

## FoLiA: Format for Linguistic Annotation

# FoLiA: Format for Linguistic Annotation

Current state of affairs



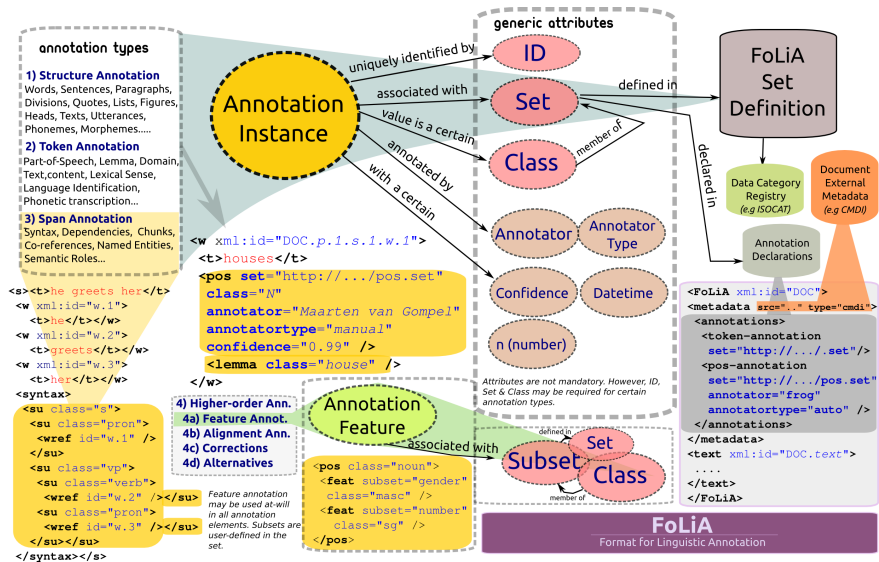
## Intended Applications

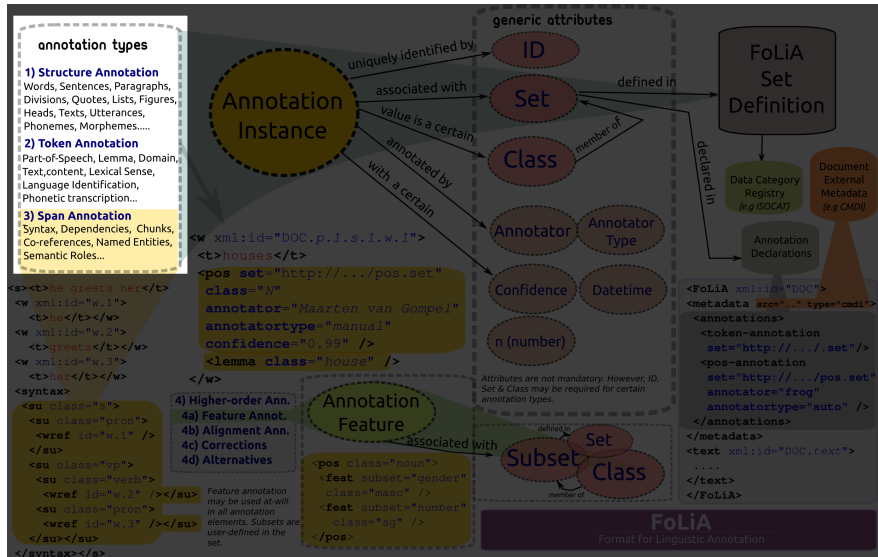
- as a corpus **storage** format
- as a language resource **exchange** format

