  - **TODO@CC**: doodle poll
- Until then, use the comment function in this document for discussion
  - Where assignments are clear, we can also already start working ;)

Minutes 2022-10-13

Telco link (please use this one): https://meet.google.com/wgw-otnz-woy

Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg - excused
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Publication & Dissemination

DH Submission [main topic]

Status:

- Coord by FK, primary topic today
    - https://docs.google.com/document/d/1VmXDrR0LwSHaxbla6L2gLqwD4-O2ut5DswyPjIG_yQk/edit
    - Details skipped
        - Abstract only
        - Focus on modelling, mention (but don't elaborate RDFa)
    - https://github.com/anasfkhan81/attestationExample/tree/main/TEIAtt
        - Romary-approved TEI modelling
        - Minor comments on OntoLex modelling (see chat)
            - Chat:
                - CC: frac:Attestation URIs should include the quoted term
                - CC: could we have multiple loci per attestation? [current modelling is possible anyway]
                - CC: better avoid blank nodes for loci:
                  frac:locus [dc:isPartOf :Codex_Exoniensis] =>
                  frac:locus :Codex_Exoniensis_16_11.
                  :Codex_Exoniensis_16_11 dc:partOf
                  :Codex_Exoniensis: [suggestion]
            - Definitions of frac:gloss and frac:Attestation/rdf:value are too close
    - authors: FK, CC, MI
        - **TODO@FK**: ask Gilles to join (for the RDFa aspect)
- Deadline 25.10

Attestation

- Frac:gloss vs. rdf:value
    - Suggestion: rdf:value is the actual text, frac:gloss is how it is represented in the resource (or derived from that)
        - **TODO**: confirm at next call
    - Keep modelling with rdf:value for DH
- Frac:locus
    - Confirmed: We had an agreement to keep both frac:locus and frac:corpus
        - **TODO@CC**: add frac:locus to TTL file [2022-10-17: done]
    - For text: avoid blank nodes for frac:locus

- "Corpus" as something that covers both NLP corpora, other data collections but also invidual texts and things described by bibliographical entries
    - Agreement that this is a problematic term
    - CC: instead of frac:Corpus, use (member of) dct:DCMIType?
        - Cf. https://www.dublincore.org/schemas/rdfs/, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
        - Actual object could then be any of Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text
            - Cf. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/dublin_core_type.nt
        - After call: frac:corpus rdfs:range [http://purl.org/dc/dcam/memberOf dct:DCMIType] .
            - In the ontology, this cannot be formalized in RDFS, but requires OWL2
    - To be discussed: what could be a better name for frac:corpus?
        - CC, After call: using dct:DCMIType strongly encourages reconsidering dc:source. The type of resource can (and should) be made explicit by its DCMIType.

Postponed

- Other publications
    - COLING: Presented yesterday by Besim
    - Embeddings@SW mailing list: https://lists.w3.org/Archives/Public/semantic-web/2022Oct/0059.html
    - MWE chapter
        - **TODO@MI,CC**: Doodle poll on meeting times
        - **TODO@WHOM**: involve OntoLex-Morph, esp. Matteo Pellegrini
- Model consolidation
- Corpus queries
- Attestations

Minutes 2022-09-29

Telco link (please use this one): https://meet.google.com/wgw-otnz-woy

2nd telco link (if needed): https://meet.google.com/trc-hgja-gzg

Time: 12:00–13:00 CET

Draft: Github

Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Max Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Publication & Dissemination

Chapter proposal: MWE in lexical resources (with Morph)

Status: CC — prefererably start in 2nd half of October
OntoLex-Morph: also mentioned there, interest by Matteo Pellegrini to join

Draft for chapter: https://www.overleaf.com/8285444258rpfnbwgwbrdp
Initial abstract: https://docs.google.com/document/d/1HLTsYScMiZE4Nb6b7mozN-SoKAIWtisdiKLa0OoXPuU/edit?usp=sharing
Slack channel: #frac-mwe-chapter

- Shall we have a separate call? Or have a focus on this next call, 13.10? **TODO@MI,CC**: Doodle poll for the call (start from 18.10)
    - FK: We need to have CC in the call to plan that. If he's available next time, focus on the
    - FK: Separate call for this, 13.10: DH submission discussion (25.10)
- What is missing?
    - MI: **TODO@all**: To look over the Overleaf draft and check if something you want to be there is not there
- Who is the responsible person?
    - CC (after Oct 15)
- Is everyone still on board? CC KG MI BK FK CT EA
- Are there any more authors? CC KG MI BK FK CT EA

DH Submission

- To be discussed next time
- Deadline: 25.10
- Topic: attestations (see minutes from 2022-09-15)
- Responsible person: FK
- TODO@MI: archive old channels, create a new channel
- FK: Will try to prepare an outline. Have an example, it might be a bit complicated (from the last time). TBD in the channel

Attestations (postponed)

Corpus queries

https://www.korpus.cz/kontext/view?viewmode=kwic&pagesize=40&attrs=word&attr_vmode=visible-kwic&base_viewattr=word&refs=%3Ddoc.aref&q=~5MqQkiMmMA8E

- Add queryLink to QueryResult (string literal)
- FK: it’s better to have representation for query semantics, not just KWIC representation
  - This imay be hard to achieve
- MI: against using a json array as a string literal in RDF
- CC: we can have a list of variables
  - CC (after call): this feedback integrated into https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md (not in the diagram yet: rename QueryResult to Query, that might be necessary because it seems to have been misunderstood)
- **TODO@CC**: modify an example according to the proposal

Minutes 2022-09-15: Attestations

Telco link (please use this one): https://meet.google.com/wgw-otnz-woy
2nd telco link (after 1 hour): https://meet.google.com/trc-hgja-gzg
Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

1. Publication & Dissemination

SIGLOD@DH2022

- LOD Special Interest Group of ADHO meeting: July 26
  - https://docs.google.com/document/d/1c8FUObyDKbZ5SMuxecmhzkQs8Ow9VBm_lYJR0sd8xj0/edit

Any outcomes?

FK: Nothing special came out of this, focus was more on DH. Not much feedback, but overall people were happy with what was presented

FK: Maybe we need to promote OntoLex more within the DH community

MI: Tooling?

FK: Maybe SPARQL endpoint would be beneficial already for some projects, some resources (see below, Portuguese dictionary in TEI+RDF)

GS: This is something to raise in Vilnius (beyong FrAC, LLOD in general). We need to promote using LLOD, maybe it’s a good moment for tooling → next project?

BK: Making a bridge between DH and FrAC is very important, this should be on our agenda (with integration to other resources)

FK: There was a proposal to have an online 24h/48h conference/workshop with demos and so on, LOD applied to DH

CC: We could keep this in mind

Chapter proposal: MWE in lexical resources

Status: CC — prefererably start in 2nd half of October
OntoLex-Morph: also mentioned there, interest by Matteo Pellegrini to join

Draft for chapter: https://www.overleaf.com/8285444258rpfnbwgwbrdp

Initial abstract:
https://docs.google.com/document/d/1HLTsYScMiZE4Nb6b7mozN-SoKAIWtisdiKLa0OoXPuU/edit?usp=sharing

Slack channel: #frac-mwe-chapter

FrAC@COLING

BK: Some formatting issues, and we need to remove the last part (addressed to reviewers). BK is going to upload the camera-ready version

2. Attestations

- Current model
    - frac:quotation => rdf:value
    - frac:attestationGloss: keep it (not gloss)
    - Keep frac:locus (=> anyURI) along with frac:corpus
    - **TODO@CC:** update diagram
    - **TODO@CC:** reformulate locus definition, degree of specificity
    - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/img#suggested-simplification
    - OWL https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
- Attestation (previous telcos):
    - Two main perspectives
        - FK: coming from perspective of retro digitized dictionaries (TEI)
            - Started working on a simple TEI based example which we can also try and encode in OntoLex to compare the two models and

            their functionality. You can find the example here: https://github.com/anasfkhan81/attestationExample/tree/main/TEIAtt
            It is taken from the Bosworth Toller Old English dictionary

        - CC: coming from corpus-based dictionaries (ePSD)

          [**TODO@CC**: make the difference more explicit in the text]

- attestation
    - locus with anyURI flexible degree of specificity [NEED EXAMPLES]
    - Corpus property with corpus object, to be used if the locus can be considered a corpus [NEED EXAMPLES]
    - Attestation can have either or both locus and corpus
- **TODO@FK:** example for attestation for Sep 15, 2022
    - Question: which attestation property to be reduced to rdf:value?
    - examples for locus and corpus
    - example from LiLa dicts?
        - **TODO@FK: talk to LiLa**
    - Cf. Perseus+Lewis-Short
- **TODO@FK:** Cf. TEI, chap. 12: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html (cited quotation), chap. 9 (attestation, in critical apparatus only)

CC: The problem is that DH is oriented at XML representation, would RDF be an acceptable format at all for them

GS: We can embed RDF inside TEI (RDFa), maybe this is the way to go, this is added value, not a replacement

CC: Would very much support this

MI: +1

FK: Proposal: overview of the two models, showing that there are no limiting differences on the level of semantics. Then RDFa and showing how to extend TEI with RDF. Then showing the functionality: querying, linking with different resources

FK: It would be a good DH submission (26.10)

GS: It’s also important to point out that RDF is general enough, there is no longer strict connection to a document, everything is combinable — e.g. author information from ontologies; Not only why going to FrAC from TEI but more generally, why going to RDF at all

Responsible person: FK

3. Corpus Queries

Summarize updates from last call

- https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md

4. Consolidation

- sieve through ontology
    - OWL model of FrAC: https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
- establish/track vs:term_status
- frac:corpus
    - Does it have to be obligatory for an observable?
    - Extend the definition so that we can also point into dictionaries that provide collocation dictionaries?
    - This was the original intention of using dc:source, instead.
- **TODO@all**: add/check issues in https://github.com/ontolex/frequency-attestation-corpus-information/issues
    - For fixes, create a pull request

5. Aob

Next call: 2022-09-29

Minutes 2022-09-01: Corpus Queries

Telco link: https://meet.google.com/wgw-otnz-woy **(IMPORTANT: this is a different link)**
Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, U Cologne (excused)
FK - Fahad Khan (excused)
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

1. Publications

Any action points?

- Chapter proposal: MWE in lexical resources
- FrAC@COLING
    - Paper accepted as a poster - Congratulations to all authors
    - Camera-ready version can have up to 9 pages excluding references and annexes
    - Deadline for submitting camera-ready version 15th September
    - CT: added Authors affiliation - needs confirmation **TODO@authors**
    - CT: added Aknowledgment section
- Anything else?
- CT: submitted version can be camera-ready version, not much to change based on the reviews
- BK: Still need to go through the paper and proof-read

6. Model consolidation

Today: no discussion, just looking onto current version

Changes:

- frac:quotation => rdf:value
- frac:attestationGloss: keep it (not gloss)
- Keep frac:locus along with frac:corpus

TODO@CC: update diagram

- https://github.com/ontolex/frequency-attestation-corpus-information

2. Corpus queries [main topic]

- Initial presentation/discussion
- https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md
- Come up with an example modeling for SkEngine wrapper

3. Aob

Next call:

- Sep 15: attestations
- DH2023, deadline mid-October. Idea: conversion between TEI and OntoLex regarding attestations. Maybe also querying parts; just a module presentation (it should be finished by then)

Minutes 2022-07-21

Telco link: https://meet.google.com/azi-ycvk-zkx (check here for new link if this one isn't working)
Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică (excused)
KG - Katerina Gkirtzou
EA - Elena Apostol

7. Publications

Chapter proposal: MWE in lexical resources

Submitted, no feedback yet
OntoLex-Morph: also mentioned there, interest by Matteo Pellegrini to join

Draft for chapter: https://www.overleaf.com/8285444258rpfnbwgwbrdp

Initial abstract:
https://docs.google.com/document/d/1HLTsYScMiZE4Nb6b7mozN-SoKAIWtisdiKLa0OoXPuU/edit?usp=sharing

Slack channel: #frac-mwe-chapter

FrAC@COLING: see model consolidation

8. Model consolidation

- CC: I would like to simplify the diagram

    - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/img#suggested-simplification
        - Replace ContextualRelation with frac:Observation, superclass also for Attestation, Frequency, Embedding and Similarity
        - Merge frac:locus and frac:corpus, define it as property of frac:Observation
        - Rename either frac:quotation or frac:attestationGloss to rdf:value (=> can be inherited from frac:Observation)
    - Christian: integration is the novel contribution of the COLING paper
        - Reviews shotr, but relatively positive
        - Currently rebuttal
    - OWL model of FrAC: https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl

2.1 Discussion

- frac:Corpus
    - Max: METASHARE Corpus class?
        - Katerina: too much overhead? Specifically designed for metadata. Frequency (total) would be buried?
        - Gilles: need only to be able to attach a total number of tokens for relative freq.
        - Tentative consensus against it
- Attestation (previous telcos):
    - Two main perspectives
        - FK: coming from perspective of retrdigitized dictionaries (TEI)
        - CC: coming from corpus-based dictinaries (ePSD)

          [maybe make the difference more explicit in the text]

- frac:corpus at attestation
    - Old question: Rename frac:corpus property?
        - Tentative consensus: No, but resurrect locus along with it
    - CC: corpus along with web annotation?
        - FK: maybe nt tie to closely to web annotation? Mre like a recmmendatin?
        - CC: just use additinal locus property for attestatins, locus: point into a work/corpus, corpus: point to cmplete (cllection of) wrks
        - FK: locus may have different levels of specificity, ranging from full wrk to passage, substring, chapter, scene, act, actual address. But sometimes, they don’t have exact pointers, but “Plato”.
    - CC: suggestion:
        - locus with anyURI flexible degree of specificity [NEED EXAMPLES]
        - Corpus property with corpus object, to be used if the locus can be considered a corpus [NEED EXAMPLES]
        - Attestation can have either or both locus and corpus
    - KG: need sample data
- **For the moment**, we have a tentative consensus for
    - frac:quotation => rdf:value
    - frac:attestationGloss: keep it (not gloss)

    - Keep frac:locus along with frac:corpus

2.2 TODOs

- **TODO@CC:** reformulate locus definition, degree of specificity
- **TODO@CC:** update diagram
- **TODO@FK:** example for attestation for Sep 15, 2022
    - Question: which attestation property to be reduced to rdf:value?
    - examples for locus and corpus
    - example from LiLa dicts?
        - **TODO@FK: talk to LiLa**
    - Cf. Perseus+Lewis-Short
- **TODO@FK:** Cf. TEI, chap. 12: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html (cited quotation), chap. 9 (attestation, in critical apparatus only)

9. Aob

- LOD Special Interest Group of ADHO meeting:

    2 meetings at/during DH2022, 2nd call open to anyone, 2nd call on 26th July

    Agenda document: https://docs.google.com/document/d/1c8FUObyDKbZ5SMuxecmhzkQs8Ow9VBm_IYJR0sd8xj0/edit

    5 min lightning talks on projects

    Discussed at Nexus Slack [several people attending 2nd call n Jule 26th]

    - florentina representing sth on Nexus DH use case
    - CC presenting sumerian [if participating]
    - **TODO:** MI,FK: FrAC [**TODO**: list at website]

Next calls:

- Sep 1: corpus queries, initial presentation
- Sep 15: attestations

Minutes 2022-07-07

Telco link: https://meet.google.com/azi-ycvk-zkx (check here for new link if this one isn't working)
Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică (excused)
KG - Katerina Gkirtzou
EA - Elena Apostol

Publications

Chapter proposal: MWE in lexical resources

Submitted, no feedback yet

Draft for chapter: https://www.overleaf.com/8285444258rpfnbwgwbrdp
Initial abstract:
https://docs.google.com/document/d/1HLTsYScMiZE4Nb6b7mozN-SoKAIWtisdiKLa0OoXPuU/edit?usp=sharing
Slack channel: #frac-mwe-chapter

Original idea:

- Intro/background (CC, CT)
- Overview of the model (FK, EA, KG)
  - How MWEs are represented in Ontolex (FK)
  - How frac:Collocation works (EA)
  - Frac:scores (CT, KG)
- Use-cases (where to use what)
  - … (BK?)
  - … (LiLa?)
    - Todo: to be discussed with them after acceptance
- Discussion

