Harmonizing Linguistic Annotations (An LD4LT workshop)

Welcome & Overview



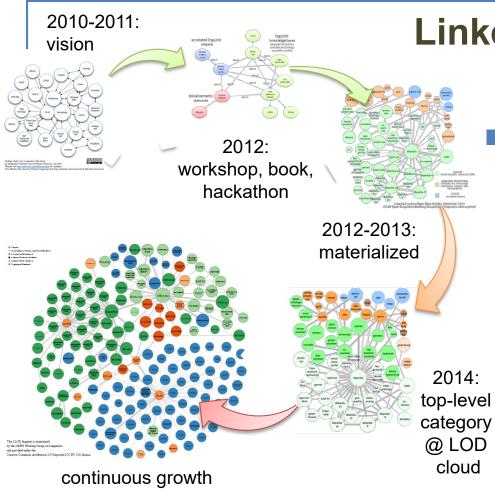




Linked Data for Language Technology (LD4LT)

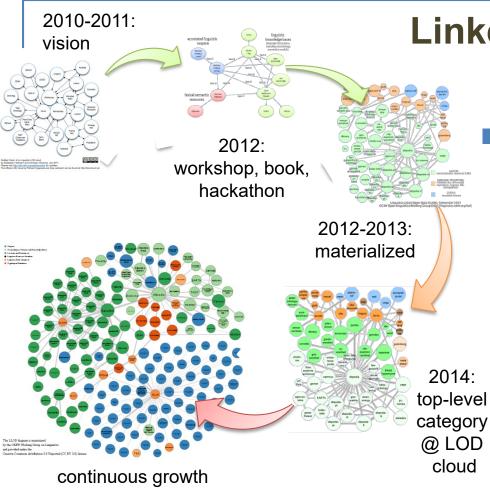
https://www.w3.org/community/ld4lt/

- W3C Community Group
 - formed 2013
 - address use cases and requirements for Language Technology
 Applications that use Linked Data
 - ⇒ interoperability
 - language technology
 - surveying needs and requirements (esp. 2013-2016)
 - language resource metadata (since 2015)
 - linguistic annotation (intensified since 2019)



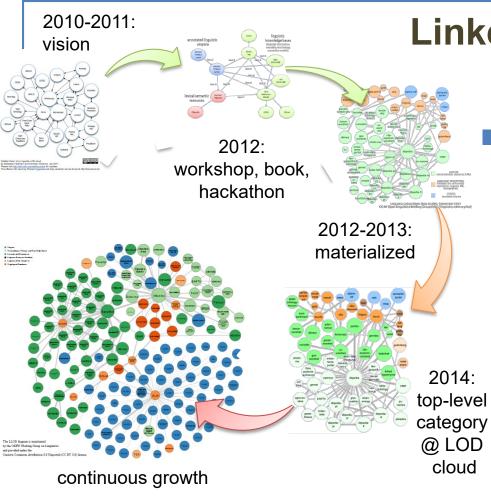
Linked Data and Language Technology

- since 2010, an LOD sub-cloud of language resources has been emerging
 - Open LinguisticsWorking Group
 - http://linguistic-lod.org



Linked Data and Language Technology

- since 2016, it has particularly grown with respect to lexical resources
 - W3C CG Ontology-Lexica (OntoLex)
 - one vocabulary, many use cases



Linked Data and Language Technology

- linguistic annotation remains a problematic area
 - some resources (blue)
 - few successful user stories
 - concurrent standards
 - USP: interoperability?
 - not in the current situation