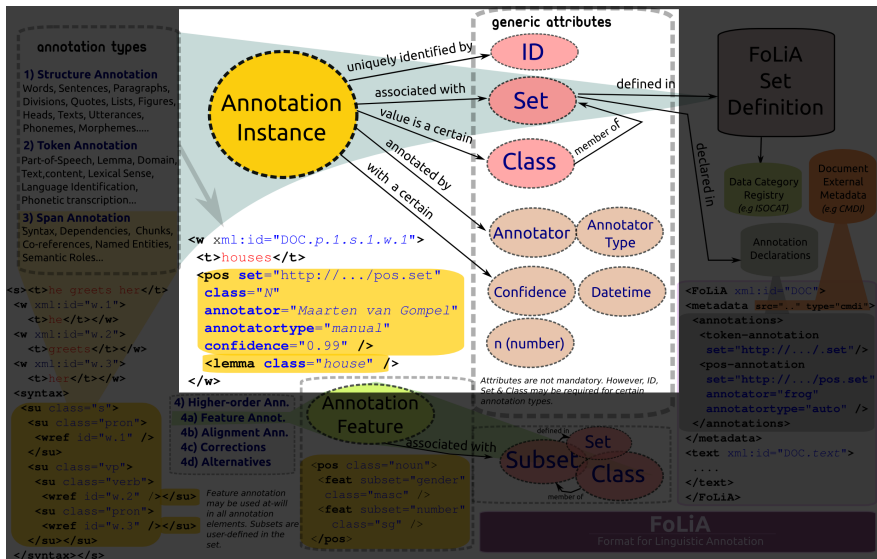
## FoLiA Characteristics

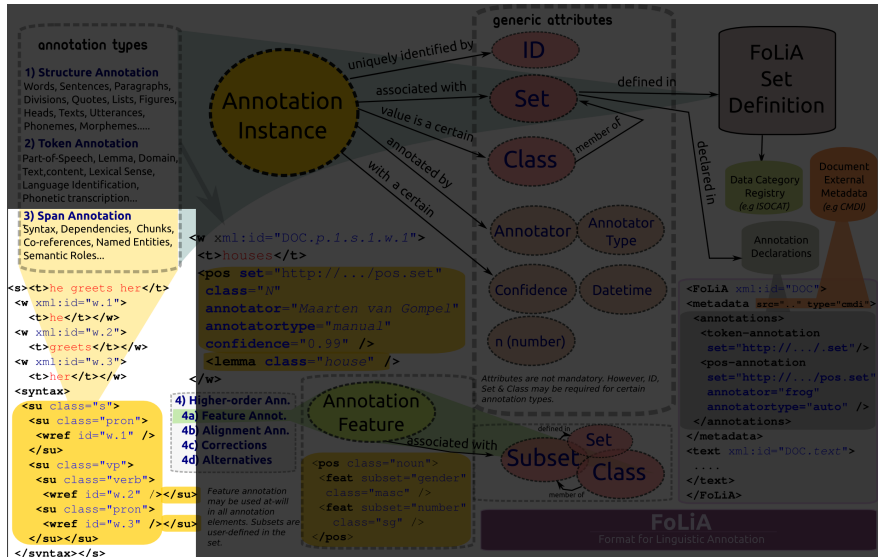
- **Uniform** setup – Consistent and uniform paradigm unifying various kinds of annotation. Not committed to any label set.
- **Extensible & Flexible** – Easy to extend
- **Expressive** – Verbose expression of annotations, their annotators, timestamps, etc... Moreover, support for *alternative* annotations.
- **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.

