

FoLiA: Format for Linguistic Annotation

FoLiA: Format for Linguistic Annotation

Current state of affairs



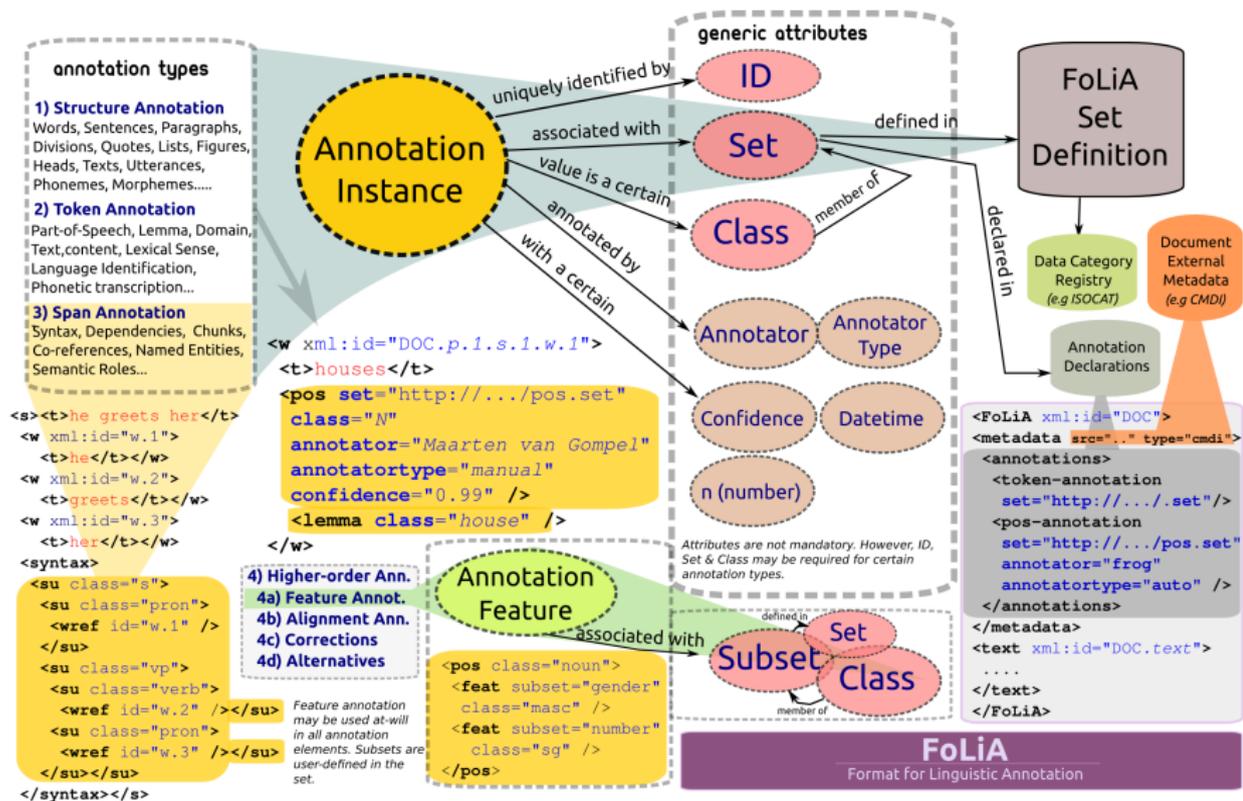