Different communities
 of practice annotate
 very different things

Interlinear Glossed Text (Toolbox)

https://software.sil.org/toolbox/

```
017_AxiskaAndasbatir280914NIA_A01.txt
                                                                                                                               _ D X
                                                                                                                                                    Dictionary.txt:4
                                                                                                                                                                     - - X
              017:026
                       sølmum
                                                                                       gendi dilimizi
                                                                                                                   zořerym
                                                                        joġ jani
                                                                                       gändi dilimizi
                                                                                                                   zölärïm
                                                                                                                                     bura
                                                                                                                                                   dyl
                                                                   -DA jox jani
                                                                                                                                                   dil:er
                                                                   -loc not in other words self
                                         there -gen language -poss.3sg
                                                                                                              -akk sav -aorist
                                                          -pron(poss) -case neg ptcl
                                                                                                                                                   dil -lAr
              pron(pers) v -tense -pers.end pron -case n
                                                                                                     -pron(poss) -case v -tense -pers.end pron
                                                                                                                                                   dil:3r
              I understand their language, but we are speaking our mother tongue.
                                                                                                                                                   dil -lAr
        \nt
                                                                                                                                             \ps
                                                                                                                                                   17/Jan/2017
        <
\id 017
\ref 017:026
\per
\tx ben
Vitr hän
                                                                                                          qändi dilimizi
                                          oranïn
                                                                                                                                              zölärïm
\mb bän
                                         ora -(n)In dil
                                                                 -(s)I(n)
                                                                                                           gändi dil
                                                                                                                             -(I)mIz
                                                                                                                                          -(j)I söjlä -(A/I)r -(I)m
                                                                                           jani
                                            there -gen language -poss.3sg -loc not
                                                                                              in.other.words self
                                                                                                                        language -poss.lpl
\ps pron(pers) v
                                                                    -pron(poss) -case neg
                                                                                                               ptcl
\ft I understand their language, but we are speaking our mother tongue.
```

Different communities
 of practice annotate
 very different things

Exmaralda

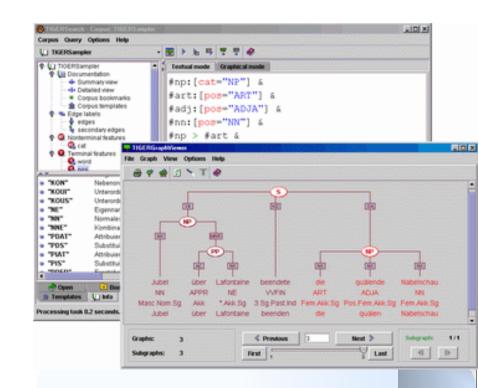
layered annotation over multimodal content (annotating dialogue, guestures, etc.)

http://exmaralda.org



Different communities
 of practice annotate
 very different things

Phrase
Structure
Syntax
(TIGERSearch/Annotate)

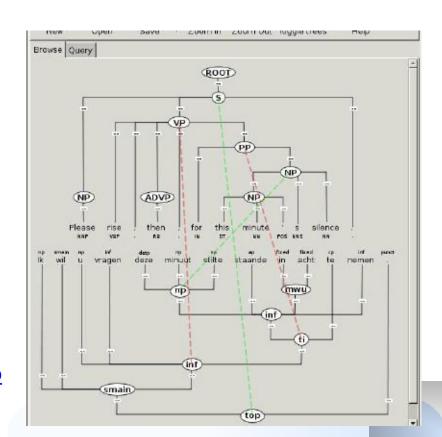


Different communities
 of practice annotate
 very different things

TreeAligner

syntax annotation for two aligned texts

https://www.ling.su.se/english/nlp/tools/stockholm-treealigner



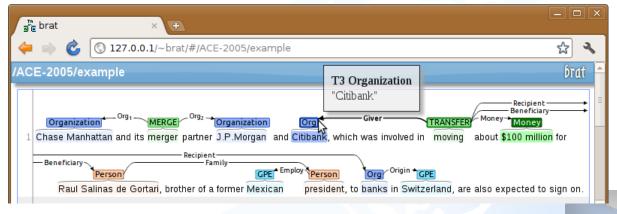
Different communities
 of practice annotate
 very different things