Model consolidation

- CC: I would like to simplify the diagram
    - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/img#suggested-simplification
        - Replace ContextualRelation with frac:Observation, superclass also for Attestation, Frequency, Embedding and Similarity
        - Merge frac:locus and frac:corpus, define it as property of frac:Observation
        - Rename either frac:quotation or frac:attestationGloss to rdf:value (=> can be inherited from frac:Observation)
- Discussion (oreviously):
    - Max (previously): looks appealing
    - Fahad (previously): looks appealing
    - We need to check all previous modelling (CC: changes only concern attestation)
    - Fahad: “observation” may be a misnomer, embeddings may be an aggregation over observations rather than a single observation
    - Several questions on whether embeddings are observations
    - Christian: likewise, “corpus” (for locus of attestations) may be a misnomer. It is, however, defined as “any body of primary data”
    - Christian: could be the novel contribution of the COLING paper (if we go for it)
    - Consensus: use that as selling point for COLING paper
- **TODO@FK:** example for attestation
    - Question: which attestation property to be reduced to rdf:value?
    - No before end of August
- Updates to attestation
    - Cf. TEI, chap. 12: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html (cited quotation), chap. 9 (attestation, in critical apparatus only)
    - Rename frac:corpus property?
    - FK: coming from perspective of retrdigitized dictionaries (TEI)
        - **TODO@FK: talk to LiLa**
            - example from LiLa dicts?
            - Cf. Perseus+Lewis-Short
        - **For the moment** go with
            - Quotation => rdf:value
            - attestationGloss: keep it (not gloss)
            - Frac:locus => frac:corpus
                - Not really happy about the ambiguity in the scope/semantics of the object
    - KG: maybe resolve ambiguity: distinguish whether pointing to corpus or part of it, a hierarchy?
        - FK: anyrthing from metashare?
        - KG: ms:mediaPart , but more on metadata level, between corpus and a part of it
    - Web annotation for specific reference, corpus for corpus or corpus metadata ?
        - https://www.w3.org/TR/annotation-model/
        - **TODO@CC**: find an example (ePSD)

        - CC: coming from perspective of corpus-based dictionaries (http://psd.museum.upenn.edu/nepsd-frame.html)
    - OWL model of FrAC
        - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
    - **POSTPONED**:
        - sieve through ontology
        - establish/track vs:term_status
        - frac:corpus
            - Does it have to be obligatory for an observable?
            - Extend the definition so that we can also point into dictionaries that provide collocation dictionaries?
            - This was the original intention of using dc:source, instead.
        - **TODO@all**: add/check issues in https://github.com/ontolex/frequency-attestation-corpus-information/issues
            - For fixes, create a pull request

Corpus queries [postponed]

https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md

Aob

Someone should join ADHO SIGLOD meeting:

> **(TODO@CC ?)**
>
> I'm writing to let you know that the LOD Special Interest Group of ADHO will be meeting virtually on July 26 8:00-10:00 (JST)/July 25 23:00-July 26 1:00 (UTC). After two years without a meeting, we plan to use this time together to revitalize this group and organize ourselves to have a web presence and central space for sharing ideas around Linked Open Data and Cultural Heritage. We'll also be looking for people to help run events and take a more active role in the SIG.
>
> If you are attending DH2022 and would like to join us, please join us! When you register you should receive a link to the workshop Zoom rooms. If you have yet to register, you can do so here: https://dh2022.adho.org/registration
>
> Warmly,
>
> Kim Martin and Susan Brown
>
> LOD-SIG Leaders, ADHO

Jul 21, 12:00-13:00 CEST

Minutes 2022-06-09

Telco link: https://meet.google.com/azi-ycvk-zkx
Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, U Cologne
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, U Cologne
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Publications

GlobaLex paper

https://overleaf.com/5768717378vzdnrzyxvfxn

Accepted

- Deadline: 27th May
- Submitted
- Discuss presentation
    - How long? 30 min: 20+10 min
    - Attending: CC, FK [not], BK, MI, KG(?)
        - Presenter: MI(+CC)
            - Todo@MI: also ask BK
    - 10:05-10:35
- Further coordination via Slack
- Nexus slide template?
    - TODO@FK

COLING paper

- Submitted, review postponed
- Think of an alternative venue
    - LLODream? (cf. Morph)
        - http://llodapproaches2022.mruni.eu/
        - https://nexuslinguarum.eu/activities/vilnius-conference-llodream2022/
        - 15.6., 350 w + postproc
        - Template dead => TODO@CC: ask Giedre
        - Google doc? TODO@CC: create one, add link here + note in slack in coling channel

Chapter proposal: MWE in lexical resources

Msg. by Verginica Barbu Mititelu

> We are contacting you to draw your attention to the call we have released for chapter outlines on "Multiword expressions in lexical resources". We are kindly inviting you to contribute the volume with a chapter that would offer state-of-the-art information on the representation of multiword expressions in Linked Data lexical resources, as well as any demonstration of the advantages offered by this technology.

Idea:

- Compare ontolex modelling of MWEs with ontolex:MWE and frac:Collocation and frac:BagOfWords
    - Describe use cases / data sets
    - MI: difference in use cases
    - CT: nothing new needed, no new ex

> We find there is a gap with respect to this aspect in the existing literature and it would be great if any of you, all of you, or any of your colleagues find(s) the time to fill it. Of course, any other idea you consider relevant and timely in connection to the call is welcome.
> From the call: “The outline should clearly express the topic and the main contribution(s) of the envisaged chapter, as well as provide an overview of the chapter's structure. The

outline should be 1-2 pages long (A4, Times New Roman, 12 points, single spacing) (references excluded)."

Contributors:

- FK, CT, KG, EA, CC

DONE@CC: create an Overleaf, outline

Skeleton + template text from Globalex:
https://www.overleaf.com/8285444258rpfnbwgwbrdp

Preliminary structure:

- Intro/background (CC, CT)
- Overview of the model (FK, EA, KG)
  - How MWEs are represented in Ontolex (FK)
  - How frac:Collocation works (EA)
  - Frac:scores (CT, KG)
- Use-cases (where to use what)
  - … (BK?)
- Discussion

Set the general idea.

Slack channel: #frac-mwe-chapter

All information about the call follows here.

%%%%%%%%%%%%

The accepted contributions will be published in an edited volume as follows:

- title: "Multiword expressions in lexical resources. Linguistic, Lexicographic and Computational perspectives"
- publication venue: series "Phraseology and Multiword Expressions" at Language Science Press
- editors: Voula Giouli, Verginica Barbu Mititelu

The proposed book is aimed at giving an account of the state-of-the-art regarding multiword expressions (MWE) representation in lexical resources in view of their robust identification and computational processing. We target both large size general lexical resources and smaller MWE-centred ones, with special focus on the representation decisions / mechanisms that facilitate their usage in NLP tasks. Possible topics to be covered include, though they are not limited to:

- MWEs in linguistic research and lexicons development; the interface between grammar and lexicon;
- MWE representation in computational lexicons (MWEs in wordnets, FrameNet, other computational lexicons, representation via word embeddings, etc.);
- MWE bi-lingual and multilingual lexical resources
- MWE lexicons for under-resourced languages and language varieties;
- Standards and formalisms for the representation of MWEs: available possibilities of representing MWE with Linked Data vocabularies; limitations;
- Acquisition of MWE lexicons and MWE lexical resources - also for less-resourced languages or non-standard language (tweets, forums, user-generated content);
- Synergies between Lexicography and NLP in view of MWE identification and discovery;
- MWE representation in corpora - linking with lexical resources /aligning resources containing MWEs - monolingual vs. cross-lingual perspective;
- MWE lexicons in NLP tasks (identification, parsing, machine translation, etc.) - limitations and possibilities;
- Automatic discovery of newly coined MWEs or new variants of existing MWEs.

Submission of outlines:

The outline should clearly express the topic and the main contribution(s) of the envisaged chapter, as well as provide an overview of the chapter's structure. The outline should be 1-2 pages long (A4, Times New Roman, 12 points, single spacing) (references excluded). The outline will be sent by email to the volume editors at: voula@athenarc.gr and vergi@racai.ro.

Tentative timeline:

- submission of the outline of the contribution: 31 May 2022
- ####Extended deadline: 15 June
- notification of acceptance: 25 July 2022
- submission of full contributions: 15 January 2023
- notification of acceptance: 31 March 2023
- submission of final versions: 30 April 2023

%%%%%%%%%%%%%

We hope you will find the call timely for publishing a new insightful contribution to the field!
Please feel free to forward this to any potentially interested party!

Best regards,
Voula Giouli
Verginica Barbu Mititelu

Verginica Barbu Mititelu
PhD, Senior researcher II
NLP Group
Research Institute for Artificial Intelligence
Romanian Academy

POSTPONED

Model consolidation

- CC: I would like to simplify the diagram
    - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/img#suggested-simplification
        - Replace ContextualRelation with frac:Observation, superclass also for Attestation, Frequency, Embedding and Similarity
        - Merge frac:locus and frac:corpus, define it as property of frac:Observation
        - Rename either frac:quotation or frac:attestationGloss to rdf:value (=> can be inherited from frac:Observation)
- Discussion (briefly):
    - Max (last time): looks appealing
    - Fahad (this time): looks appealing
    - We need to check all previous modelling (CC: changes only concern attestation)
    - Fahad: “observation” may be a misnomer, embeddings may be an aggregation over observations rather than a single observation
    - Several questions on whether embeddings are observations
    - Christian: likewise, “corpus” (for locus of attestations) may be a misnomer. It is, however, defined as “any body of primary data”
    - Christian: could be the novel contribution of the COLING paper (if we go for it)
    - Consensus: use that as selling point for COLING paper
- **DONE@CC**: provide OWL model of FrAC

        - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
- **POSTPONED**:
    - sieve through ontology
    - establish/track vs:term_status
    - Frac:corpus
        - Does it have to be obligatory for an observable?
        - Extend the definition so that we can also point into dictionaries that provide collocation dictionaries?
        - This was the original intention of using dc:source, instead.
    - **TODO@all**: add/check issues in https://github.com/ontolex/frequency-attestation-corpus-information/issues
        - For fixes, create a pull request

Corpus queries

https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md

No call 2022-05-26

Cancelled for datathon preparations, etc.

Minutes 2022-05-12

Telco link: https://meet.google.com/azi-ycvk-zkx
Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, GU Frankfurt
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, GUF (excused)
TD - Thierry Declerck
MP - Matteo Pellegrini

CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

Publications

GlobaLex paper

https://overleaf.com/5768717378vzdnrzyxvfxn

Accepted
Deadline: 27th May

- merge into single main.tex
    todo@christian: DONE by 2022-05-13 10:30 CET

- style: too technical, too difficult to read, and contains a number of typos
    As a minor note the authors are very inconsistent about the use of fixed-width fonts for RDF entity names. I would also expect names to be capitalized in the citations section (OntoLex-Lemon, Sketch Engine), the last citation is also missing the authors (Colin McIntosh, Ben Francis & Richard Poole).
    - **todo@besim**: pre-final read
    - **todo@fahad**: final read

- examples on how corpora will be accessed for describing collocations how the resulting information will be encoded in FrAC
    Cf. attestation paper:
        NIF (return a context + offsets)
        Web Annotation (standoff annotation over documents which are online)
        Attestation discussion: citation
    Modelling of corpus queries:
        CC: not yet
    todo@christian

- know more about the potential applications which could benefit from the RDF representation of collocations
    - information integration
    - **todo@ciprian**: recommendation systems

- higher level description of the model and the justification for its design
    todo@Besim + Katerina

Moreover, there is a real lack of comparison with other models and I would have expected the authors to at least make some comparison with competing models such as TEI, which has a collocation tag ().
compare other models, at least TEI

todo@fahad [first draft] & christian [second, revision]

The authors also justify the modelling with respect to the Oxford Collocation Dictionary, which is good, but a second example would show that the model is more flexible to a wider range of use cases and would improve the evaluation of the model.

TODO: second example
We have that already => task: select and elaborate use cases
TODO@christian: restore original use cases
DONE 2022-05-13 10:30 CET
[Elenas use case: later, problematic to add as co-author now]

- restore metrics

todo@christian: restore in document (DONE 2022-05-13 10:30 CET)
todo@ciprian+katerina
todo@besim: motivate selection (omission) of metrics
Reformulation; probability, not frequency

Suggestion:

- Maxi version by next call
- Cut it within two days
- If all ready, we can start cutting before, coordinate via slack
- Nothing gets lost, but content not going directly into the publication must reside in a separate folder in overleaf

COLING paper?

17.5.!

- Anonymous!
    - Private github?
- Overview over the entire model
    - Short paper? Small focused contrib
    - Owl ontology (no defs yet)
    - Describe all parts of the module
        - **TODO@Christian**: create the initial skeleton with re-used text
            - https://www.overleaf.com/8837131433wpdbfddjmntz
            - DONE 2022-03-13 01:00: skeleton
            - TODO: re-used text
        - Background:

      - moviating the use of ontolex / RDF
      - Motivsating the general approach [Observable/Observation]
        - After call: separate section ;)
      - Historical things
      - [some of that taken form existing text, needs paraphrasing] **[Fahad?]**
    - Attestation
      - Paraphrasing **[Fahad]**
    - Frequency
      - Paper excerpt [**Elena**, tbc. Via Slack]
      - **DONE@Christian**: sent paper
    - Embeddings **[Christian(+Max?)]**
    - Collocations **[Ciprian, Besim]**
    - Similarity
      - Paper excerpt [**Elena**, tbc. Via Slack]
      - **DONE@Christian**: sent link to section in GitHub draft (correction: not in any paper so far)
    - Linking corpora **[Christian]**
  - **No coding**
    - Single example?
    - Brown corpus
    - American Heritage Dictionary
      - Screenshot or anything => attestation example?
    - Frequency
      - Brown corpus can be queried via SketchEngine => example?
    - Embeddings
      - SemCor corpus (=> Brown corpus), WordNet annotation + Autoextend embeddings
    - Collocations
      - See slack conversation
    - Similarity
      - Sketch Engine? [Christian+Elena]
    - **TODO@all**: confirm via Slack until midnight tomorrow whether these examples work
      - For questions: also ask christian directly (via slack)
- OLD NOTES
  - ? use cases that involve multiple aspects of the current modelling
    - Attestation
    - Frequency
    - Embeddings
    - Collocations
    - ?Links with other ontolex vocabularies
    - Think until May 2nd, evening

Use case:

  - data from some shared task?
  - Two data sets from similar shared tasks?
  - Max: Coref OntoNotes?
    - Attestations (PTB)
    - Lexical data (sense groupings, propbank)
    - Embeddings: coref features [=> not a task, but a system]
      - AutoExtend
    - Similarity clusters: brown clusters?
    - Referent detection (collocations?)
    - Benefit: improved re-usability
  - Resource integration:
    - AutoExtend: WordNet+Embeddings
    - OntoNotes: sense grouping + attestation
    - Wortschatz collocation data over PTB? collocation, frequency, attestation
    - Brown clusters over PTB
    - Problem: cannot distribute PTB
    - Work on a single PTB file?
    - Evaluation: we can query across all those resources
=> work on conversion
- ? overall consolidation [restructure of diagram to make it more readable etc.]
  - More like a motivation/position paper
    - “Linking language resources with OntoLex-FrAC”
      - https://overleaf.com/8837131433wpdbfddjmntz
      - Planned Coordination call: Wed May 4, 12:00 CEST (Berlin)
        - **CANCELLED**
    - Cf. resource integration, but don’t implement
    - Show queriability
    - Contributors:
      - CC
      - Rather not: MI (but will proof-read and check)
      - CO
      - EA
      - Rather not: KG
      - ?Besim => to be contacted by MI
      - ?Fahad => to be contacted by CC

Chapter proposal: MWE in lexical resources [mentioned, to be decided at next call]

Msg. by Verginica Barbu Mititelu
Postponed to next call

Overall consolidation [=> COLING?]

- CC: I would like to simplify the diagram
    - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/img#suggested-simplification
        - Replace ContextualRelation with frac:Observation, superclass also for Attestation, Frequency, Embedding and Similarity
        - Merge frac:locus and frac:corpus, define it as property of frac:Observation
        - Rename either frac:quotation or frac:attestationGloss to rdf:value (=> can be inherited from frac:Observation)
- Discussion (briefly):
    - Max (last time): looks appealing
    - Fahad (this time): looks appealing
    - We need to check all previous modelling (CC: changes only concern attestation)
    - Fahad: “observation” may be a misnomer, embeddings may be an aggregation over observations rather than a single observation
    - Several questions on whether embeddings are observations
    - Christian: likewise, “corpus” (for locus of attestations) may be a misnomer. It is, however, defined as “any body of primary data”
    - Christian: could be the novel contribution of the COLING paper (if we go for it)
    - Consensus: use that as selling point for COLING paper
- **DONE@CC**: provide OWL model of FrAC
    - https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/owl/frac.ttl
- **POSTPONED**:
    - sieve through ontology
    - establish/track vs:term_status
    - Frac:corpus
        - Does it have to be obligatory for an observable?
        - Extend the definition so that we can also point into dictionaries that provide collocation dictionaries?
        - This was the original intention of using dc:source, instead.
    - **TODO@all**: add/check issues in https://github.com/ontolex/frequency-attestation-corpus-information/issues
        - For fixes, create a pull request

Minutes 2022-04-28

Telco link:	https://meet.google.com/mjn-svyu-fdy (one-time link)
Time:	12:00–13:00 CET
Draft:	Github
Diagram	https://github.com/ontolex/frequency-attestation-corpus-information note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, GU Frankfurt
FK - Fahad Khan (excused)
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, GUF
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou
EA - Elena Apostol

GlobaLex paper

https://overleaf.com/5768717378vzdnrzyxvfxn

submitted

COLING paper?

17.5.?