Intended Applications

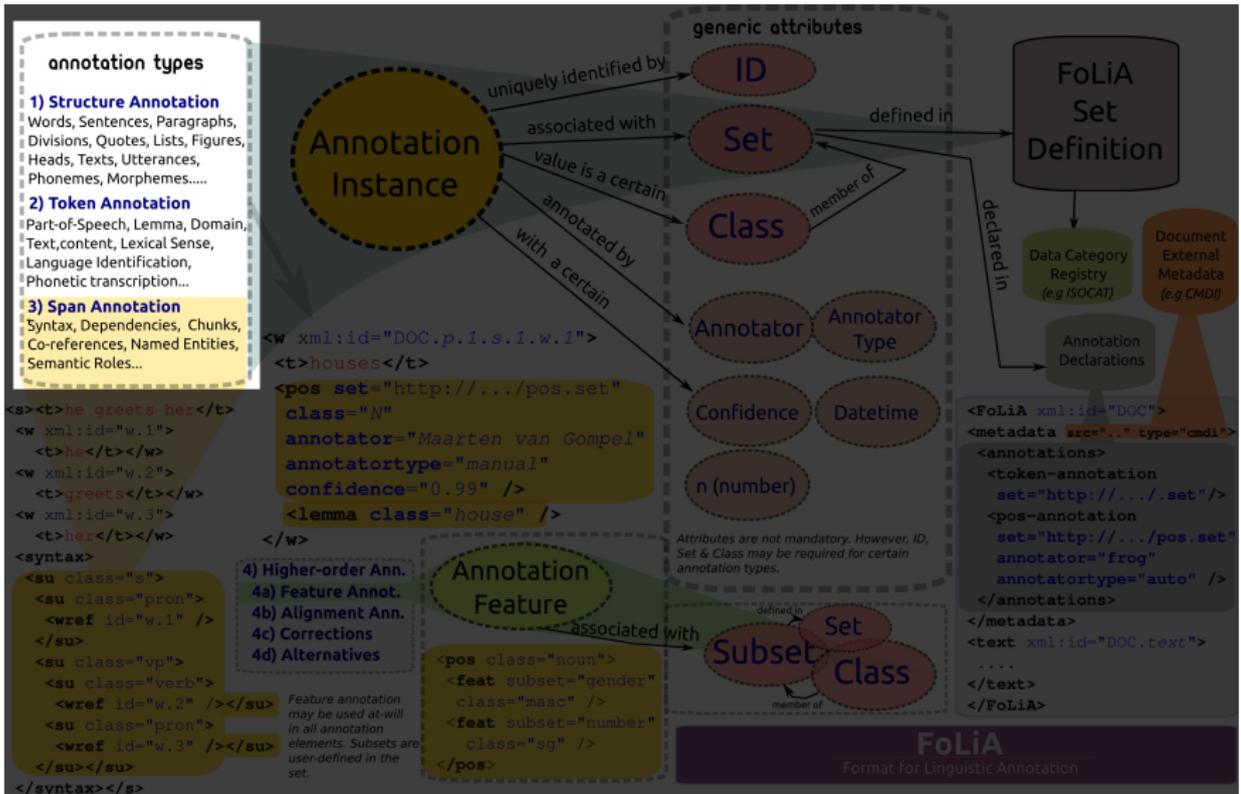
- as a corpus **storage** format
- as a language resource **exchange** format

