

FoLiA: Format for Linguistic Annotation

Maarten van Gompel
Radboud University Nijmegen

20-01-2012



Introduction

What is FoLiA?

- Generalised XML-based format for a wide variety of linguistic annotation

Characteristics

- Generalised** paradigm – Single universal paradigm applicable to all kinds of annotations; as few ad-hoc provisions as possible. Not committed to any label set.
- Extensible** – Unsupported annotation types can be added fairly easily.
- Expressive** – Verbose expression of annotations, their annotators, timestamps, etc... Moreover, support for *alternative* annotations.
- Formalised** – Validation on two levels: shallow and deep. The latter validates the used label set and allows for links with for instance ISOcat.

Intended Applications

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

Properties

- One document, one text, one XML file – containing all annotations.
- Annotation types and label sets must be declared in the document header
- Document metadata can be either included in the file (limited), or by reference to external CMDI or IMDI (preferred)

Trade-off: Expressivity versus Computing Efficiency

- FoLiA aims at expressivity rather than computing efficiency.
 - XML and FoLiA overhead: Not ideal for real-time or resource-constrained applications
 - **Conversion** to less expressive, more efficient, formats.

Why (yet) another format?

- Many ad-hoc and legacy annotation formats (CGN, Tadpole column format)
- Many theoretic and specialised annotation formats with limited scope (LAF, SynAF, MAF, TEI)
- Bottom-up rather than top-down development: FoLiA arose from practical need, immediately developed alongside practical programming libraries and applications.
- De-facto-standard: D-COI XML

Dissemination

- SoNaR
- TTNWW
- DutchSemCor
- Valkuil.net
- Frog & Ucto

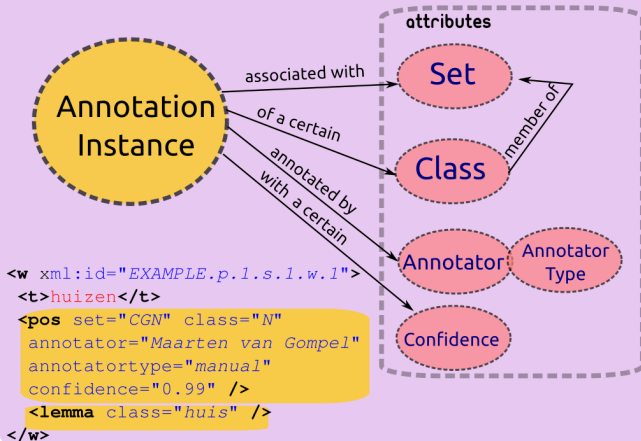
Paradigm

Paradigm: Annotation Categories

Four categories of annotation:

- **Structure Annotation** - Elements denoting document structure
 - E.g: Divisions, Header, Paragraphs, Sentences, Lists, Figures, Gaps, Quote
- **Token Annotation** - Linguistic Annotations pertaining to a single token (inline annotation)
 - E.g: Part of Speech Annotation, Lemma Annotation, Lexical Semantic Sense Annotation
- **Span Annotation** - Linguistic Annotations spanning over multiple tokens (standoff annotation)
 - E.g: Syntactic Parses, Dependency Relations, Entities/Multi-word Units
- **Subtoken Annotation** - Linguistic Annotations pertaining to a subpart of a token (standoff annotation)
 - E.g: Morphology

- uniform paradigm -



Format

```
<?xml version="1.0" encoding="utf-8"?>
<FoLiA xmlns="http://ilk.uvt.nl/FoLiA"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xml:id="example">
  <metadata type="cmdi" src="example.cmdi">
    <annotations>
      ...
    </annotations>
  </metadata>
  <text xml:id="example.text">
    ...
  </text>
</FoLiA>
```

Example

```
<p xml:id="TEST.p.1">
<t>This is a test. It has two sentences.</t>
<s xml:id="TEST.p.1.s.1">
  <t>This is a test.</t>
  <w xml:id="TEST.p.1.s.1.w.1"><t>This</t></w>
  <w xml:id="TEST.p.1.s.1.w.2"><t>is</t></w>
  ..
</s><s xml:id="TEST.p.1.s.2">...</s>
</p>
```

Characteristics of basic structure

- ① **Structure Elements:** Paragraphs, Sentences, Words/Tokens
- ② More: Division, Head, List, ListItem, Figure, Gap...
- ③ Unique identifiers
- ④ Text content element (t) holds actual text.