- Anonymous!
    - Private github?
- Overview over the entire model
    - Short paper? Small focused contrib
- ? use cases that involve multiple aspects of the current modelling
    - Attestation
    - Frequency
    - Embeddings
    - Collocations
    - ?Links with other ontolex vocabularies
    - Think until May 2nd, evening

Use case:

- data from some shared task?
- Two data sets from similar shared tasks?
- Max: Coref OntoNotes?
    - Attestations (PTB)
    - Lexical data (sense groupings, propbank)
    - Embeddings: coref features [=> not a task, but a system]
        - AutoExtend
    - Similarity clusters: brown clusters?
    - Referent detection (collocations?)
    - Benefit: improved re-usability
- Resource integration:
    - AutoExtend: WordNet+Embeddings
    - OntoNotes: sense grouping + attestation
    - Wortschatz collocation data over PTB? collocation, frequency, attestation
    - Brown clusters over PTB
    - Problem: cannot distribute PTB
    - Work on a single PTB file?
    - Evaluation: we can query across all those resources
      => work on conversion

- ? overall consolidation [restructure of diagram to make it more readable etc.]
    - More like a motivation/position paper
        - "Linking language resources with OntoLex-FrAC"
            - https://overleaf.com/8837131433wpdbfddjmntz
            - Coordination call: Wed May 4, 12:00 CEST (Berlin)
        - Cf. resource integration, but don't implement
        - Show queriability
        - Contributors:
            - CC
            - Rather not: MI (but will proof-read and check)
            - CO
            - EA
            - Rather not: KG
            - ?Besim => to be contacted by MI
            - ?Fahad => to be contacted by CC

Overall consolidation [=> COLING?]

- Todo: provide OWL model of FrAC
    - => COLING
- CC: I would like to simplify the diagram

  - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/img#suggested-simplification
  - Replace ContextualRelation with frac:Observation, superclass also for Attestation, Frequency, Embedding and Similarity
  - Merge frac:locus and frac:corpus, define it as property of frac:Observation
  - Rename either frac:quotation or frac:attestationGloss to rdf:value (=> can be inherited from frac:Observation)
- Frac:corpus
  - Does it have to be obligatory for an observable?
  - Extend the definition so that we can also point into dictionaries that provide collocation dictionaries?
  - This was the original intention of using dc:source, instead.

Minutes 2022-04-14

Telco link: https://meet.google.com/yeq-fsgu-pzy (if it doesn't work, look here for updated link)
Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg (excused)
CC - Christian Chiarcos, GU Frankfurt
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, GUF
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou

GlobaLex paper status

- GlobaLex Paper status
    - https://de.overleaf.com/5768717378vzdnrzyxvfxn
        - NOTE: a fragment on metrics in main.tex, but actual draft in abstract.tex !
        - Too long, but not complete yet

Other

- COLING paper postponed
    - Todo: provide OWL model of FrAC
- Consolidation postponed

Minutes 2022-03-31

Telco link: https://meet.google.com/yeq-fsgu-pzy (if it doesn't work, look here for updated link)
Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg (excused)
CC - Christian Chiarcos, GU Frankfurt
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, GUF
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică

Remaining collocation things

- GlobaLex Paper status
- Updates from modelling excercises (FK+BK)
- Status other data
    a. CC: ePSD2: yet to come [postponed to summer]
    b. Wortschatz: sample in github draft
        i. Requirement: distinguish left and right context. Can be solved by
            1. **Either** pattern/query [=> novel vocabulary needed]
            2. **Or** defining two different subclasses of Collocation that specialize for different context windows [does not require any new vocabulary]
    c. Sketch engine
        i. ?Ranka, ?Max

Overall consolidation

- COLING paper status
    - Todo: provide OWL model of FrAC
- CC: I would like to simplify the diagram
    - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/img#suggested-simplification
    - Replace ContextualRelation with frac:Observation, superclass also for Attestation, Frequency, Embedding and Similarity

        - Merge frac:locus and frac:corpus, define it as property of frac:Observation
        - Rename either frac:quotation or frac:attestationGloss to rdf:value (=> can be inherited from frac:Observation)
    - Frac:corpus
        - Does it have to be obligatory for an observable?
        - Extend the definition so that we can also point into dictionaries that provide collocation dictionaries?
            - This was the original intention of using dc:source, instead.

Syntactic Patterns and Corpus queries

- See minutes 2022-02-17 and 2022-02-03
- Proposal: https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md

Minutes 2022-03-17

Telco link: https://meet.google.com/yeq-fsgu-pzy (if it doesn't work, look here for updated link)
Time: 12:00–13:00 CET
Draft: Github
Diagram https://github.com/ontolex/frequency-attestation-corpus-information
note that we cannot fully disable internal caching for the diagram, so it might take a few days to update after changes

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, GU Frankfurt
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, GUF
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică
KG - Katerina Gkirtzou

Publications

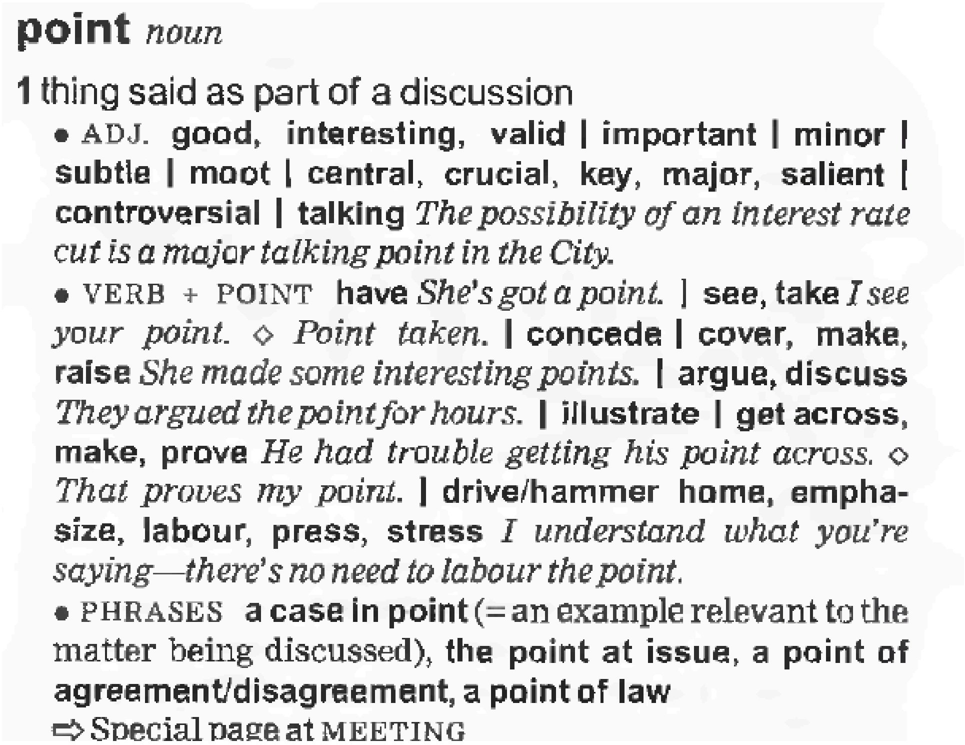
1. collocation modelling for a lexicography ws (GlobaLex) based on ODZIC, ?Wortschatz, ?Fahad, ?Besim
    - https://de.overleaf.com/5768717378vzdnrzyxvfxn
    - Deadline: https://globalex2022.globalex.link/lrec2022/
        - April 8th, 1,000-word abstract
        - May 27: full or short LREC paper
        - authors
            - Christian: example ODZIC, overall structure, poss. Wortschatz, etc.
            - Max: -"-, contrib metrics; SketchEngine
            - Fahad: example Oxford Dictionary of Collocations
            - Ciprian: collocation metrics
            - Besim: background collocations, rel. Research, contrib to metrics; usage of CWB
            - hesitant: katerina
        - Structure, todo@CC => overleaf
            - Background: CC,MI, BK
            - Collocations in FrAC:
                a. CC+MI+all [definitions + diagram
                b. Metrics: CT+others - I think that the MI^2 and MI^3 are wrong, as presented in Rychly 2008. I base this assessment on Role 2011
            - Application to Data

            a. ODZIC: CC,MI (if not already in preceding section)
            b. Ox: FK, BBI comb. Dict: BS
        - Lexicographic workflows:
            a. Sketch Engine, CQP/CWB: Max, Besim
        - Discussion and outlook
            a. CC+all
    - does julia join? (passively)
        - prefers to be mentioned in acknowledgments
2. FrAC overview @ COLING
    - https://de.overleaf.com/8837131433wpdbfddjmntz
    - Deadline May 17
    - To be discussed after Globalex submission
    - We should provide an OWL model of FrAC (also for internal validation)

Collocations

- Ciprian: Status collocation scores
    - N to be fixed (total number of tokens, not weight)
    - Provide both probabilities and O-notation
    - **Todo**: All to be double-checked:
        - Also to be out into overleaf [=> doc link into markdown]
        - First-pass Ciprian => primarily Manning [until Wednesday]
        - Second-pass Katerina => after Wed
        - Third-pass Max => internal slides
        - Paper reference(s) for every score!
    - Max: no tools for calculation provided, just names
        - Tool-specific scores won't be punished if they have variant formulas
    - If someone spots anything exotic, please mark it as such, we can discuss removal
        - Also mark additions

- Fahad's example
    - example of the modelling of part of an entry from the Oxford Dictionary of Collocations



point *noun*

1 thing said as part of a discussion

- ADJ. **good, interesting, valid | important | minor | subtle | moot | central, crucial, key, major, salient | controversial | talking** *The possibility of an interest rate cut is a major talking point in the City.*
- VERB + POINT **have** *She's got a point.* | **see, take** *I see your point.* ◇ *Point taken.* | **concede | cover, make, raise** *She made some interesting points.* | **argue, discuss** *They argued the point for hours.* | **illustrate | get across, make, prove** *He had trouble getting his point across.* ◇ *That proves my point.* | **drive/hammer home, emphasize, labour, press, stress** *I understand what you're saying—there's no need to labour the point.*
- PHRASES **a case in point** (= an example relevant to the matter being discussed), **the point at issue, a point of agreement/disagreement, a point of law**

⇨ Special page at MEETING

```
# for the purposes of this example we define a new subclass of lexicographic component
# this represents groupings of collocations which we find in dictionaries

:CollocationPattern rdfs:subClassOf lexicog:LexicographicComponent .

# the point_entry individual is a container for the collocation information contained in the dictionary enty
# it is linked to the :point lexical entry via the :describes property
# in this encoding we mention two collocation pattern sub components (see above)

 :point_entry a lexicog:Entry;
  lexicog:describes :point;
  lexicog:subComponent :point_colloc_pattern_1, :point_colloc_pattern_2 .

# we only have encoded one of the senses here
# I don't know where the definition goes
# or whether its better to make this a LexicalConcept
# and use SKOS definition

:point a ontolex:LexicalEntry;
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:sense :point_sense_1 .
```

```
:point_sense_1 a ontolex:LexicalSense .

#the next few entries are collocates of point

:have a ontolex:LexicalEntry;
  lexinfo:partOfSpeech lexinfo:verb .

:see a ontolex:LexicalEntry;
  lexinfo:partOfSpeech lexinfo:verb .

:take a ontolex:LexicalEntry;
  lexinfo:partOfSpeech lexinfo:verb .


# here we define the first two collocation patterns in the entry
# the first one is point and adjectives
# the second one is verb and point

:point_colloc_pattern_1 a :CollocationPattern;
  dct:description "ADJ. " .

# the second collocation pattern is broken up into two others
# the first deals with have, the second see and take
:point_colloc_pattern_2 a :CollocationPattern;
  dct:description "VERB + POINT";
 lexicog:subComponent :have_coll, :see_take_coll .

# the next level of collocation patterns is described here
# these are related to the frac:collocations defined below
:have_coll a :CollocationPattern;
 lexicog:describes :col_1 .
:see_take_coll a :CollocationPattern;
 lexicog:describes :col_2_1, :col_2_2  .

:col_1 a <http://www.w3.org/ns/lemon/frac#Collocation>;
  lexinfo:example "She's got a point";
 <http://www.w3.org/ns/lemon/frac#head> :point_sense_1;
  rdf:first :point_sense_1;
  rdf:rest [ rdf:first :have;
  rdf:rest rdf:nil ] .

:col_2_1 a <http://www.w3.org/ns/lemon/frac#Collocation>;
```

```
  lexinfo:example "I see your point";
<http://www.w3.org/ns/lemon/frac#head> :point_sense_1;
  rdf:first :point_sense_1;
  rdf:rest [ rdf:first :see;
  rdf:rest rdf:nil ].


col_2_2 a <http://www.w3.org/ns/lemon/frac#Collocation>;
  lexinfo:example "Point taken";
 <http://www.w3.org/ns/lemon/frac#head> :point_sense_1 ;
  rdf:first :point_sense_1;
  rdf:rest [ rdf:first :take;
  rdf:rest rdf:nil ].
```

Modelling ok, minor remarks:

- Works, and almost equivalent with ODZIC example
- illustrates use of lexicog module for structuring
- Replace list modelling (rdf:first, etc.) by seq modelling (rdf:_1, etc.)
- LexicographicPattern also used for distinguishing , and | separators
- Maybe do not provide a named class :LexicographicPattern, but instead use lexicog:LexicographicComponent directly (, vs. | is more like a semantic difference, but pattern sounds very syntactic)

- Besim’s 4 types
    - we need excerpts of a real-world resource for each of these examples
    - (How) does Fahad’s example, Julia’s, Wortschatz and Google n-grams map to the 4 types?
    - **TODO@Besim+FK**
        - Find and model example data for “4 types”, for own data, see minutes 2022-02-17 => add to overleaf sect.
        - Discuss, results to be reviewed at next meeting
            - Date is important if we want to include this in the GlobaLex paper
            - Besim might not be available next time

Minutes 2022-03-03

Venue: telco, **NEW** link
https://meet.google.com/yeq-fsgu-pzy (one time link only)

Time: 12:00–13:00 CET

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, GU Frankfurt
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, GUF
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică

Agenda:

- Wrapup Collocations (except for syntactic patterns/corpus queries)
- Aob

Collocations:

- Draft updated for current consensus (done, CC): https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#collocations
    - Draft diagram under https://github.com/ontolex/frequency-attestation-corpus-information (note that we cannot fully disable internal caching for the diagram, so the cscore property appeared in the image only a few days *after* updating the diagram)
        - BK: some collocations might be symmetric, so no head
            - CC: head is optional so that won't be a problem
        - MI: score should also be optional.
            - **TODO@CC**: Change definition of Collocation "and characterized by their collocation weight (frac:score)" → "and can be characterized by their collocation weight (frac:score)" [done]
        - FK: corpus property is required but sometimes corpus is unknown/undefined
            - CC: open world assumption allows to overcome this
            - FK: but validation would fail then. It's possible to create a dummy corpus?
            - **TODO@FK**: look at the front matter of the Oxford collocation dictionary and try creating a dummy corpus and see if this makes sense

        - CC: maybe just use the print dictionary (=> DOI [as URL] or URN) as “corpus”, but then “corpus” may need to be redefined).
            - This will create a situation where a dictionary is a corpus. **TODO**: return to this when the draft is ready
        - **POSTPONED UNTIL FINAL WRAPUP**: reconsider this, see what can be done. Maybe reconsidering Corpus, frac:corpus definition? (this is not specific to collocations, but applies throughout, so postponed)
    - **FK:** redundancy of specifying corpus for each collocation
        - **CC**: using subclasses for specific corpora (like with frequencies).
        - **TODO**: fix EPSDFrequency example in the draft (hasValue URI) (and all the others <>-URIs in examples)
        - **TODO**: provide analoguous example for corpus-specific collocations [added note in draft]
    - **TODO**: add symmetric or asymmetric to the description of different scores [done for relative frequency]
- Modelling
    - model/update examples OZDIC (from Julia):
        - CC: Done, see https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/samples/collocations/odzic/odzic.md
            - No major difficulties, some presentational issues (grouping) should be handled via lexicog:Component
        - FK: did analoguous modelling structure with lexicog (=> example, discussing that postponed to next week)
- Querying/generation
    - CC: Confirmed this can be queried (for ODZIC): https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/samples/collocations/odzic/queries.md
        - Queries are complex, so these are not tailored to end users but to developers
        - CC: cannot be simplified with direct transitions, we would need to reify, and these queries would be equally complex
        - FK complex queries, maybe simplify with having a string pattern and then aligning variables
        - CC: yes, this is exactly the idea of the syntactic pattern/query extension

Summary on current draft:

https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#collocations

modulo minor changes identified above and ContextualRelation (which we excluded), this is **approved**

- “Inverse head” property:
    - **OLD TODO@Besim**: Is there any external data set that demonstrates the need for a property that points from observable (lexical entry) to collocation?
        - And that is different from (the inverse of) head and (the inverse of) rdfs:member?
            - Head: pointing from one collocation to at most one of its member words, needed for asymmetric metrics
            - rdfs:member: pointing from one collocation to all its member words
        - And for use cases that are *cannot* be covered by lexicog:subComponent
            - pointing from one lexicog entry to one lexical entry and **selected** collocations to be displayed with the lexical entry
            - Must be information that goes beyond lexicography (dictionary layout/structure)
                - unless such data is found, we do not introduce such a property. We can query ^rdfs:member or ^frac:head, instead. (Lexicographic use case is handled by lexicog, e.g., *lexicog:subComponent*).
                    - Can we confirm this **now ?**