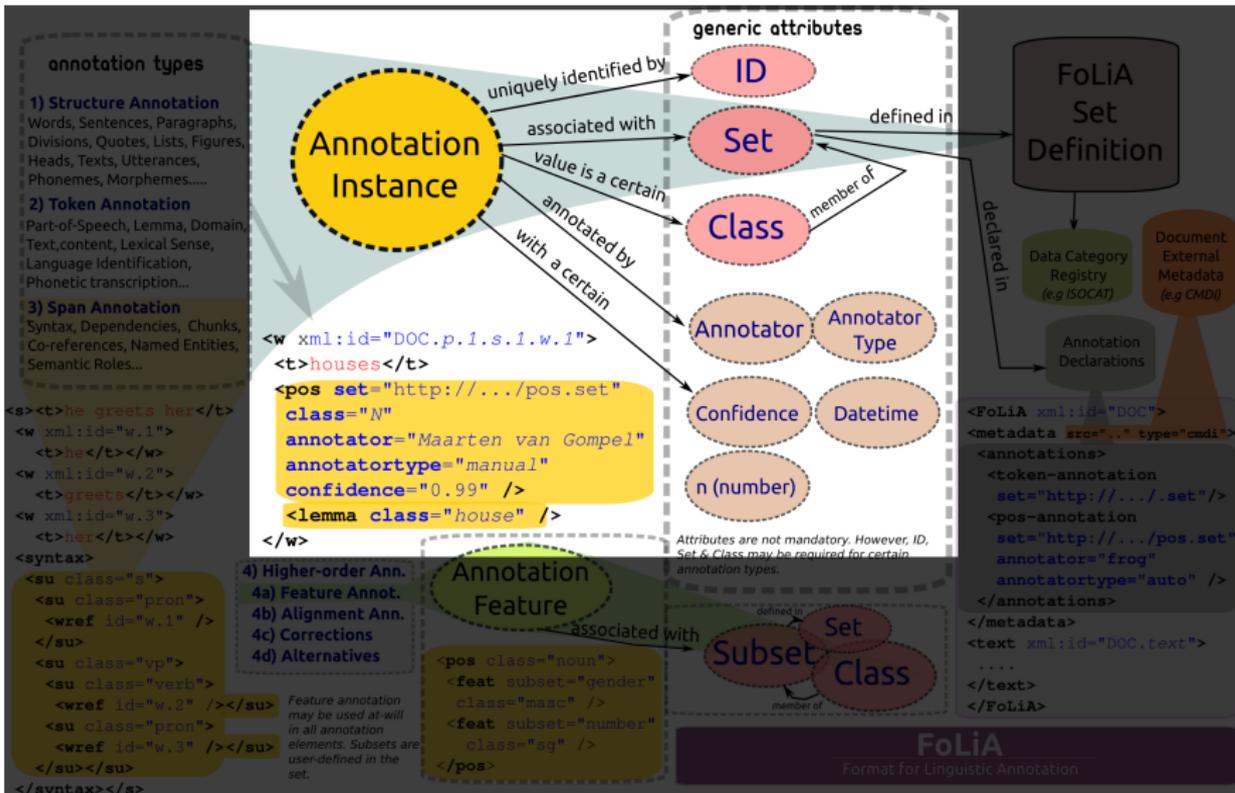
FoLiA Characteristics

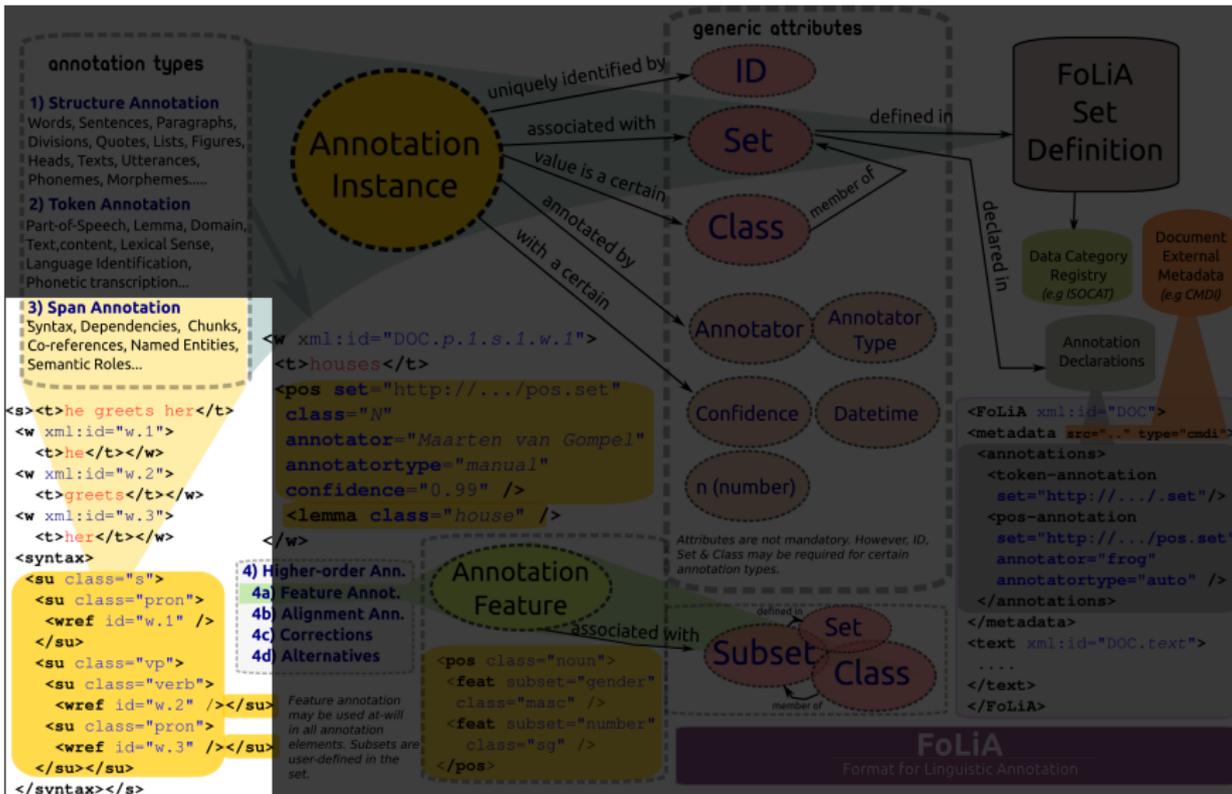
- **Uniform** setup – Consistent and uniform paradigm unifying various kinds of annotation. Not committed to any label set.
- **Extensible & Flexible** – Easy to extend
- **Expressive** – Verbose expression of annotations, their annotators, timestamps, etc... Moreover, support for *alternative* annotations.
- **Formalised** – Validation on shallow (structural) or deep level. The latter validates the label set and allows for links with for data category registries (e.g. ISOcat).
- **Practical** – Bottom-up development alongside libraries, various applications, different projects.