Recogito
Georeference
(Entities)

```
Όροι ἡπείρων·
Εὐρώπης μὲν καὶ Λιβύης αἴ Ἡρακλέους στῆλαι, Λιβύης δὲ καὶ Ἀσίας
ό Νείλος· οἱ δ' ἰσθμὸν τὸν ἀπὸ Σερβωνίδος λίμνης καὶ Άραβίου
<mark>κόλπου, Άσίας</mark> δὲ καὶ <mark>Εὐρώπης</mark> οἱ μὲν ἀρχαῖοι Φᾶσιν ποταμὸν καὶ <mark>τὸν</mark>
<u>ἔως Κασπίας ἰσθμόν, οἱ δ' ὕστερον νεώτεροι Μαιῶτιν λίμνην καὶ</u>
Τάναϊν ποταμόν.
Έκλήθησαν δ' ἤπειροι ἄπειροί τινες οὖσαι δι'
άγνοιαν Ασία δ' ἀπὸ τοῦ ἆσσον ἰέναι τοῖς ἀπ' Εὐρώπης ἀπιοῦσι καὶ
πεζῆ καὶ νήσοις στιχηδὸν κειμέναις. Εύβοια Άνδρος Τῆνος Μύκονος
Ικαρία Σάμος Μυκάλη. ἡ δὲ <mark>Εὐρώπη</mark> ἀπὸ τοῦ εὔρους ἀνομάσθη,
Λιβύη δ' ὑφ' Ἑλλήνων ἦν ἄγνωστος πάνυ, ἀπὸ δὲ ἔθνους ἐπισήμου
Φοίνικες ἀνομάσθησαν πλέοντες, <mark>ἀκεανὸς</mark> δὲ διὰ τὸ ἀκέως ἀνύειν
κύκλω την γῆν.
```

Different communities
 of practice annotate
 very different things

Relations and Entities (Brat)

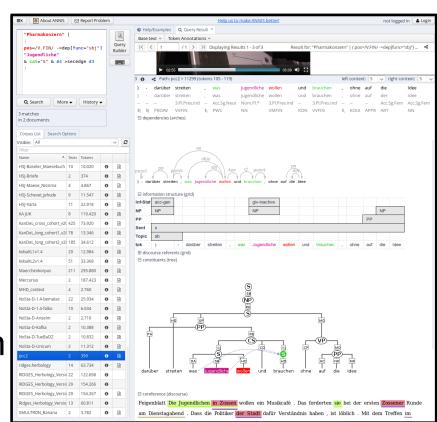


https://brat.nlplab.org/manual.html

Different communities
 of practice annotate
 very different things

All at once ?
Multi-Layer Annotation

https://corpus-tools.org/annis/



09:00 Welcome

09:10 - 10:30 Background: Linguistic Annotation on the Web

W3C Standard Web Annotation Christian Chiarcos

NLP Interchange Format Milan Dojchinovski

Text Encoding Initiative Fahad Khan

ISO TC37 standards
 Thierry Declerck

Text Fragids Joel Kalvesmaki

10:30 - 11:00 Break

11:00 - 11:40 Discussion

- QA: What is missing? What is unclear? Where are problems?
- Summary of LD4LT Discussions on Linguistic Annotation

11:40 - 12:30 Use Cases, Experiences, Extensions

Linking Latin
Francesco Mambrini

Distributed Text Services Christian Chiarcos

Interlinear Glossed Text
Maxim Ionov

Discourse Research
 Giedre Valunaite Oleskevicienė

Transforming Language Resources Christian Fäth

12:30-13:00 Brainstorming & Next Steps



- primary goals
 - provide and collect background information
 - present and discuss use cases and requirement
 - initiate and plan future discussions
- we record background presentations
 - for future reference
 - if you don't want to be recorded, please switch off camera and join the discussion via chat

- held in conjunction with
 - 3rd Conference on Language, Data and Knowledge (LDK-2021)
 - Face-to-face meeting of the W3C CG Ontology-Lexica
- organized in cooperation between
 - W3C CG LD4LT
 - Cost Action Nexus Linguarum, T1.1

09:00 Welcome

09:10 - 10:30 Background: Linguistic Annotation on the Web

W3C Standard Web Annotation Christian Chiarcos

NLP Interchange Format Milan Dojchinovski

Text Encoding Initiative Fahad Khan

ISO TC37 standards
 Thierry Declerck

Text Fragids
Joel Kalvesmaki

10:30 - 11:00 Break