        => we skipped that during discussion of overall model for collocations, but the overall model for collocations was approved, so, I (CC) take this to mean that we have no such data at the moment **and thus do not introduce such a property**

- Updates on the list of collocation metrics (frac:cscore sub-properties, https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#collocations)
    - datatype properties [extensible list], frac:cscore subproperty of rdf:value
    - Ciprian: skipgrams, other association metrics => Document with new Metrics: https://docs.google.com/document/d/1pKOEjiLIVZQKkcebkSdw-bkAgPPjbwFE/edit?usp=sharing&ouid=103970418235573946532&rtpof=true&sd=true
    - **TODO@Ciprian**: integrate into draft
    - Note: do not put all properties into the diagram, only frac:cscore
- Aob:
    - CC: we should do a paper on collocation modelling for a lexicography ws (GlobaLex!) based on ODZIC, ?Wortschatz, ?Fahad, ?Besim
        - **TODO@CC**: overleaf. DONE: https://de.overleaf.com/5768717378vzdnrzyxvfxn [done]
            - TODO@MI: ask julia to join
    - FrAC overview
        - BK: COLING? [May 17]
        - **TOOD@CC**: overleaf. DONE: https://de.overleaf.com/8837131433wpdbfddjmntz
    - Next telco on collocation examples, again, we postpone queries/syntactic patterns

Minutes 2022-02-17

Venue: telco, **NEW** link
https://meet.google.com/yeq-fsgu-pzy (one time link only)

Time: 12:00–13:00 CET

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, GU Frankfurt
JBG - Julia Bosque-Gil, UZAR
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, GUF
TD - Thierry Declerck
MP - Matteo Pellegrini
CT - Ciprian-Octavian Truică

Agenda:

- Update on Collocations
- Corpus queries: postponed
- Aob

Collocations:

- CC: Description and examples in GitHub updated according to results of the last call: https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#collocations
  Consensus: see last time

  Open todos for collocations:

    - Question: Do we need an inverse head property?
        - To mark which collocations are to be listed with an observable
        - No cardinality restrictions, not a substitute for head
        - Suggestion last week:
            - unless a need for such a property is demonstrated that goes beyond lexicography (dictionary layout/structure), we do not introduce such a property. We can query ^rdfs:member, instead. (Lexicographic use case is handled by lexicog, e.g., *lexicog:subComponent*).
        - Decision to be confirmed ~~**today**~~ **next time**
        - https://jamboard.google.com/d/1naSvzOCxO1elqO0tb_qaqCCekqrWzk6Wx0xjmfe4bFk/edit?usp=meet_whiteboard
        - **TODO@Besim**: find data that demonstrates the need
  - List of collocation metrics?
    - **TODO@all**: find more metrics, check whether symmetric, check formulas
        - Ciprian: skipgrams, other association metrics

        - Draft for cscore properties by CC (added as such to https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#collocations):
            - frac:pmi: pointwise mutual information (Sketch Engine: “MI-score”), “association ratio”
            - frac:mi3: SketchEngine: modified pmi score, see https://www.sketchengine.eu/wp-content/uploads/2015/03/Lexicographer-Friendly_2008.pdf
            - Frac:pmi_logfreq: SketchEngine, see https://www.sketchengine.eu/wp-content/uploads/2015/03/Lexicographer-Friendly_2008.pdf, https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf “MI.log-f”, formerly “salience”
            - frac:student_t: t-score
            - frac:chi_sq: Chi²
            - frac:likelihood_ratio (log-likelihood)
            - frac:rel_freq: collocation frequency relative to head frequency (note that SketchEngine returns percent rather than the actual score)
            - frac:dice Dice coefficient, https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf
            - frac:logDice (SketchEngine, https://www.sketchengine.eu/wp-content/uploads/2015/03/Lexicographer-Friendly_2008.pdf, this is the SketchEngine default score)
            - frac:minSensitivity (Pedersen, Dependent Bigram Identification, in Proc. Fifteenth National Conference on Artificial Intelligence, 1998, https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf)
            - We omit all scores based on grammatical relations (cf. https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf)
            - Datatype properties [extensible list]
                - All subproperies of frac:cscore, a subproperty of rdf:value
                - Note: do not put all properties into the diagram, only frac:cscore
    - Examples OZDIC (Julia): apply, analyze => ***planned for end of Feb***
                - **todo@CC**: add attestation information to RDF snippet
    - Examples Besim (“4 types”):
        - **TODO**: we need excerpts of a real-world resource for each of these examples
            - Type 1 (fixed/idiom), “not see the wood for the trees”
            - Type2 (typical collocations, more than two words)

              “be free to choose”

“see the point”
“by the light (of the moon)”
“a beam/ray of light”
“watch a film” vs. “*see a film”

- Type3 (typical collocations, two words)
  “see danger”
  “heavy rain” /vs./ *”strong rain”
  “strong wind” / *”heavy wind”
  “strong taste”
  “strong coffee”
  “strong tea”
- Type 4 (free combinations)
  “high temperature”
  “sandy beach” (domain specific / context)
  “buy a house/book”
- **TODO@FK** (in two weeks): find examples in real-world dictionaries and start working on them
- **TODO@Besim:** same thing for other dictionaries

- For querying/generation (in SPARQL)
  - See last time
  - Note: postponed until we have example data
    - **TODO@CC**: test [not done yet]
- Other data
  d. CC: ePSD2: yet to come [not before mid-March]
  e. Wortschatz: sample in github draft
     1. Requirement: distinguish left and right context
  f. Sketch engine
     1. Ranka, maybe Max
-
- Further discussion points postponed to next call

Corpus queries: [postponed]

- CC: updated diagram and description under https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md
- **TODO**: example, e.g., SketchEngine

Minutes 2022-02-03

Venue: telco, **NEW** link
https://meet.google.com/yeq-fsgu-pzy (one time link only)

Time: 12:00–13:00 CET

Participants (please list yourself with initials; optional: affiliation)
CC - Christian Chiarcos, GU Frankfurt
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
MI - Maxim Ionov, GUF
TD - Thierry Declerck
MP - Matteo Pellegrini

Agenda:

- Collocations
- Corpus queries
- Publication plans
- Aob

Minutes

2. Collocations
   Consensus
    - Goal: **Both** collocation dictionaries **and** corpus metrics (DB or on-the-fly)
        - Collocation is intentionally ambiguous: frequency counts (“anglophone”) vs. multi-word expression (“continental”), applicable to both kinds of resources. In the context of frequency and corpus information, the frequency sense is more prominent
    - Collocations are collections of observables
        - unordered collocation: rdf:Bag (rdfs:member properties)
        - Ordered collocation: rdf:Seq (rdfs:member and rdf:_nnn)
    - Collocations are observables (can have attestations, frequency, embeddings)
      => nested collocations
    - Collocations can have a(t most one) head
        - If collocation *metrics* are asymmetric, this is the element that the metrics are about; Head marking is optional
        - Note: “head” motivated by “head word” in dictionaries, not to be confused with head used in syntax nor with any term from collocation studies
    - We need syntactic patterns (as for Julia’s and Ranka’s examples)
    - Not to be confused with colligation, just filters for syntactic criteria
    - Collocation of lexical entries? (= current modelling)
        - Seq of underspecified lexical entries: instead of a LexicalEntry, we just write a blank node with a POS)
        - Works, doesn’t require extensions, is limited, no support for gaps
    - synsem? (https://www.w3.org/2016/05/ontolex/#syntax-and-semantics-synsem)
        - Rather not: collocations are not necessarily syntactic. We cannot specify order
    - Use an external mechanism to specify patterns

- Pattern: string (contains a series of variables)
- Variables: list of variables, sequentially aligned with a sequence in the collocation
- For patterns, we can (but don’t have to) use corpus query languages, e.g., CQP
- Consensus here: postpone after discussing corpus queries

Open todos for collocations:

- Question: Do we need an inverse head property?
  - To mark which collocations are to be listed with an observable
  - No cardinality restrictions, not a substitute for head
  - Suggestion last week:
    - unless a need for such a property is demonstrated that goes beyond lexicography (dictionary layout/structure), we do not introduce such a property. We can query ^rdfs:member, instead. (Lexicographic use case is handled by lexicog, e.g., *lexicog:subComponent*).
  - Decision to be confirmed **at next call**
- List of collocation metrics?
  - addenda by CC (after last call; added as such to https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#collocations):
    - frac:pmi: pointwise mutual information (Sketch Engine: “MI-score”), “association ratio”
    - frac:mi3: SketchEngine: modified pmi score, see https://www.sketchengine.eu/wp-content/uploads/2015/03/Lexicographer-Friendly_2008.pdf
    - Frac:pmi_logfreq: SketchEngine, see https://www.sketchengine.eu/wp-content/uploads/2015/03/Lexicographer-Friendly_2008.pdf, https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf “MI.log-f”, formerly “salience”
    - frac:student_t: t-score
    - frac:chi_sq: Chi^2
    - frac:likelihood_ratio (log-likelihood)
    - frac:rel_freq: collocation frequency relative to head frequency (note that SketchEngine returns percent rather than the actual score)
    - frac:dice Dice coefficient, https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf
    - frac:logDice (SketchEngine, https://www.sketchengine.eu/wp-content/uploads/2015/03/Lexicographer-Friendly_2008.pdf, this is the SketchEngine default score)

        - frac:minSensitivity (Pedersen, Dependent Bigram Identification, in Proc. Fifteenth National Conference on Artificial Intelligence, 1998, https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf)
        - We omit all scores based on grammatical relations (cf. https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf)
        - Datatype properties [extensible list]
            - All subproperies of frac:cscore, a subproperty of rdf:value
            - Note: do not put all properties into the diagram, only frac:cscore
- Examples OZDIC (Julia): apply, analyze
    - ***Discussion planned for call end of Feb***
    - Tbc: Were these satisfactorily solved? (see RDF below)
    - Tentative consensus: yes, except for syntactic pattern
        - grouping in the dictionary/ selection of collocations can be modeled within lexicog
        - Example sentences: attestation
            - **todo@CC**: add attestation information to RDF snippet
- Examples Besim ("4 types"):
    - **TODO**: we need excerpts of a real-world resource for each of these examples
        - Type 1 (fixed/idiom), "not see the wood for the trees"
        - Type2 (typical collocations, more than two words)

          "be free to choose"
          "see the point"
          "by the light (of the moon)"
          "a beam/ray of light"
          "watch a film" vs. "*see a film"

        - Type3 (typical collocations, two words)

          "see danger"
          "heavy rain" /vs./ *"strong rain"
          "strong wind" / *"heavy wind"
          "strong taste"
          "strong coffee"
          "strong tea"

        - Type 4 (free combinations)

          "high temperature"
          "sandy beach" (domain specific / context)
          "buy a house/book"

        - **TODO@FK** (4 weeks): find examples in real-world dictionaries and start working on them
- For querying/generation (in SPARQL)
    - Note: we need example data first
        - Functionalities to be tested:

            - Can we retrieve : by looking onto some examples
                - Symmetric case: rdfs:member
                - Single head: frac:head
                - N:m case: ?*frac:collocation* to point from observable to Collocation
            - Can we generate (see algorithm above, TODO@CC)
        - Get all members
        - Get their numbered properties
        - Use BIND(STR(prop)) to strip position from numbered properties
        - Order according to position
        - **TODO@CC**: test [not done yet]
    - Other data
        a. CC: ePSD2: yet to come [not before mid-March]
        b. Wortschatz: sample in github draft
            1. Requirement: distinguish left and right context
        c. Sketch engine
            1. Ranka, maybe Max
    - Suggestion:
        - Modelling Besims and checking Julia's examples on real-world data
        - To be discussed in 4 weeks
        - Other than that, we consider collocation to be more or less stable (no new features) => next time we discuss sth else

3. Modeling corpus queries
   https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/ideas/query.md
    - To model syntactic patterns by means of an established (external) formalism
        - Can represent order, alternations, context restrictions
        - Take CQP as an example, but don't create a dependency
    - QueryResponse object
        - Query string (rdf:value)
        - ID for query language (frac:queryLanguage)
        - (JSON) array of variable names (frac:variables)
        - Links to responses (frac:result) -> frac:Collocation (or, more generally, a new superclass, frac:Cooccurrence); if this is a sequence, we can align it with variable names from frac:variables
    - FK: suggestion to keep Collocation (sub Cooccurrence)
    - TD: possible ties to synsem discussion
    - CC: also cf. LD4LT

4. Publication plans
   Steps towards community report?
   Start writing by June, feature freeze by early May, May: discuss simplifications/naming, esp., "Observation" as generalization over things linked with an Observable.

5. Aob
   Next call: Two weeks, 12:00.

Minutes 2022-01-20

Venue: telco, **NEW** link
https://meet.google.com/yeq-fsgu-pzy (one time link only)

Time: 11:00–12:00 CET

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, GU Frankfurt
JBG - Julia Bosque-Gil, UZAR
FK - Fahad Khan
GS - Gilles Sérasset - Grenoble
RS - Ranka Stanković
Excused: MI - Maxim Ionov, GUF

Agenda:

- SWJ paper
- Wrapup collocations
- Publication plans
- aob

SWJ paper:

- Fahad: Nexus T1.1 paper accepted at SWJ, need to update section on FrAC

Wrapup collocations:

- Goal: **Both** collocation dictionaries **and** corpus metrics (DB or on-the-fly)
- Collocations are collections of observables
    - unordered collocation: rdf:Bag (rdfs:member properties)
    - Ordered collocation: rdf:Seq (rdfs:member and rdf:_nnn)
- Collocations can have attestations
    - => collocation are observables
      (=> nested collocations)
    - Tbc: are we ok with that?

      - FK: +1
  - Collocations can have a head
    - If collocation metrics are asymmetric, this is the element that the metrics are about
    - rdfs:Seq/Bag + at most one frac:head
      - Note: “head” motivated by “head word” in dictionaries, not to be confused with head used in syntax nor with any term from collocation studies
    - Head marking is optional
    - there is at most one single head
  - As these questions raised no objections, we go on with these modelling decisions for the moment

Notes:

- Ranka: need syntactic patterns
  - With the current modelling, we can express combinations of lexemes and parts of speech (i.e., a sequence of lexical entries; these may have up to one part of speech)
  - Need confirmed by Ranka’s and Julia’s examples
  - Not to be confused with colligation, just filters for syntactic criteria
  - Cf. https://www.w3.org/2016/05/ontolex/#syntax-and-semantics-synsem
    - (not fully clear how to use it, need an example)
    - Possibly, synsem:SyntacticFrame could just be an observable (no collocation required)
- Gilles: Collocation is ambiguous: frequency counts (“anglophone”) vs. multi-word expression (“continental”)
  - No firm conclusion, but depends on resource
  - More frequency sense, but intentionally left ambiguous
- Open problem: sequences with “gaps”
  - Approx: left and right context?
    - Howto model?
  - Wild cards?
    - Need real-world example
  - Machine-readable version of CQP?
    - TODO@Max? CQP operators
    - Queries could be written as collocations, then !?

POSTPONED: todos/questions:

  - For querying/generation (in SPARQL)
    - Functionalities to be tested:
      - Can we retrieve : by looking onto some examples
        - Symmetric case: rdfs:member
        - Single head: frac:head
        - N:m case: ?*frac:collocation* to point from observable to Collocation
      - Can we generate (see algorithm above, TODO@CC)
    - Get all members
    - Get their numbered properties

        - Use BIND(STR(prop)) to strip position from numbered properties
        - Order according to position
        - **TODO@CC**: test [not done yet]

    - Do we need an inverse head property?
        - To mark which collocations are to be listed with an observable
        - No cardinality restrictions, not a substitute for head

- Examples OZDIC (Julia):
    - Examples below: apply, analyze
        - Were these satisfactorily solved?

- model Besim’s 4 types

Type 1 (fixeed/idiom)

not see the wood for the trees (fixeed/idiom)

Typ2 (typical collocations, more than two words)

be free to choose
see the point
by the light (of the moon)
a beam/ray of light

watch a film / vs./ *see a film

Typ3 (typical collocations, two words)

see danger

heavy rain /vs./ *strong rain
strong wind / *heavy wind

strong taste
strong coffee
strong tea

Type 4 (free combinations)

high temperature

sandy beach (domain specific / context)

buy a house/book

- Other data
    - ePSD2: yet to come
    - Wortschatz: sample in github draft
    - Sketch engine

Publication plans
Steps towards community report?

Any other business

- Goodbye and thanks to Julia !
- Next call: Two weeks, 12:00.

Minutes 2021-12-xx

Venue: telco, **NEW** link
https://meet.google.com/yeq-fsgu-pzy (one time link only)

Time: 11:00–12:00 CET

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
JBG - Julia Bosque-Gil, UZAR
MI - Maxim Ionov, GUF

Minutes to be recovered …
CC: there was a December call that I missed.