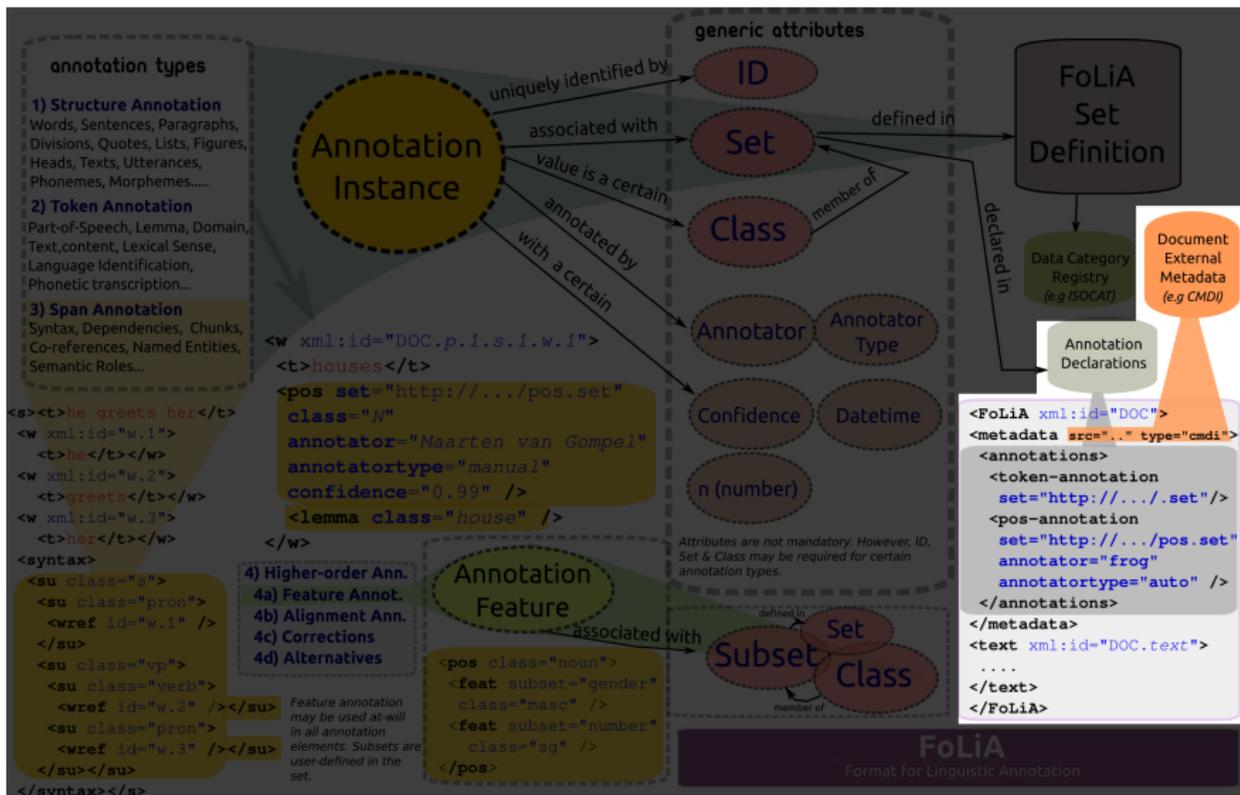
Paradigm

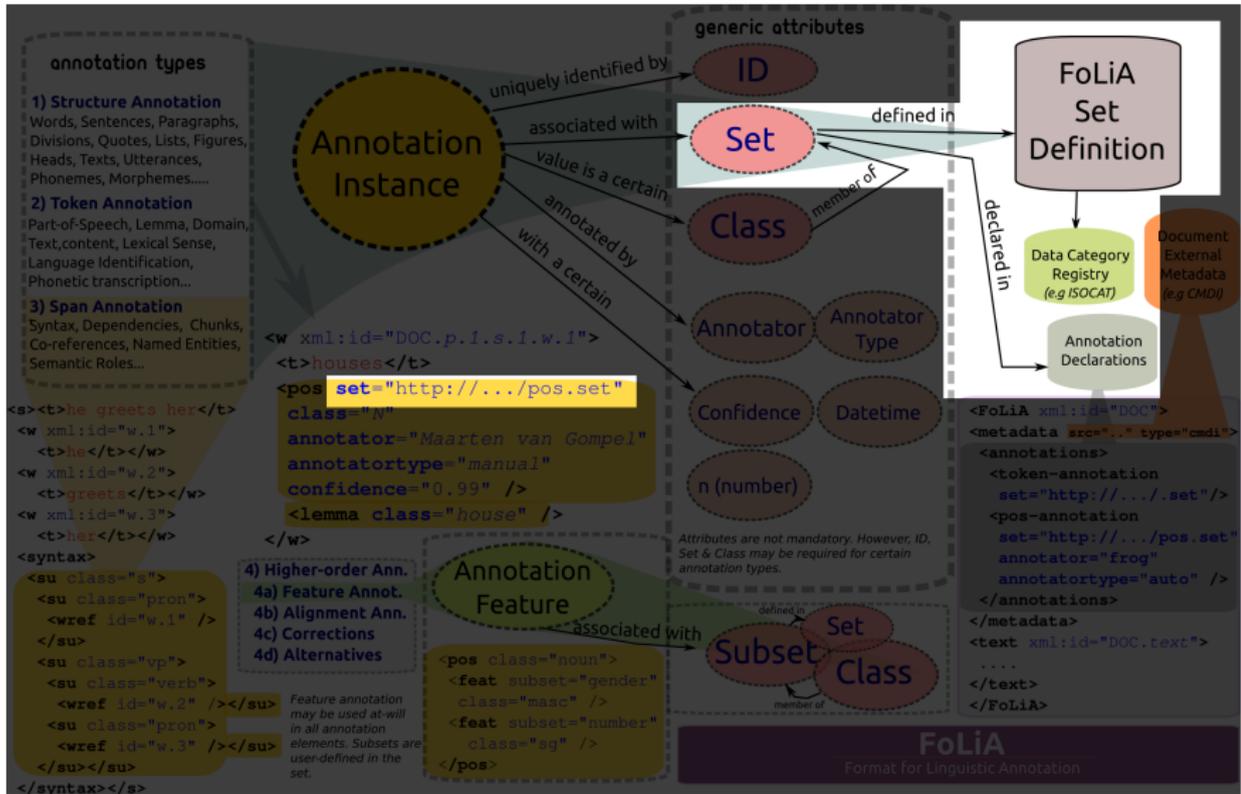


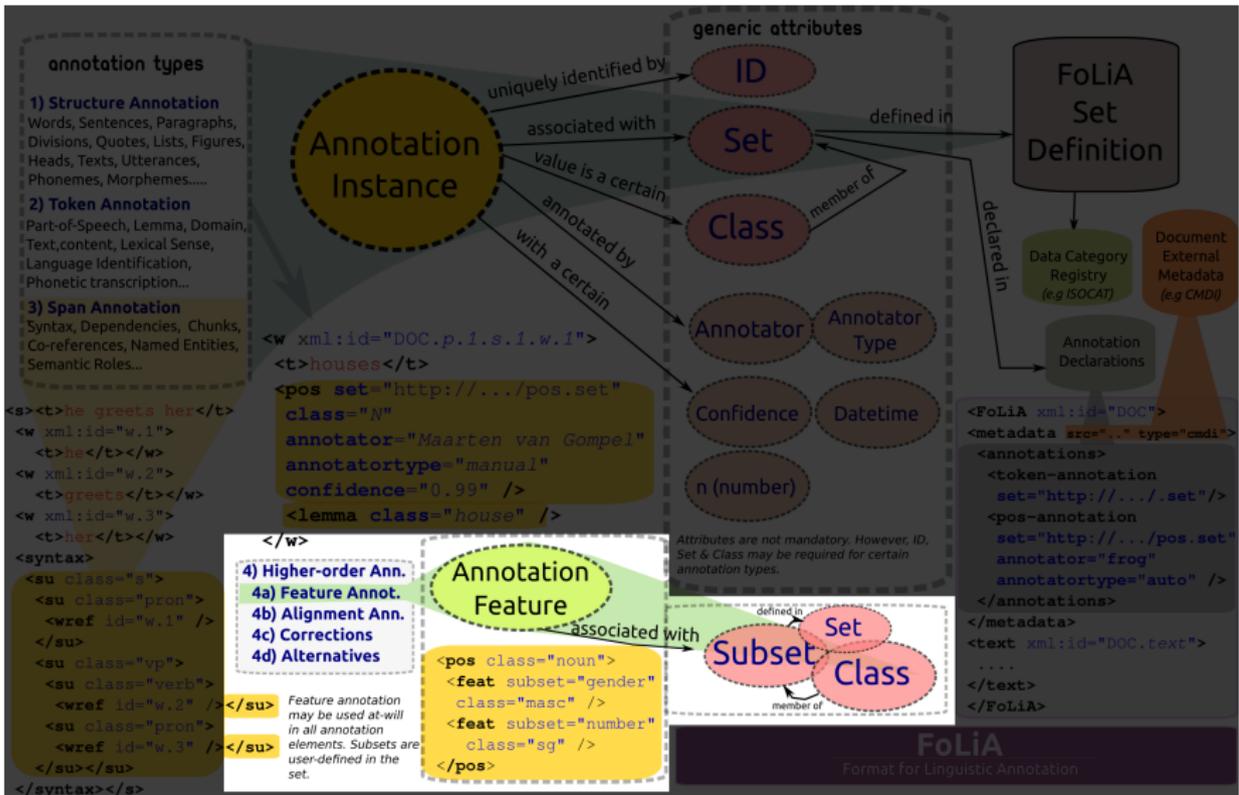