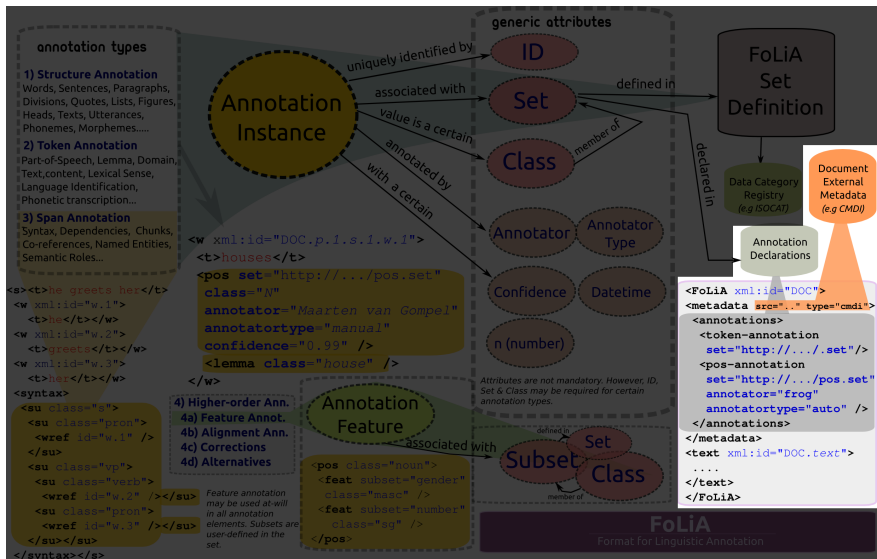
## Paradigm

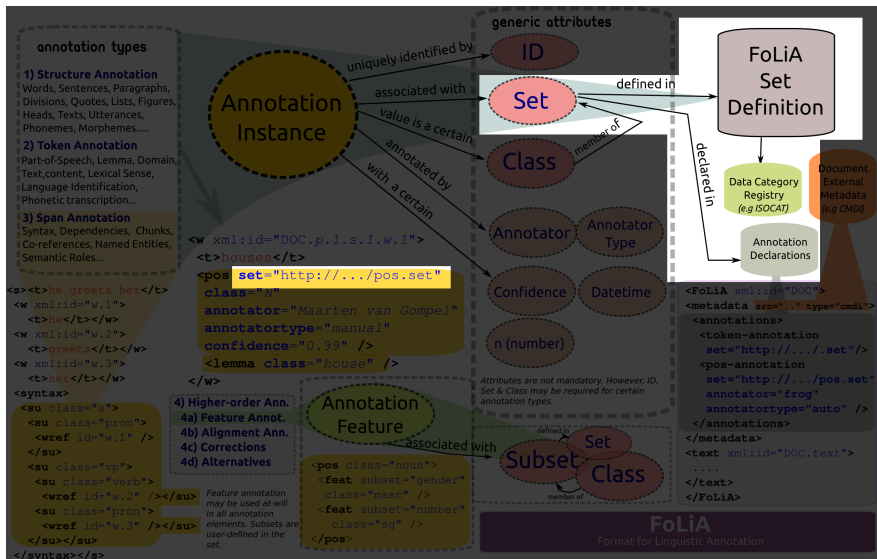




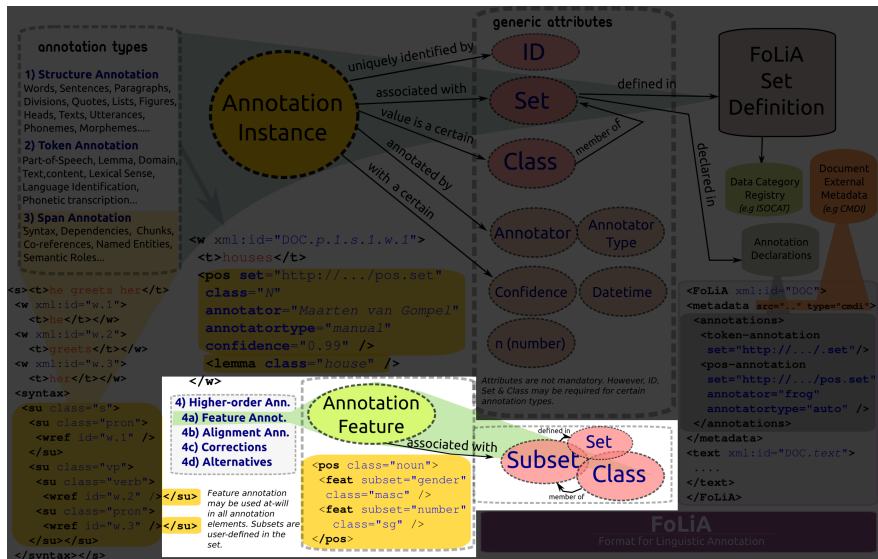


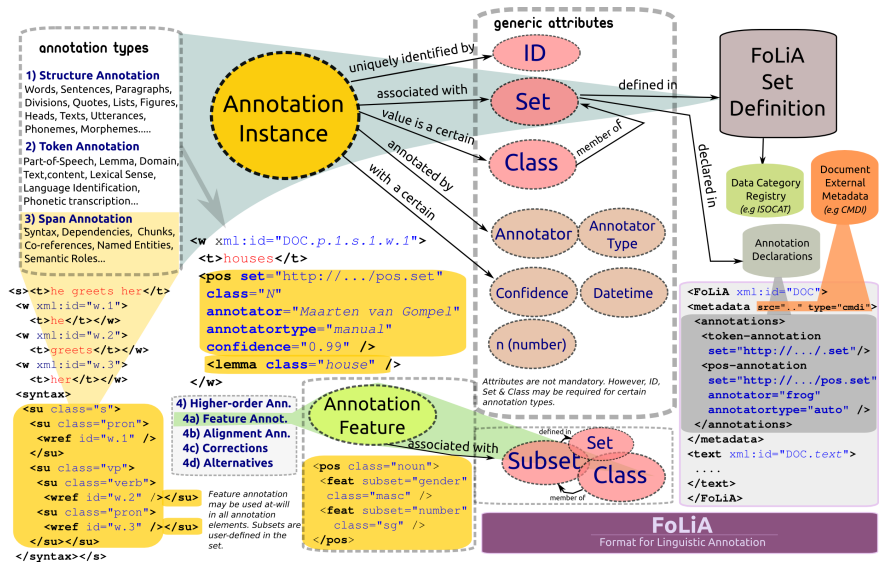












## Tools for working with FoLiA

- 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/folia_doc.xml')

#grab a specific sentence from the index
sentence = doc['folia_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['folia_doc.s.1.w.5']
word.append( folia.PosAnnotation, cls="N",
             set="http://some/url/cgn.set")

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

The screenshot shows a web browser window with the URL `ilk.uvt.nl/folia/example.xml`. The page title is "Stemma nl". The main content area displays a document page titled "Stemma".

**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 weer te geven.

**Werkwijze**

Hiervoor worden de handschriften genummerd en gedateerd zodat ze op de juiste plaats van hun afstammingsgeschiedenis 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 stippelijntje op een onzekere verwantschap duidt.

- Eerste testitem
- Tweede testitem

A family tree diagram is shown at the bottom of the page. It illustrates the relationships between various individuals, with names enclosed in colored boxes (blue, red, or purple) and connected by lines. The diagram shows a central figure, Lucas Grey, branching into Mary Grey and Jason Grey. Mary Grey is connected to Fred Smith, Jane Smith, and Sean Grey. Jason Grey is connected to Jessica Grey, who is further connected to Joseph Wetter, John Wetter, and Laura Wetter.

A metadata popup is visible on the right side of the page, displaying the following information:

ID	WR-P-E-J-000000001.p.1.s.2.w.7
PoS	N(soort,ev,basis,zijd,stan)
Lemma	stamboom

## 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?