Minutes 2021-11-25

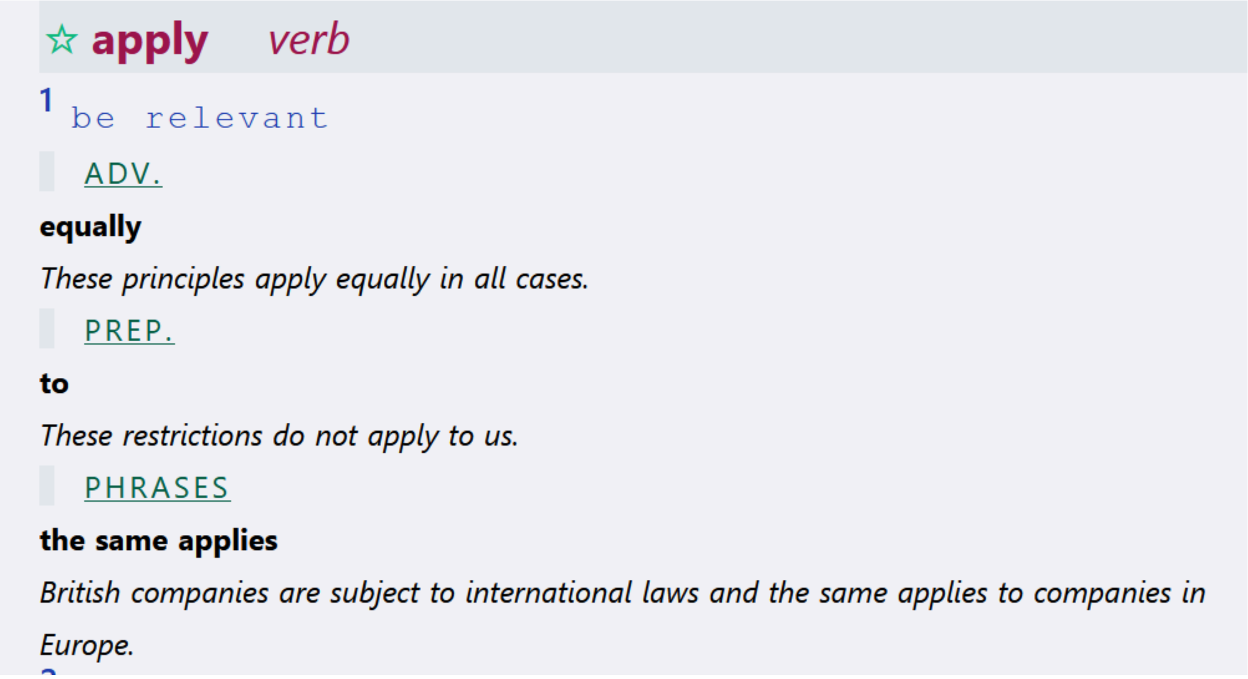
Venue: telco, **NEW** link
https://meet.google.com/yeq-fsgu-pzy (one time link only)

Time: 11:00–12:00 CET

Participants (please list yourself with initials; optional: affiliation)
BK - Besim Kabashi, FAU Erlangen-Nürnberg
CC - Christian Chiarcos, GU Frankfurt
JBG - Julia Bosque-Gil, UZAR
Excused: MI - Maxim Ionov, GUF

- wrapup
    - Besim:
        - collected examples, four types discussed, different degrees of freedom
        - How to model collocations with more than two parts
            - Possibilities
                - Mutliple modifiers pointing to the head
                    - We cannot express order, then
                    - The same head cannot occur in multiple collocations
                - list /agregator object, modifiers and head pointing to it
        - More a technical problem, not a linguistic one
        - Suggestion: brain storming during an afternoon, then cut, running in cycles
            - CC: at a Nexus meeting?
    - Besim: Goals?
    - **Both** collocation dictionaries **and** corpus metrics (DB or on-the-fly)

- OZDIC (Julia):



☆ **apply** *verb*

1 be relevant

ADV.

equally

These principles apply equally in all cases.

PREP.

to

These restrictions do not apply to us.

PHRASES

the same applies

British companies are subject to international laws and the same applies to companies in Europe.

:apply-v a ontolex:LexicalEntry.

:equally-adv a ontolex:LexicalEntry.

:apply-equally a frac:Collocation:

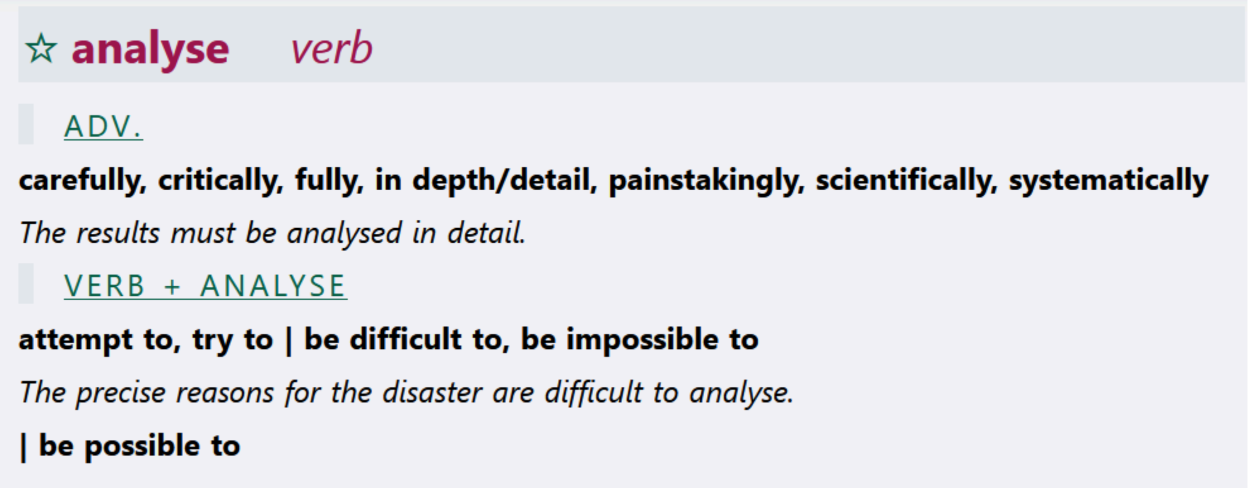
rdf:_1 :apply-v, rdf:_2 :equally-adv.

(Same procedure with "apply to")

What do we do with "the same applies"?

:the_same_applies a frac:Collocation;   (?)

rdf:_1 :the-dt, rdf:_2 :same-pron, rdf:_3 :apply-form-sg. (???)



☆ **analyse** *verb*

ADV.

carefully, critically, fully, in depth/detail, painstakingly, scientifically, systematically

The results must be analysed in detail.

VERB + ANALYSE

attempt to, try to | be difficult to, be impossible to

The precise reasons for the disaster are difficult to analyse.

| be possible to

Same situation as in the previous example BUT here we have two differentiated sets (attempt to, try to) and (be difficult to, be impossible to). Do we want to somehow “keep” those two sets differentiated?

Here the “Verb + analyse” tells us the order of the rdf container.

Julia; Questions:

How to model the example

- There is an attestation for the collocations (I think)
    - CC: +1
- Attestations go with observables, but a collocation is not an observable.
    - CC: We can make them observables, works for attestation, embedding
        - BUT: can collocations also be part of other collocations?

        - Yes; “nested collocation”: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=12FE3797A92010AD1644C05435FB0706?doi=10.1.1.14.4173&rep=rep1&type=pdf
      - JBG: “Stand in strong opposition to” -> stand in opposition to; strong opposition.
    - Cf. overall diagram https://github.com/acoli-repo/ontolex-frac/blob/master/img/ontolex-frac-2019-03.png (dated)
- In lexicog, examples go with lexical senses.
- So what do we do with the example?

NB:

- unordered collocation: rdf:Bag (rdfs:member properties)
- Ordered collocation: rdf:Seq (rdfs:member and rdf:_nnn)
- For querying/generation (in SPARQL)
    - Get all members
    - Get their numbered properties
    - Use BIND(STR(prop)) to strip position from numbered properties
    - Order according to position
    - **TODO@CC**: test

Discussion:

- Besim: can we agree of having the frac:head property plus list-based model
- Julia: not sure head is always best choice, collocation dictionaries will *sometimes* provide the same collocation, but not always, using head for that would violate uniqueness
- CC: three types
    - Fully symmetric (no head needed, just rdfs:member/_:1)
    - Single head (for asymmetric metrics, to express what the scores are about)
    - n:m relations, but not fully symmetric (this is effectively a filter over rdfs:member, EITHER: duplicate [copy] the collocation object in data modelling, OR: introduce a custom property -- if demand is demonstrated this can be added to the model)
        - Unless there is a strong case that such a custom property is needed, we stay with rdf:Seq/Bag + frac:head
- Tentative consensus
    - Collocation is a Observable
    - rdfs:Seq/Bag + at most one frac:head
        - Note: “head” motivated by “head word” in dictionaries, not to be confused with head used in syntax nor with any term from collocation studies
- To be tested
    - Julia’s data (entries from the OZDIC): could work, tbc
    - ePSD2: yet to come
    - Wortschatz: sample in github draft
    - Sketch engine
    - Besim: 4 types data [below]
    - Functionalities to be tested:

- Can we represent (i.e., revert the modelling)?: by looking onto some examples
  - Symmetric case: rdfs:member
  - Single head: frac:head
  - N:m case: ?*frac:collocation* to point from observable to Collocation
- Can we generate (see algorithm above, TODO@CC)

- Paper?
  - To publicly discuss the modelling
  - ?LREC
    - Julia contributing passively
    - Besim interested
    - Christian: interested if we find more collaborators
- Homework for next time: model Besim’s types

Type 1

not see the wood for the trees (fixeed/idiom)

Typ2 (typical collocations, more than two words)

be free to choose
see the point
by the light (of the moon)
a beam/ray of light

watch a film / vs./ *see a film

Typ3 (typical collocations, two words)

see danger

heavy rain /vs./ *strong rain
strong wind / *heavy wind

strong taste
strong coffee
strong tea

Type 4 (free combinations)

high temperature

sandy beach (domain specific / context)

buy a house/book

Minutes 2021-11-11

Venue: telco, link https://meet.google.com/wgw-otnz-woy
Time: 12:00–13:00 CET

Participants (please list yourself with initials; optional: affiliation)

- ~~CC - Christian Chiarcos, GU Frankfurt~~
- MI - Max Ionov, GU Frankfurt
- BK - Besim Kabashi, FAU Erlangen-Nürnberg
- JBG - Julia Bosque-Gil, UZAR
- FK - Fahad Khan, CNR

Agenda:

1. Collocations, examples
2. Renaming properties (postponed because CC is not in the call)

not see the wood for the trees (i.e. fixeed > idiom)

be free to choose
see the point
by the light (..of the moon)
a beam/ray of light

see danger

watch a film / vs./ *see a film

heavy rain /vs./ *strong rain
strong wind / *heavy wind

strong taste
strong coffee
strong tea

high temperature

sandy beach (domain specific / context)

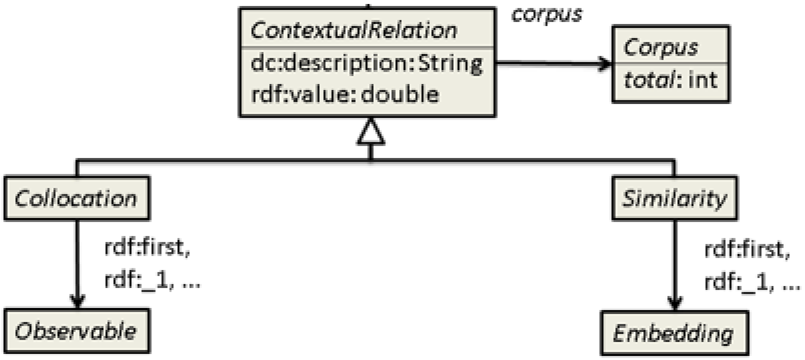
buy a house/book (i.e. free)

TODO: ACTION: pick 5 examples from the dictionary and put here

MI> Are completely free combinations considered as collocations?
BK> In some schools, yes, we might need to be able to model it

CC (from the off): possibly, the modelling of collocations should be similar to the modelling of compounds in OntoLex-Morph. Both (can) have what is sometimes called a head (cf. https://www.sfs.uni-tuebingen.de/GermaNet/documents/compounds/split_compounds_from_GermaNet16.0.txt). The modelling of this information in OntoLex-Morph is still unclear, though.

Reminder: current proposal of collocation modeling:



:kill_cf frac:hasContext :colloc1.
:colloc1 a rdf:Seq, frac:Collocation; rdf:_1 :kill_cf; rdf:_2 :switch_cf; rdf:value “0.002”; dct:Description “portion of collocations per total of :kill_cf”.

:switch_cf frac:hasContext :colloc2.
:colloc2 a rdf:seq, frac:Collocation; rdf:_1 :kill_cf; rdf:_2 :switch_cf; rdf:value “0.0305”.

“Absolute” collocation (value not specific to *one* word)
:colloc3 a rdf:seq, frac:Collocation; rdf:_1 :kill_cf; rdf:_2 :switch_cf; rdf:value “50”; dc:description “total freq”.

Minutes 2021-10-28

Venue: telco, link https://meet.google.com/wgw-otnz-woy
Time: 12:00–13:00 CET

Participants (please list yourself with initials; optional: affiliation)

- ~~CC - Christian Chiarcos, GU Frankfurt~~
- MI - Max Ionov, GU Frankfurt
- BK - Besim Kabashi, FAU Erlangen-Nürnberg
- FK - Fahad Khan. CNR
- EI - Elena Irimia, RACAI, Bucharest
- TD - Thierry Declerck, DFKI

Agenda:

3. Collocations, how to proceed
4. Renaming properties

Jamboard:
https://jamboard.google.com/d/1iVj7pfph_avBT87Nosb9v_E4xTrtFX09vpNFApW4LUA/edit?usp=sharing

Besim: candidate definitions (from last time):

1. *(original FrAC draft definition added by CC): A Collocation is a frac:ContextualRelation that holds between two or more ontolex:Elements based on their co-occurrence within the same utterance and characterized by their collocation weight (rdf:value) in one or multiple source corpora (dct:source).*

2. Collocation is the way words combine in a language to produce natural-sounding speech and writing, i.e. word associations. (Oxford Collocations dictionary for students of English, p.vii, OUP, 2002)

3. In Corpus Linguistics, a collocation is a series of words or terms that co-occur more often than would be expected by chance. (Wikipedia (EN) https://en.wikipedia.org/wiki/Collocation)

4. Collocations are partly or fully fixed expressions that become established through repeated context-dependent use.

MI> Do we use only quantitative (firthian) approach to collocations or also support the lexicographical way (British vs. continental tradition)
FK> We don't always have connection to the corpus in coll. dictionaries, so it might be good to support it
TD> Yes, it's often not present
FK> OntoLex is tightly connected to lexicography, so maybe we should support this, but the module is connected to the corpus so maybe we stick to the Firthian tradition
JBG> Yes, maybe if the module is connected to corpus tradition, we stick to that

JBG> In an example from the last time: head was unclear, there were two possible analysis grounded on different approaches, for example the second one was from the point of view of lexical selection

Definitions (postponed to next time, for CC to be here):

- frac:corpus/frac:locus vs. dc:source ?
- frac:annotationGloss?
- frac:quotation?
- (from the last time) frac:Collocation / frac:collocates ?

ACTIONS for the meeting on 11.11.2021:

- Convert a few examples from corpora query output

ACTIONS for the meeting on 25.11.2021:

- Convert a few examples from Ana Salgado's data
- Convert a few examples from Oxford collocations dictionary
- Convert a few examples from Besim's data

Next meeting: 11.11.2021, 11:00am

Minutes 2021-10-07

Venue: telco, link https://meet.google.com/wgw-otnz-woy
Time: 12:00–13:00 CET

Participants (please list yourself with initials; optional: affiliation)

- CC - Christian Chiarcos, GU Frankfurt
- MI - Max Ionov, GU Frankfurt
- BK – Besim Kabashi, FAU Erlangen-Nuremberg
- JBG - Julia Bosque-Gil, UNIZAR
- KG - Katerina Gkirtzou, ARC

Agenda:

- To the definition of collocations (BK)
- Identifying aspects to be revisited/verified before writing [postponed from last week]
- Planning the writing process

Collocations

Goal compare/align draft definitions with definitions of relevant terms such as ‘collocation’ in the literature

Besim: candidate definitions:

5. *(original FrAC draft definition added by CC): A Collocation is a frac:ContextualRelation that holds between two or more ontolex:Elements based on their co-occurrence within the same utterance and characterized by their collocation weight (rdf:value) in one or multiple source corpora (dct:source).*

6. Collocation is the way words combine in a language to produce natural-sounding speech and writing, i.e. word associations. (Oxford Collocations dictionary for students of English, p.vii, OUP, 2002)

7. In Corpus Linguistics, a collocation is a series of words or terms that co-occur more often than would be expected by chance. (Wikipedia (EN) https://en.wikipedia.org/wiki/Collocation)

8. Collocations are partly or fully fixed expressions that become established through repeated context-dependent use.

Examples:
For def 2

strong wind (Ger. Starker Wind) ,
heavy rain (Ger. Starker Regen) ;
*heavy_wind ,
strong_rain ;

For def 4

terms as 'crystal clear', 'middle management', 'nuclear family'<, and 'cosmetic surgery' are examples of collocated pairs of words.