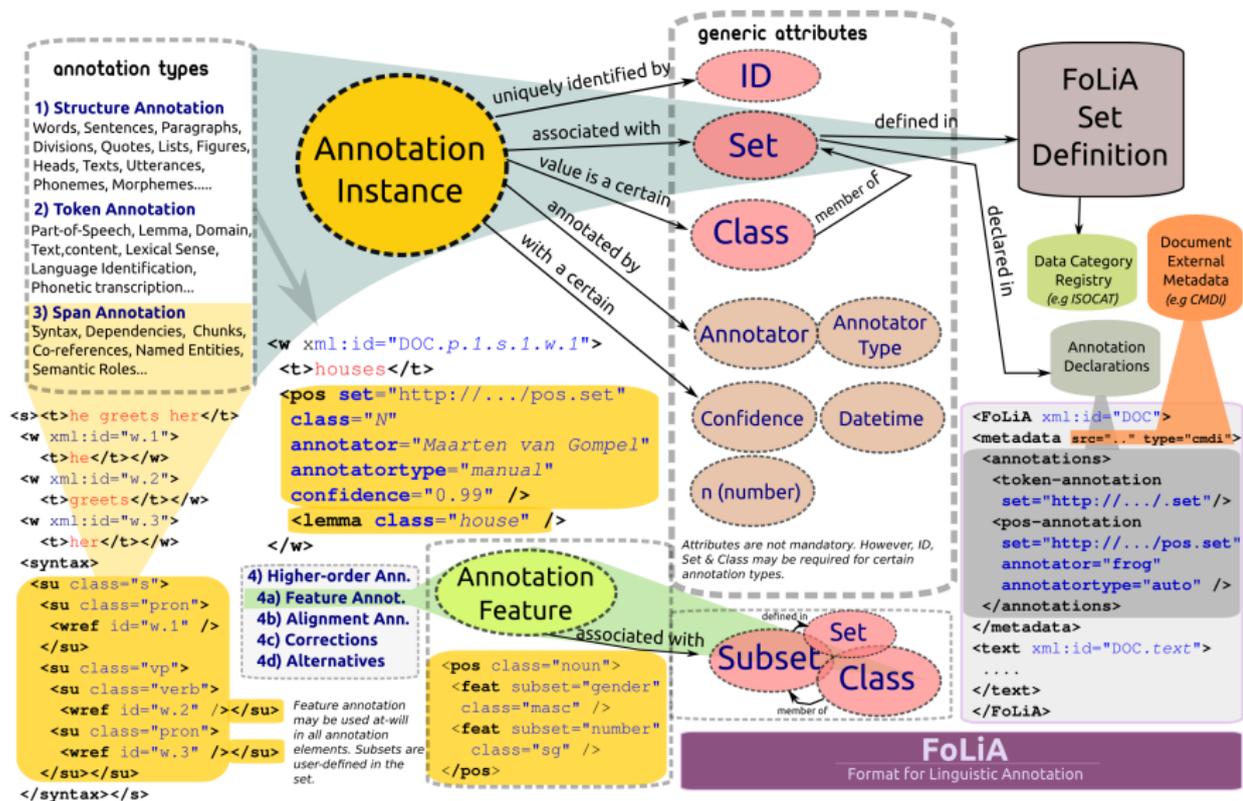








Paradigm



Tools for working with FoLiA

- Standard **XML** facilities: XSLT, XPath
- **Python** library: `pynlpl.formats.folia`
- **C++** library: `libfolia` (*Ko van der Sloot*)