Notes:

a. Combination of words in a language can be ranged on a cline from the totally free – see aman/car/book – to totally fixed and idiomatic – not see the wood for the trees.
b. Collocations can be in a syntactic relation (such as verb–object: 'make' and 'decision'), lexical relation (such as antonymy), or they can be in no linguistically defined relation.
c. Knowledge of collocations is vital for the competent use of a language: a grammatically correct sentence will stand out as awkward if collocational preferences are violated.
d. Collocation measures with different parameters: —> Collocations crated the measure of association based on (1) sentence level ? (2) Maximum window span of x (=5?) words? // Skip-grams

Additional literature:

- FK (last time) For an extended definition of lexical collocations (archives-ouvertes.fr) (TODO@FK: read this paper)
- JB: More bibliography here: http://www.stefan-evert.de/PUB/BartschEvert2014.pdf, Bartsch, S., & Evert, S. (2014). Towards a Firthian notion of collocation. Vernetzungsstrategien Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern, 2(1), 48-61.

Discussion

- Towards Definition: to be merged from those above
- CC: we need to state that this is an aggregate/series/group of of elements, not a way of combination (must be a class)
- Discussion: “words” [=> LexicalEntry?] ? could also be individual forms => observables?
    - PRO observable would be if not just specific lexical entries, but also specific forms [tbc with example]
    - CON: word ist most readable (and mostly, more what people would expect)
- Approach:
    - Find a resource (one dictionary entry, plus corpus-derived data)!
    - Book
        - http://docplayer.org/43558587-Uwe-quasthoff-woerterbuch-der-kollokationen-im-deutschen-berlin-new-york-walter-de-gruyter-isbn.html
        - TODO@Besim: copy sample page to github
        - https://www.google.de/books/edition/Das_A_und_O/vnd3moYdHYAC?hl=de&gbpv=1&dq=Redewendungen:+W%C3%B6rterbuch+der+deutschen+Idiomatik+(Duden&printsec=frontcover
    - Corpus-derived
        - cf. Wortschatz: https://corpora.uni-leipzig.de/en/res?corpusId=deu_news_2020&word=Halse
        - BK: dynamically created via CQP
            - CQP only returns information on pairs of words, not on larger groups

Suggested vocabulary elements:

- Collocation object
- Some encoding of sequence
    - rdf:Seq or binary property?
        - BK: cannot be binary : pro rdf:Seq

- MI: how to distiguish left and right context (CC: if ordered, by position relative to the the head)
- => Collocation would be an aggregate (rdf:Collection), user can say it's ordered, but defining it as Seq [exactly how it is right now]
- Not other encoding of sequence needed (?)

- Collocates, head properties between the elements
  - German *den Hals nicht vollkriegen* "cannot get the neck full"
  - Variants: *den Rachen nicht vollkriegen*, *die Hütte nicht vollkriegen, den Hals nicht voll bekommen*
  - Head: *Hals* (or *voll+kriegen*)
  - Intuitive modelling: Hals -collocates-> voll+, Hals -collocates-> +kriegen, Hals -collocates-> den, Hals -collocates-> nicht
  - Problems
    - Hals nicht vollkriegen
    - Analysis 1: Hals is head
      - Hals -collocates-> **vollkriegen**
    - Analysis 2: vollkriegen is head
      - **vollkriegen** -collocates-> Hals
    - Result (in RDF):
      - Hals <-collocates-> vollkriegen
      - Not clear what the scores refer to, scores of analysis 1 and 2 may be different
        - Or, are they always symmetric? --> Log ratio is not
      - To distinguish them, we need to reify
    - **Suggestion: reified collocates is exactly a collocation object**
      - **However: we need to (be able to) make the head explicit (this would support both analyses)**
      - **Property that points from collocation object to head => todo: find a name**
      - **=> general approval (tbc next time)**

=> [if approved] Answers to questions from last time:

- Collocation → Observable is an aggregation, not necessarily rdf:seq
  - container/collection, user can define it as Seq (ordered) or Bag (unordered)
- relations between collocates?
  - No, relation between collocation object and collocates or head
- "next" property? (how to encode end of collocation?)
  - No, rdf:Seq if necessary
- properties for specific metrics
  - Still unresolved, check list

Note: on property to identify the head, we had last time a tentative consensus to name it **"frac:collocationNode"** (tbc)

aspects to be revisited/verified before writing [postponed to next week]

- [postponed from last week]
  - see call from 2021-07-29
  - anything to be renamed?
    - frac:corpus/frac:locus vs. dc:source ?

        - frac:annotationGloss?
        - Frac:quotation?

TODOS:

- Next telco: 3 weeks, Oct 28th
- then : confirm abandonment of collocates property in favor of Collocation object (~ reified, n-ary "collocates" with scores attached)

Minutes 2021-09-23

Venue: telco, https://meet.google.com/wgw-otnz-woy **[NEW <WRONG LINK IN THE EMAIL, APOLOGIES>]**
Time: 12:00-13:00 CET

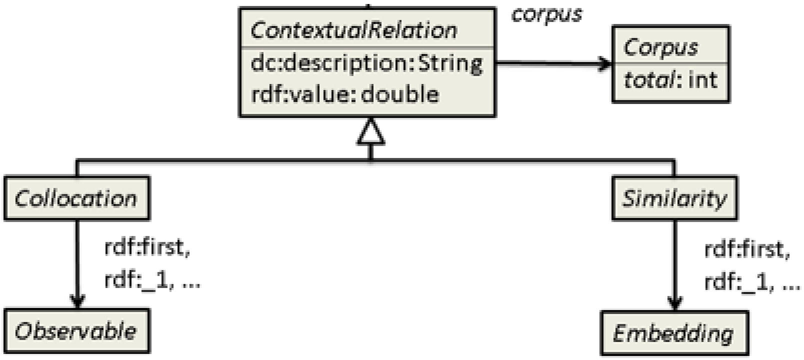
Participants (please list yourself with initials; optional: affiliation)

- CC - Christian Chiarcos, GU Frankfurt
- MI - Max Ionov, GU Frankfurt
- TD - Thierry Declerck, DFKI, Saarbruecken
- BK - Besim Kabashi, FAU, Erlangen
- KG - Katerina Gkirtzou, ARC,
- FK - Fahad Khan, CNR
- FM - Francesco Mambrini, Università Cattolica del Sacro Cuore, Milan

Agenda:

- A short overview of collocations modeling
- Input on collocations from Besim
- Planning the writing process
- Next call?

Collocations:



- *context* property
    - Problematic, because traditionally, it refers to the elements other than the base word
    - baseWord [Hausmann; German]? Node [oxford collocation dict, SketchEngine, Lancaster]?
    - Definition: **TODO@Besim**
    - **Consensus “frac:collocationNode” (!)**
    - Direction: node to collocate
        - Relation: between all words
        - BK: want direct “next” property
            - CC: how to encode the end of a collocation?
    - Direction:
        - collocate -> node: ^rdfs:member/frac:collocationNode
        - Node -> collocate: ^frac:collocationNode/rdfs:member
        - (plus Filter : node != collocate)
    - “Schuhe einlaufen” (German):
        - In this context, einlaufen has a special meaning, shoe is more important (=> collocationNode)
            - Normally, it’s about words, not other kinds of observables
            - CC: doesn't this mean that we have a collocation of a word **sense** here ?
    - TD: relation between similarity and collocation is indirect, nothing really in common, creating a common superclass may give a wrong impression
        - Maybe something to vote about
- rdf:value for metrics, subproperties for specific ones
- Collocation → Observable is an aggregation, not necessarily rdf:seq
- FK: Compare draft definitions with definitions of relevant terms such as ‘collocation’ in the literature, e.g., For an extended definition of lexical collocations (archives-ouvertes.fr) (FK volunteers to read this paper for next meeting)

Minutes 2021-09-16

Venue: telco, https://meet.google.com/zww-afdi-hdp **[NEW]**
Time: 12:00-13:00 CET

Participants (please list yourself with initials; optional: affiliation)

- CC - Christian Chiarcos, GU Frankfurt
- MI - Max Ionov, GU Frankfurt
- KG - Katerina Gkirtzou (ARC)
- JM - John McCrae, NUIG

Decision to postpone to **next week**

- Upcoming discussions require a broader consensus and more participants
  - We forgot no invitation email
- This is
  - Planning the writing process
  - Identifying aspects to be revisited/verified before writing
    - see last call
    - anything to be renamed?
    - Any more input on collocations?

Minutes 2021-07-29

Venue: telco, https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC - Christian Chiarcos, GU Frankfurt
- MI - Max Ionov, GU Frankfurt
- JBG - Julia Bosque-Gil (UZAR)
- GVO - Giedre Valunaite Oleskeviciene (MRU)
- FK - Fahad Khan (FK)
- TD - Thierry Declerck (DFKI)
- KG - Katerina Gkirtzou (ARC)

Agenda and minutes:

- Logistics:
  - JBG to step down as FrAC telco hostess -> need to change regular telco link
- Modeling requirements from SWC
- Some points for discussion:
  - CC: Discuss locus & corpus property -> maybe come up with a single property
  - CC > attestationGloss and another datatype property (maybe quotation?) -> maybe go into that discussion again to revisit the differences
    - **frac:quotation** (range: xs:String) This contains the text content of the dictionary quotation.
    - **frac:attestationGloss** (domain: frac:Attestation, range: xs:String) This contains the text content of an attestation as represented within a dictionary. *This may be different from a direct quotation because the target expression may be omitted or normalized.*
  - CC> What happens with ungrammatical examples/hypothetical, non attested examples -> maybe use lexicog:UsageExample, and stick to Observables for (observed) examples
  - CC> rdf:value for collocation, which metrics do we add. Agreement on the name of the property linking Collocation to the score? (please revise these minutes)
  - CC> How and when to write the community report? Suggestion: coordinate at/after the Zaragoza F2F. Before that we prepare the status as is.
    - Coordinate a separate telco to prepare the slides (and share the slides with everyone)
- Next call: ~~12-08-2021~~ ~~26-08-2021~~ [skipped because of physical meeting at LDK]

Minutes 2021-07-15

Venue: telco, ~~https://meet.google.com/rsx-mbkr-oxi~~ https://meet.google.com/pky-mbhm-ddd
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC - Christian Chiarcos, GU Frankfurt
- MI - Max Ionov, GU Frankfurt
- TD - Thierry Declerck, DFKI
- KG - Katerina Gkirtzou, ILSP-ARC
- FK - Fahad Khan, CNR
- GVO - Giedre Valunaite Oleskeviciene, Mykolas Romeris University
-

Agenda:

- Collocations / colligations
  - Reconsider all the places that we might want to reconsider
    - E.g. attestation locus
  - Think about when hasContext is necessary (for which metrics) and maybe whether to rename it

- check if this is still the case. Reasoner used to break on collections like this. Test if it breaks on some collocation data + some owl in the graph
- TODO: change Collocation→Observable arrow from rdf:first to an aggregation symbol, put in the documentation: “rdf:Seq is a preferred way to model this, but we are aware of the potential discussions on rdf sequences, and if there is an official decision/recommendation, it should become a recommended way” [the other way around]
-

check if this is still the case. Reasoner used to break on collections like this. Test if it breaks on some collocation data + some owl in the graph
FK: list still does cause problems, rdf:Seq doesn’t
CC: Nothing prohibits using a list and a sequence at the same time (double inheritance).
So: using rdf:List and if needed defining it as a sequence too
JMC: decomp also uses rdf:Bag

TODO: change Collocation→Observable arrow from rdf:first to an aggregation symbol, put in the documentation: “rdf:Seq is a preferred way to model this, but we are aware of the potential discussions on rdf sequences, and if there is an official decision/recommendation, it should become a recommended way” [the other way around]

We model this as a sequence, but if a user needs to reason over it, they can define a collocation instance as a List. This is not a recommended practice

Multiple collocation scores: recommended to introduce subproperties for rdf:values TODO@Max

- T-Score
- log-likelihood
- log-dice
- MI/MI2

hasContext property:
TD: In the Terminology module, we will have contexts as instances of a class. We need a property to link to those instances.
CC: what is the definition of context
TD: Is a simple sentence example that reflects the "correct" use of a term, established in TBX/Terminology (DataCat)
CC: is that different from an attestation?
TD: Very few contexts then. Not a corpus driven approach. It is a kind of attestation, in my opinion
TD: Or two sublcasses of attestation: observed, and elicited (KG: +1)
TD: Name: Context?
TD: Illustration? Example is good too. Is already in Lexicog I think
CC: elicited examples as usageExampe, maybe also in terminology
TD: It could be. We are thinking of it

As for collocations, illustrative “context” examples have different range, so “our” hasContext property will be very different. But if terminology module uses UsageExample, there is no problem with this.

JMC: avoid small words, context rather than hasContext (etc.)

hasContext:
Because of the non-symmetric scores we need to indicate “focus” words for collocations. So a collocation should have at most 1 hasContext property: pointing from a word (Observable) to a Collocation. If no asymmetric scores are used, it’s unnecessary because we can always use rdfs:member to get words of a collocation.
With the terminology discussion we recommend using lexicog:UsageExample for elicited examples (and optionally frac:Attestation for observed examples, but usageExample can be elicited or observed) and to avoid the term “context” because of the risk of confusion. Alternatively, hasContext might be renamed (if they need this word reserved)

TODO: Inverse the property hasContext

TODO: Rename it “definedFor” (for now)

Colligations: grammatical pattern illustrating the context of a word
MI: corpus query as representation of the grammatical context? => next time

TODO: update examples, descriptions and diagram for collocations to the new modelling

Next time:

- Colligations as corpus queries
- Renaming things / reassessing older module components **TODO@all**: collect ideas
- Meet in two weeks, 29.07.2021

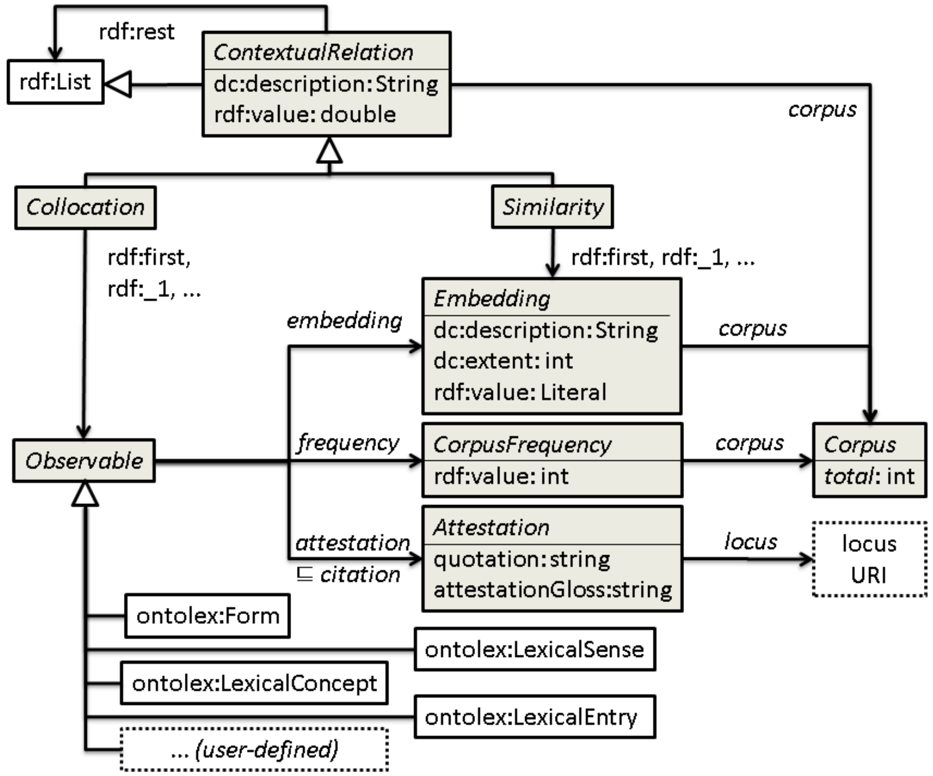
Minutes 2021-07-01

Venue: telco, ~~https://meet.google.com/rsx-mbkr-oxi~~ https://meet.google.com/qod-bkmh-nga
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC - Christian Chiarcos, GU Frankfurt
- MI - Max Ionov, GU Frankfurt
- FK- Fahad Khan, ILC-CNR
- RS - Rachele Sprugnoli, Unicatt
- FM - Francesco Mambrini, Unicatt
- TD - Thierry Declerck, DFKI
- GVO - Giedre Valunaite Oleskeviciene, MRU
- KG - Katerina Gkirtzou, ILSP-ARC

Agenda:

- Collocations / colligations
- Where do we stand and further plans



Collocations

- Sequential data for collocations: We can’t leave the core feature underspecified.

- CC: Preference for rdf:first, rdf:rest
  FK: Doesn’t it create a problem with OWL?
  CC: Yes, but we need to check if this is still the case for OWL2
  TD: Is it a problem for Frac? Do we want to have reasoning over OntoLex?
  CC: Highly unlikely for it, but people might want to do it on the same graph which might lead to unexpected behaviour.

  TODO: check if this is still the case. Reasoner used to break on collections like this. Test if it breaks on some collocation data + some owl in the graph

  TODO: change Collocation→Observable arrow from rdf:first to an aggregation symbol, put in the documentation: “rdf:List is a preferred way to model this, but we are aware of the potential discussions on rdf sequences, and if there is an official decision/recommendation, it should become a recommended way”

  FK: RDF sequence is recommended by W3C, a subclass of containers
  CC: This would make sense, this should be a preferred way. rdfs:Container
  CC: It has disadvantages when representing in Turtle and querying for order, but this is not too bad
  CC: Container is problematic because it is open-ended.
  TD: Can we specify size for the container?
  CC: We can introduce a new property for that
  MI: Are we talking about a relation between words in one collocation or a list of collocates for a word?
  CC: We can have both

  :kill_cf frac:hasContext :colloc1.
  :colloc1 a rdf:Seq, frac:Collocation; rdf:_1 :kill_cf; rdf:_2 :switch_cf; rdf:value “0.002”; dct:Description “portion of collocations per total of :kill_cf”.

  :switch_cf frac:hasContext :colloc2.
  :colloc2 a rdf:seq, frac:Collocation; rdf:_1 :kill_cf; rdf:_2 :switch_cf; rdf:value “0.0305”.

  “Absolute” collocation (value not specific to *one* word)
  :colloc3 a rdf:seq, frac:Collocation; rdf:_1 :kill_cf; rdf:_2 :switch_cf; rdf:value “50”; dc:description “total freq”.

  Multiple scores: recommended to introduce subproperties for rdf:value

For the next time:

1. Reconsider all the places that we might want to reconsider
   a. E.g. attestation locus

2. Think about when hasContext is necessary (for which metrics) and maybe whether to rename it

Minutes 2021-06-17

Venue: telco, https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC - Christian Chiarcos, GU Frankfurt
- TD Thierry Declerck, DFKI
- MI - Max Ionov, GU Frankfurt
- JBG - Julia Bosque-Gil, UNIZAR
- FK - Fahad Khan, CNR
- KG - Katerina Gkirtzou, ARC
- RS, Rachele Sprugnoli (UNICATT)
- SK - Saurav Karmakar, NUIG

1. Quick overview on recent updates (similarity, collocations, any updates elsewhere?)
    a. Similarity modelling approved?
        i. Modelling as sequential data structure approved, modelling as rdf:List to be discussed together with collocations
2. TD on representing signed language
    a. Sign writing systems: is it Unicode or images → to check
    b. First two layers in slides — corpus excerpts, other layers — annotations?
    c. Add subproperty to attestations: authoritative attestation
        i. GS> Not preferred, because there might be many authoritative, we don’t want to limit it to just one
        ii. GS> Maybe also the difference in attestations vs. elicitations in sign dictionaries is not that it’s authoritative but in context vs. not in context
        iii. CC> maybe we should have both: ±authoritative, ±in-context
        iv. FK> wouldn’t it be better to specify context as null for non-contextual attestations? Since we have context property anyway
        v. CC> we have only locus for now, it is a multistep process to figure out the context
    d. Summary:
        i. create a subproperty of attestation, authoritativeExample/representativeExample. Create a property of Attestation: observed (bool, in vivo/in vitro). Lexinfo property?
        ii. Other than that: beyond frac, whether sign structures can be presented in writtenRep (phoneticRep) or need another property (in the core module or a new module?)
3. Collocations
    a. Updates on Zaida’s data
        i. This is 100% in line with the GitHub draft

        ii. CC: some comments below
    b. How do we want to query for a collocation?
        i. Modelling with rdf:List means that collocations are blank nodes
            1. “Two” collocations with the same elements will be different objects (even if having the same order of elements)
            2. Retrieve all collocations that contain :x1 and :x2; order-insensitive:
                a. ?coll rdf:rest*/rdf:first :x1, :x2.
                b. But this includes also those that contain yet another element
            3. Retrieve all collocations that contain exactly :x1 and :x2, order-insensitive:

```
?conll rdf:rest*/rdf:first :x1, :x2.
MINUS {
?conll rdf:rest*/rdf:first ?x3.
FILTER( ?x3 not in (:x1,:x2) )
```

            This is QUITE some query, and SPARQL 1.1 only. Any way for a more easy equality check?
        ii. Also, we have no canonical order defined

4. aob

Minutes 2021-05-20

Venue: telco, https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- MI -- Max Ionov, GU Frankfurt
- JBG -- Julia Bosque-Gil, UNIZAR
- FK -- Fahad Khan
- KG -- Katerina Gkirtzou, ARC

Agenda

- Collocations

Minutes

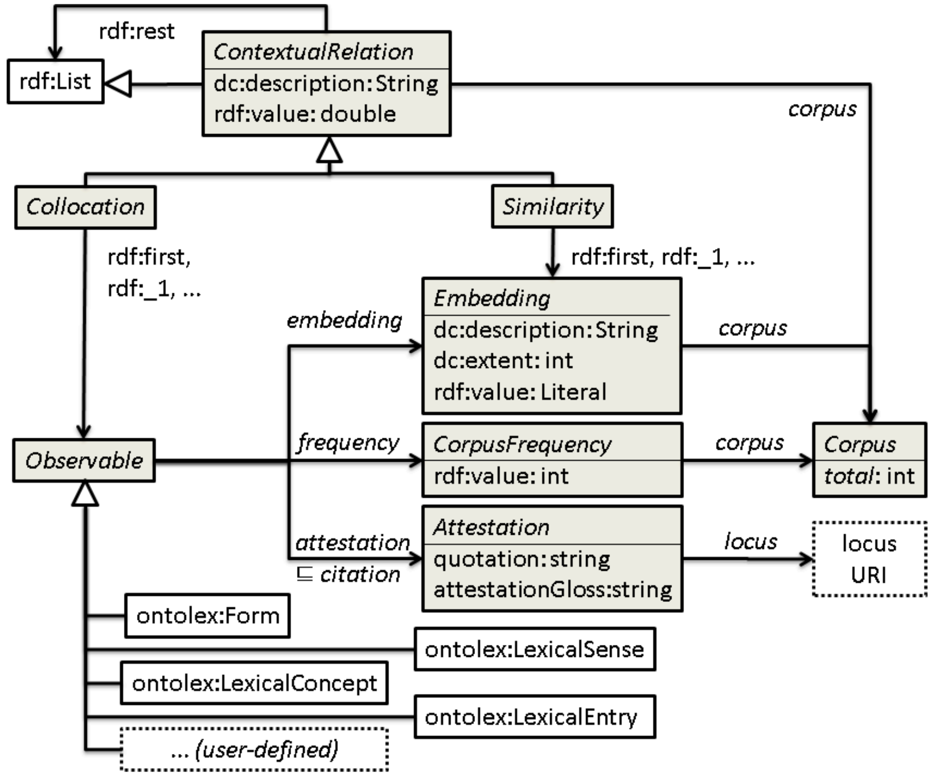
<presentation by ZBD?>

Representation needs mentioned by Zaida:

- Frequency
- Collocations of the terms
- Order

- Attestations in the corpus
- Colligations?
-

Current proposal in Frac:



- https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#collocations
- using rdf:List
- "Collocations are lists of ontolex:Elements, and formalized as rdf:List. Collocation elements can thus be directly accessed by rdf:first, rdf:_1, rdf:_2, etc. The property rdf:rest returns a rdf:List of ontolex:Elements, but not a frac:Collocation.
  By default, frac:Collocation is insensitive to word order. If a collocation is word order sensitive, it should be characterized by an appropriate dct:description, as well as by having lexinfo:termType lexinfo:idiom".

```
# kill (verb)
:kill_v a ontolex:LexicalEntry;
  lexinfo:partOfSpeech lexinfo:verb;
  ontolex:canonicalForm :kill_cf.

# kill (canonical form)
```

```
:kill_cf ontolex:writtenRep "kill"@en.

# switch (noun)
:switch_n a ontolex:LexicalEntry;
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:canonicalForm :switch_cf.

# switch (canonical form)
:switch_cf ontolex:writtenRep "switch"@en.

# form-form bigrams
(:kill_cf :switch_cf) a frac:Collocation;
  rdf:value "199";
  dct:description "2-grams, English Version 20120701, word frequency";
  dct:source ;
  dct:temporal "2008"^^xsd:date;
  lexinfo:termType lexinfo:idiom.

(:kill_cf :switch_cf) a frac:Collocation;
  rdf:value "121";
  dct:description "2-grams, English Version 20120701, document
frequency";
  dct:source ;
  dct:temporal "2008"^^xsd:date;
  lexinfo:termType lexinfo:idiom.

# form-lexeme bigrams
(:kill_cf :switch_n) a frac:Collocation;
  rdf:value "187";
  dct:description "2-grams, English Version 20120701, word frequency";
  dct:source ;
  dct:temporal "2008"^^xsd:date;
  lexinfo:termType lexinfo:idiom.

(:kill_cf :switch_n) a frac:Collocation;
  rdf:value "115";
  dct:description "2-grams, English Version 20120701, document
frequency";
  dct:source ;
  dct:temporal "2008"^^xsd:date;
  lexinfo:termType lexinfo:idiom.`
```

Minutes 2021-05-06

Venue: telco, https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- MI -- Max Ionov, GU Frankfurt
- JBG -- Julia Bosque-Gil, UNIZAR
- TD -- Thierry Declerck, DFKI
- RS -- Ranka Stanković, UB Serbia
- FK -- Fahad Khan, CNR
- Rachele Sprugnoli, UniCatt Italy
- IK -- Ilan Kernerman, KD
- SK - Simon Krek, ELEXIS, Jozef Stefan Institute, Slovenia
- DG - Dagmar Groman, University of Vienna
- SK -- Saurav Karmakar, NUIG

Agenda

- Organisational and updates
- SketchEngine API modelling

Minutes

Organisational and updates

- JBG > Zaida Bartolomé-Díaz contacted me via Slack in Nexus with questions on how to represent different collocations for a given term in FrAC. She asks, for example,
    - how to represent which verbs occur most frequently with a given term.
    - she wonders about the use of frac:collocation to express lexicalised multiword expressions
    - how to express more specificities about the collocations, for example, that they are noun + noun, or noun + adjective collocations.
    - Action on JBG -> to invite Zaida to present her representation needs in next session?
    - She is using VocBench
    - Check the results of a working group at the Dagstuhl Datathon (was working on collocations)

SketchEngine API modelling

API endpoint *Thesaurus*: an automatically generated **list of synonyms or words belonging to the same category** (semantic field). The list is produced based on the context in which the words appear in the selected text corpus. Only nouns, adjectives, verbs and adverbs are supported in most corpora

Input:

- Corpus
- Main word
- Part of speech (optional)

Returns:

- Word (synonym)
    - DG > Maybe we can get non-lemmatized forms from the Sketch Engine API as well
- Frequency (absolute)
    - MI > It’s absolute frequency, this fact was checked by Max
- Similarity score computed based on word sketches (≈ collocations)
    - MI > https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf, section 4
- Word ID (mysterious, looks like an inner ID of the word but can’t be extracted / referenced)

Example:
https://app.sketchengine.eu/#thesaurus?corpname=preloaded%2Fbnc2_tt21&tab=advanced&lemma=risk&lpos=-n&showScores=1&showresults=1

Current draft:
https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md#similarity

MI> Do we want to stick to this modelling? Is this pair-wise modelling of similarity scalable? Do we want to take another approach, with clusters of similarity, and use containers?

Frequency and similarity pairs:

```
skell_bnc2_threat a ontolex:LexicalEntry .

skell_bnc2_3209 a ontolex:LexicalEntry ;
        frac:frequency [
            a frac:CorpusFrequency;
            rdf:value "7440"^^xsd:int;
            frac:corpus
<https://app.sketchengine.eu/#dashboard?corpname=preloaded/bnc2> .
] .

skell_bnc2_threat_3209 a frac:Similarity ;
      dc:description “danger” ;
      rdf:value “0.305”^^xsd:float ;
      rdfs:member skell_bnc2_threat ;
      rdfs:member skell_bnc2_3209 .

skell_bnc2_4796 a ontolex:LexicalEntry ;
        frac:frequency [
            a frac:CorpusFrequency;
            rdf:value "6243"^^xsd:int;
            frac:corpus
<https://app.sketchengine.eu/#dashboard?corpname=preloaded/bnc2> .
] .

skell_bnc2_threat_4796 a frac:Similarity ;
      dc:description threat ;
      rdf:value 0.238 ;
      rdfs:member skell_bnc2_threat ;
      rdfs:member skell_bnc2_4796 .
```

FK: Daga et al. Paper on Sequential Linked Data (can maybe help us with deciding how to represent ordered clusters): swj2737.pdf (semantic-web-journal.net)

Minutes 2021-04-08

Venue: telco, https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC -- Christian Chiarcos, GU Frankfurt
- MI -- Max Ionov, GU Frankfurt
- JBG -- Julia Bosque-Gil, UNIZAR
- FK -- Fahad Khan, CNR
- GVO - Giedre Valunaite Oleskeviciene, MRU, LT
- KG - Katerina Gkirtzou, ARC
- FM - Francesco Mambrini, UNICATT, Milan
- AL - Ana Luís, Coimbra, PT

Agenda

- Organisational and updates

Minutes

- Organisational and updates:.
    - JBG (I estimate 5-7 minutes to go over this list in the telco)-> Update on five points after co-chair meeting on the W3C Day Schedule:
        - Schedule. The schedule will be posted on the Wiki, 4 hours including coffee breaks.
        - FrAC Slot. Allocated 30min for FrAC module status update/presentation, **would any of you be available (e.g. @Max, Christian) to give this presentation on FrAC?** There are 30min allocated for *morph* as well (JBG to contact Bettina), 30min for Terminology (Patricia UPM), and 30min for Syntax and Semantics (Alexander Kirillovich, if he is available).
        - ontolex:Form. Allocated 20min for suggested updates (in general) to OntoLex [slot led by chairs]. This includes ontolex:Form, and any other pending update (past suggestions on *vartrans*, for example). When ontolex:Form is discussed in this slot, it makes sense to summarize the FrAC discussion and e-mail thread that motivated this suggestion. Jorge will contact Fahad for discussion on versioning and sustainability.
        - Other prospective modules. There is a slot for 15min discussion (in total) on potential modules and the **multimodality** one. This could include a brief status update on the multimodality module, along with some mentions to the interest in etymology and maybe phonology (if time

allows). **@Christian, would you be available to address briefly the multimodality topic?**

        - LD4LT. **@Christian**, in the last LD4LT telco I remember we mentioned that the boundaries between the Linguistic Annotations goals/scope and OntoLex would need to be clear for the audience. This came up as a question by someone not involved in OntoLex but interested in annotations (I *think*, but do not recall exactly). Do you think this is **a point you could address in the LD4LT workshop** that same day, if you consider it appropriate?
    - JBG> *morph* & *decomp* (**Not for this telco**, but as a **heads-up** to the involved people). JBG will contact Bettina to set-up a doodle for a telco addressed at explicitly discussing decomp vs. morph in detail with a couple of examples - other chairs interested in attending in order to decide on what to do with respect to the relation of these two modules more informed and in light of examples.