FoliaTools

Installation: `$ easy_install folia`

- **Converters:** `folia2dcoi`, `dcoi2folia`, `folia2html`, `folia2columns`, `alpino2folia`
- **Validator:** `foliavalidator`
- **Simple search tool:** `foliaquery`
- **Other:** `foliamerge`, `foliafreqlist`

Applications using FoLiA

- **Frog** – tagger/lemmatisation/parser suite: FoLiA input & output
- **ucto** – tokeniser: FoLiA input and output
- **Valkuil.net** – Dutch spelling checker
- **Fowlt.net** – English spelling checker
- **Ticcl** – Spelling normalisation

Corpora delivered in FoLiA

- SoNaR (*STEVIN*), DutchSemCor (*NWO*), VU-DNC (*CLARIN*), Basilex (*NWO*)

```
from pynlpl.formats import folia

#Load FoLiA document
doc = folia.Document(file='/path/to/fofia_doc.xml')

#grab a specific sentence from the index
sentence = doc['fofia_doc.s.1']

#print words in sentence along with PoS and lemma
for word in sentence.words():
    print word.text() + ", " + word.pos() + ", " +
          word.lemma()

#add PoS Annotation to a specific word
word = doc['fofia_doc.s.1.w.5']
word.append( folia.PosAnnotation, cls="N",
             set="http://some/url/cgn.set")

doc.save() #Save edited document
```

Working with FoLiA: visualisation

The screenshot shows a web browser window with the URL `ilk.uvt.nl/folia/example.xml`. The page content is as follows:

Stemma

Stemma is een ander woord voor stamboom. In de historische wetenschap wordt zo'n **stamboom**, onder de naam **stemma codicum** (handschriftelijke genealogie), gebruikt om de verwantschap tussen handschrift en **weerde geven**.

Werkwijze

Hiervoor worden de handschriften genummerd en gedateerd zodat ze op de juiste plaats van hun afslammingsgeschiedenis geplaatst kunnen worden. De hoofdletter A wordt gebruikt voor het originele handschrift. De andere handschriften krijgen ook een letter die verband kan houden met hun plaats van oorsprong of plaats van bewaring. Verdwenen handschriften waarvan men toch vermoedt dat ze ooit bestaan hebben worden ook in het stemma opgenomen en worden weergegeven door de laatste letters van het alfabet en worden tussen vierkante haken geplaatst. Tenslotte gaat men de verwantschap tussen de handschriften aanduiden. Een volle lijn duidt op een verwantschap, terwiel een stippelijen op een onzekere verwantschap duidt.

- Eerste testitem
- Tweede testitem

The family tree diagram shows the following structure:

```

graph LR
    LG[Lucas Grey] --- MG[Mary Grey]
    LG --- JG[Jason Grey]
    MG --- FS[Fred Smith]
    MG --- JS[Jane Smith]
    MG --- SG[Sean Grey]
    JG --- JG2[Jessica Grey]
    JG2 --- JW1[Joseph Wetter]
    JG2 --- JW2[John Wetter]
    JG2 --- LW[Laura Wetter]
  
```

Names in red boxes (Mary Grey, Jane Smith, Jessica Grey) represent the first test item, and names in blue boxes (Fred Smith, Sean Grey, Jason Grey, Joseph Wetter, John Wetter, Laura Wetter, Lucas Grey) represent the second test item.

Recent developments

- Co-reference resolution
- Semantic roles
- Improvements in morphological annotation
- Speech annotation: first proposal

Work in progress for the future

- Speech annotation, phonemes
- FoLiA Set Definitions and deep validation
- More tools and applications



New website: <http://proycon.github.com/fofia>
Later today: FoLiA demo session including Frog, ucto, Python
library.

Questions?