[minutes to order] SketchEngine API + FrAC

- We are discussing the results of a query to get the similarity scores of *risk* and other lemmas in the corpus
    - Max to check the source of the “freq” value, are these absolute frequencies?
    - CC> Are there similarity clusters?
    - CC> Are these properties symmetric?
    - CC> What exactly does the “ID” value represent in these results?
- Grammar relations->
    - CC > The relations themselves should be in the scope of synsem -> syntatic frames? But we can treat them as Observable elements
        - JM > Yes, not sure exactly whether these relations would be syntactic frames according to synsem
        - CC > These seem to be dependency relations, with less information than a syntactic frame...but we would need a way to represent these -> we would need a mapping of the SketchEngine definition that fits the current synsem definitions
        - Max > Need to check whether there is a list of such patterns in SketchEngine
    - Max > Try to draft a representation of these collocations for the next time -> pointer to the collocation vs. colligation discussion
        - Fahad > We didn’t have concrete examples
        - CC > Isn’t the combination of the specific word + the pattern given by SketchEngine in fact a colligation?
            - Fahad> word + association with a certain POS tag?
            - Max > _risk of_ would be then a colligation?
        - CC > We need a modelling for this kind of data
        - Max > Will try to come up with a draft modelling for this? (to confirm)
- Concordances:

  - CC > There are different options:
    - **A**. Model every response as an attestation
    - **B**. Model the query as an attestation
      - Max> But option (b) would be locus, no?
      - CC> No, you have a filter in the query
    - **C**. Delegate this to the modelling of annotations, e.g. with NIF in a way that the context is included
- Max> what would be the next steps?
  - E.g. to create a library with utilities that take queries and produce FrAC RDF output?
  - CC> We could pre-compile word sketches and store them along with dict data
  - CC > If we only have dependency data...we could have a UD data and write something on our own (for the moment)
  - Max> so, _if we are working with a corpus management API_, what exactly do we do? This would be a use case of dynamic FrAC RDF data generation...what would we use that for? (what would be the use case for this)
    - CC> Save space, increase the level of granularity, e.g. for the case of Wiktionary.
    - Max > How would the next service in line use it?
      - CC> We can talk about _services_ here in this context, they need to return RDF. We would be providing a service.
      - Max> Who would be responsible for the creation? Added value -> this could be a wrapper for an existing corpus infrastructure.
      - CC > Need to discuss this in terms of challenges and capacity. Potential contributors but this needs further discussion.
    - ALL: think of the added value of this for particular use cases
    -

Minutes 2021-03-25

Venue: telco, https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC -- Christian Chiarcos, GU Frankfurt
- MI -- Max Ionov, GU Frankfurt
- FM -- Francesco Mambrini
- TD -- Thierry Declerck, DFKI
- JBG -- Julia Bosque-Gil
- FK -- Fahad Khan
- KG -- Katerina Gkirtzou, ARC

- AL - Ana Luís
- SK - Saurav Karmakar, NUIG

Agenda

- Organizational & Updates
- Metashare ontology
- Embeddings?

Minutes

- Organizational & Updates
    - How to join mailing list
        - https://www.w3.org/community/wp-login.php?redirect_to=%2Fcommunity%2Fontolex%2Fjoin
        - If doesn't work, coordinate with chairs (here: Julia and John)
    - Update on OntoLex form discussions
        - JBG: email, no feedback, discussion at f2f on Sep 4th
            - 1st week of april telco of co-chairs to discuss f2f agenda
            - Email asking for limitations of form
                - JM: which?
                - CC: one lexical entry per lexical form?
                    - JM: not explicitly, should be permitted (but is not intended)
                - TD: some problems, expressing form-specific ambiguity, specific senses (but found solutions)
                    - JBG: We have lexicog:FormRestriction, not sure whether that helped, but just in case! ;)
                - FM: no major problems, but occasional questions. What if we want to say that a writtenrep is at a particular time, etc. (cf. Fahad's request/question)
    - Update on multimodality discussions
        - Possible that these extend beyond OntoLex, still in the process of defiing the scope (many contributors from Nexus/non-OntoLex background), call for use cases to narrow that down
        - JBG: Sent Call for Use Cases
            - **TODO** > TD + FK: To present lexicographic multimodal examples for next Multimodal Telco (to make sure there are designated
        - TD: Currently working on possible FrAC/multimodality paper
- KG: Metashare ontology
    - The metashare ontology repo: https://github.com/ld4lt/metashare
    - *segmentationLevel* property
        - Specifies the segmentation unit in terms of which the resource has been segmented or the level of segmentation a tool/service requires/outputs
    - *Corpus* class
        - Does not need to be accessible, can be something completely abstract
        - A structured collection of pieces of data (textual, audio, video, multimodal/multimedia, etc.) typically of considerable size and selected according to criteria external to the data (e.g., size, type of language, type of text producers or expected audience, etc.) to represent as comprehensively as possible the object of study

    - *Size* class
        - The size of the resource with regard to the SizeUnit measurement in form of a number
        - → SizeUnit: The unit of measurement used for determining and describing the size of a resource (part)
    - *Distribution* class
        - Def in metashare: any form with which a language resource is distributed; for software, this can refer to web services, executable or code files, etc.; for datasets, it can be a downloadable form in a specific format (e.g., spreadsheet, plain text, etc.) or an API with which it can be accessed
    - *Dataset distribution* class
        - Each language resource can have multiple dataset distributions, no inverse relation from distribution to language resource). CC > We would need this back link. We need a way to provide there is a total
        - Any form with which a dataset is distributed, such as a downloadable form in a specific format (e.g., spreadsheet, plain text, etc.) or an API with which it can be accessed
    - CC> **TO-DO**: Maybe Ask Penny to rename *Distribution* to *Edition*, not referring to the distribution of datasets only, but also something more abstract. We can keep *distribution* as a label. If this were not an option, this would be still an important factor to consider...we would need to reify *total* (JBG> CC please check my minutes, thank you!). KG to ask Penny about this.
    - CC> Possible discussion in LD4LT telcos
- Embeddings (postponed)
    - TD data? To come after Easter (a bit delayed, as I was on holiday)

- Next call: 2 weeks
    - possible topic: TD on relation between older multimodality models and FrAC (= MMS paper)

Minutes 2021-03-11

Venue: telco, https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC -- Christian Chiarcos, GU Frankfurt
- MI -- Max Ionov, GU Frankfurt
- FK -- Fahad Khan, ILC
- JBG - Julia Bosque-Gil, UZAR
- KG -- Katerina Gkirtzou, ARC
- GS -- Gilles Sérasset

Agenda

- Organizational
- [From the mailing list]
- Metashare ontology
- TD data?
- LiLa data: segmentation

- Fahad’s Faliscan example (from mailing list)

```
 :ekupetaris a ontolex:Form ;
    lexinfo:case lexinfo:nominativeCase ;lexinfo:gender lexinfo:masculine ; lexinfo:number
 lexinfo:singular ;
    frac:attestation :att_0, :att_1, :att_2, :att_3,..., :att_9 ; ontolex:writtenRep "ECVPETARIS"@xfa, "EQUPETARS"@xfa, ...
 "eppetaris"@xfa .
  att_0 a frac:Attestation ;
    frac:attestationGloss "Pa2 lines 2-3, Certainty: certain, Bibliography:
  Pellegrini-Prosdocimi 1967, pp. 328-331" ;
    frac:quotation "ekupetaris" .
  :att_5 a frac:Attestation ;
    frac:attestationGloss "Pa6, Certainty: certain, Bibliography:Pellegrini-Prosdocimi 1967,
  pp. 344-348" ;
    frac:quotation "EQUPETARS" .
```

(see discussion in the OntoLex mailing list)

Summary of questions we are addressing in the telco:

- What to do when we have different attested forms but all of them -> same morphological features
- What to do when we have two forms, with the same morphological features, but both the the phonetic and the orthography is different:
    - CC: but here you could turn to the language tags (IF they are sufficient)
- What to do when we have more than one orthographic variant of a form and we want to provide more information about that orthographic variant specifically (e.g. to say it is not widespread)

- What to do when we have an attested form which could have different morphological features but the form is indistinguishable in these cases ->
    - single out examples in which you cannot say for sure whether something is a nominative or an accusative form, and don’t want to decide.
    - We can also have a lexical resource providing a form with different morphological interpretations

      Also, in the Rayfield Georgian dictionary: information of which orthographic norm is more widely used → two forms. Changing the guidelines in OntoLex about one or different forms to indicate this depends much on the use case?

- CC: discuss this in the F2F meeting. JBG writes the email about all this to the mailing list

“Different forms are used to express different morphological forms of the entry."
->

"Different forms of the same entry are used to express forms that differ in their structure."

Minutes 2021-02-18

Venue: telco, https://meet.google.com/fnx-ctad-dcs [**NEW! ONLY THIS TIME**]
FOR MARCH 11 PLEASE GO HERE: https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC -- Christian Chiarcos, GU Frankfurt
- MI -- Max Ionov, GU Frankfurt
- FK -- Fahad Khan, ILC-CNR
- FM -- Francesco Mambrini, LiLa - Università Cattolica
- RS -- Rachele Sprugnoli, LiLa - Università Cattolica
- TD -- Thierry Declerck, DFKI
- KG -- Katerina Gkirtzou, Athena Research Center

Agenda

- Organizational
- LiLa
- Embeddings
- Similarity
- Other topics?

Minutes

- Organizational
    - GitHub updated with publications (SemDeep-2021), presentations (EUROLAN-2021) and bibliography (bib.bib): https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/doc
    - Wikipage updated: https://www.w3.org/community/ontolex/wiki/Frequency,_Attestation_and_Corpus_Information, checked that all documentations are listed and all documents are linked
        - declerck@dfki.de : What's SemDeep proceeding status?
        - Should go to ACL anthology
    - Anything we need to know/plan for the OntoLex meeting at LDK?
- LiLa
    - LiLa sample data [FM, MP] under https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/samples/LiLa

- **OLD TODO@all**: Until next time: look at this and comment either in minutes document or as github issues
    - CC: some comments in https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/samples/LiLa/sample-commented.ttl
        - There is no way to say what the segmentation for the total is
            - CC: reify total?
            - FK: reify segmentation or tokenization instead?
            - CC: introducing segm. Or tok. Would be this reification already
            - FK: this is the preferred way
            - FM: yes, this is good
            - MI: can we reuse some vocab for that?
            - CC: not really. NIF but with limitation (also, NIF:String/NIF:Token is not really suitable for linguistic objects as words)
            - CC: use dc:Dataset instead of frac:Corpus?
            - FM: let's **leave the question open**
        - How to better approach modelling citations, maybe not CITO?
            - CC: CTS/DTS might be good for that, there are discussions about this in LD4LT telcos
            - FK: locus must be specified for each attestation, right?
        - Embeddings, rdf:value doesn't have a way to tell a user how to parse it. If we use CSV2RDF we don't need to specify the embeddings in RDF in the first place
            - CC: we need to ask someone with more experience in CSV2RDF
        - If we won't reuse anything for describing segmentation, how will we model this?
            - There was a lengthy discussion: many options
                - Don't use Corpus, use dc:Dataset, use Corpus as a certain segmentation for a dataset, use Dataset as a manifestation of a corpus, etc.
            - CC: I want to stay as agnostic as possible in the definition of a Corpus, OntoLex only on *lexical resources*, this would be beyond scope
            - KG: reuse METASHARE vocabulary?
                - **TODO@KG**:
                  next time — to present a part of the Metashare ontology (segmentation class) to see if it could be used within the frac:Corpus class
                  The repo of Metashare ontology
                  https://github.com/ld4lt/metashare

- As a result: TODO@FM — update the example, add frac:locus to attestations

- Updates on embeddings
  - Gender-Number Ambiguities in Wiktionary with BERT [TD] **[=> postponed to next call]**
    - Status of encoding?
    - **OLD TODO@TD**: please share SemDeep slides and recordings
  - Alternatives to rdf:value
    - SemDeep question: if we have a lot of embeddings, do we want to store them all in RDF or in a database? How to make that scale?
      - **OLD TODO@TD**: engage with colleague from the netherlands who asked that question at SemDeep, put CC and MI in CC
    - Consensus so far: focus on exchange format, not interactive access to it => binary encoding of the full RDF graph can be implemented independently from modelling decisions [=> CC&MI working on this]
    - Open question: → can we point into a database?
      - CC: if stored as a textfile, via line URIs according to https://tools.ietf.org/html/rfc5147 (cf. RFC 5147 *string* URIs used in NIF): **but** it normally is not provided as textfile, but in a compressed (gzip) or non-portable binary format (pickle!)
      - **TODO@all**: ideas?
  - Cosine similarity between words in contexts [requested by RS] [=> postponed]
    - RS: frequency with these similarities would be useful; threshold on frequencies to compute similarities only on more that N times occurring
    - CC (offline): Attestation -instanceEmbedding-> Embedding <-rdfs:member- Similarity
- Updates on similarity
  - Python implementation of SketchEngine API [RS+MI] [=> postponed]
    - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/samples/sketch-engine
    - **OLD TODO@RS, CC, MI**: wrap it into FrAC RDF
  - distributional similarity with word2vec [CC]
    - postponed until we have a use cases with a concrete application or API
    - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/samples/similarity/word2vec (just w2v docu)
  - Word similarity in psychological association tests (used as evaluation data for distributional similarity) [CC]
    - Postponed until representative data can be found
    - preliminary information under https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/samples/similarity/word-associations
- Next meeting
  - Next call in three weeks to avoid clash with morphology?

    - Feature a specific use case then?
        - TD data
        - KG presentation
        - Discuss LiLa data: segmentation
        - Python implementation of SketchEngine API? [RS (+MI)]
    - We use the “old” telco link again, then

Minutes 2021-01-28

Venue: telco, https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC -- Christian Chiarcos, GU Frankfurt
- JB -- Julia Bosque Gil, Uni Zaragoza
- MI -- Max Ionov, GU Frankfurt
- TD -- Thierry Declerck, DFKI
- FK -- Fahad Khan, CNR
- RS -- Ranka Stanković, UB
- F
- JM - John McCrae, NUIG

Minutes

- Similarity (additional samples)
    - RS: Python implementation of SketchEngine API
        - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/samples/sketch-engine
        - **TODO@RS, CC, MI**: wrap it into FrAC RDF
    - CC: distributional similarity with word2vec
        - https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/samples/similarity/word2vec (just w2v docu)
        - Better would be a tool that wraps this into an end user application or via an API, postponed until we have that
    - CC: Word similarity in psychological association tests [used as evaluation data for distributional similarity]
        - preliminary information under https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/samples/similarity/word-associations
        - data seems incomplete — check if better data can be found, postpone usecase until then
- FM: LiLa sample data
    - We’ll work on this use-case now; progress should be done before March

  - FM: no major difficulties in developing sample data
  - Sample data in the Frac GH under https://github.com/ontolex/frequency-attestation-corpus-information/tree/master/samples/LiLa
    - **TODO@all**: Until next time: look at this and comment either in minutes document or as github issues
- TD: Gender-Number Ambiguities in Wiktionary, showing BERT results, sample data
  - Dataset: Words in context: try to encode it in FrAC?
  - **TODO@TD**: please share SemDeep slides and recordings
  - SemDeep question: if we have a lot of embeddings, do we want to store them all in RDF or in a database? How to make that scale?
    - CC: it's an important question. One possibility is to use the current model + a compact binary encoding. We (GUF) already experiment with such an implementation. This will facilitate exchange of data only, not interactive access to it
      - realistically, focus on RDF as an exchange format for the moment
    - Open question: → explore how can we point into a database?
      - CC: if stored as a textfile, via line URIs according to https://tools.ietf.org/html/rfc5147 (cf. RFC 5147 *string* URIs used in NIF)
      - CC: but it normally is not provided as textfile, but in a compressed (gzip) or binary format (pickle!)
      - **TODO@all**: ideas?
    - **TODO@TD**: engage with colleague from the netherlands who asked that question at SemDeep, put CC and MI in CC
  - Dataset part: Cosine similarities between words in contexts
    - RS: frequency with these similarities would be useful; threshold on frequencies to compute similarities only on more that N times occurring
- Next meeting
  - Discuss LiLa modelling decisions and todos above
  - Next call in 3 weeks due to overlap with EUROLAN "summer" school

Minutes 2021-01-07

Venue: telco, https://meet.google.com/rsx-mbkr-oxi
Time: 12:00-13:00 CET
Participants (please list yourself with initials; optional: affiliation)

- CC -- Christian Chiarcos, GU Frankfurt
- JB -- Julia Bosque Gil, Uni Zaragoza
- MI -- Max Ionov, GU Frankfurt
- RS -- Ranka Stanković, University of Belgrade
- FK -- Fahad Khan, ILC-CNR

- SK -- Saurav Karmakar, NUIG

Minutes

- Updates on similarity
    - SketchEngine
- collocation/colligation?
    - Fahad: start with collocation, possibly extend to collligation, not actively seek examples
        - No objections, approved
    - Ranka: Multi-word units linked with
        - Excavator -- XY excavator
            - Lexical information more in ontolex core, incl. paradigm
            - Frequency / corpus-based information within FrAC
            - MI: criterion to model sth with FrAC should be whether it emerges from a corpus or a metric or algorithm that constructs the observable. E.g.: MWE can be modeled with other modules, but if they have scores (e.g. T-score) or a source corpus → this is a good fit to model it with Frac
                - No objections, approved
- Use Cases?
    - SketchEngine
        - See comments about their API in comments below
        - TODO: Ranka: deposit Markdown/Jupyter NB in Github
- SemDeep presentation
    - ?how to record/attend discussion
- OntoLex f2f?
    - June 17th, post-conf
    - No fixed program, yet. Proposal: https://docs.google.com/document/d/1vAJx-WQNN5h__f6-xLT3X-vmqGwt7TCJ_yWboElFaRs/edit
    - Can include frac status update
    - Topics beyond frac:
        - Morphology module (to be resumed?)
        - Expressions of interest for other potential modules (terminology, diachrony, multimodality)
        - Discussion points: function words? Language tags?
    - Update morphology:
        - Bettina Klimek preparing Doodle poll on morph and review of what’s ready and not
- Update on SWJ plans
    - No FrAC paper, but FrAC in a Nexus T1.1 paper
    - Jan 25th, 2021

- Next Telco Jan 28th, then biweekly
    - Topics:
        - similarity (sketch engine example, then in GitHub)
        - LiLa sample data
        - (regular topics)