Token Annotation

Token annotation occurs within the scope of a word/token (*w*) element.

Example

PoS and Lemma Annotation:

```
<w xml:id="example.p.1.s.1.w.2">
  <t>boot</t>
  <pos set="cgn" class="n"
    annotator="Maarten_van_Gompel" annotortype="manual" />
  <lemma set="english-lemmas" class="boot" />
  <sense set="cornetto" class="D_N12345" annotator="supwsd1"
    annotortype="auto" confidence="0.65" />
</w>
```

Token Annotations with subsets

For all annotation types; **subsets** can be used for more refined annotations.

Example

```
<w xml:id="example.p.1.s.1.w.2">
  <t>boot</t>
  <pos set="cgn" class="N(soort,ev,basis,zijn,stan)">
    <feat subset="head" class="N" />
    <feat subset="ntype" class="soort" />
    <feat subset="number" class="ev" />
    <feat subset="degree" class="basis" />
    <feat subset="gender" class="zijd" />
    <feat subset="case" class="stan" />
  </pos>
</w>
```

Span Annotation

- Token Annotation is not sufficient, some annotations span over multiple tokens (not necessarily consecutive)
- Spanning multiple tokens can produce nesting problems (e.g. $A(BC)D$ and $AB(CD)$)
- **Solution:** Span Annotation using standoff notation

Properties

- **Applications:** Syntactic Parses, Chunking, Dependency Relations, Entities/Multi-Word Units
- **Layers:** Each type of span annotation is placed within an *annotation layer*, annotation layers are usually embedded within *sentences* (s))
- Same paradigm: Set, class, annotator, confidence, etc...

```

<s xml:id="example.p.1.s.1">
  <t>The Dalai Lama greeted him.</t>
  <w xml:id="example.p.1.s.1.w.1"><t>The</t></w>
  <w xml:id="example.p.1.s.1.w.2"><t>Dalai</t></w>
  <w xml:id="example.p.1.s.1.w.3"><t>Lama</t></w>
  <w xml:id="example.p.1.s.1.w.4"><t>greeted</t></w>
  <w xml:id="example.p.1.s.1.w.5"><t>him</t></w>
  <w xml:id="example.p.1.s.1.w.6"><t>.</t></w>
  <entities>
    <entity xml:id="example.p.1.s.1.entity.1" class="person">
      <wref xml:id="example.p.1.s.1.w.2" />
      <wref xml:id="example.p.1.s.1.w.3" />
    </entity>
  </entities>
</s>

```

```

<syntax>
<su xml:id="example.p.1.s.1.su.1" class="s">
  <su xml:id="example.p.1.s.1.su.1_1" class="np">
    <su xml:id="example.p.1.s.1.su.1_1_1" class="det">
      <wref xml:id="example.p.1.s.1.w.1" />
    </su>
    <su xml:id="example.p.1.s.1.su.1_1_2" class="pn">
      <wref xml:id="example.p.1.s.1.w.2" />
      <wref xml:id="example.p.1.s.1.w.3" />
    </su>
  </su>
</su>
<su xml:id="example.p.1.s.1.su.1_2" class="vp">
  <su xml:id="example.p.1.s.1.su.1_1_1" class="v">
    <wref xml:id="example.p.1.s.1.w.4" />
  </su>
  <su xml:id="example.p.1.s.1.su.1_1_2" class="pron">
    <wref xml:id="example.p.1.s.1.w.5" />
  </su>
</su>
</syntax>

```

Tools for working with FoLiA

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

Applications

- **Frog** – tagger/lemmatisation/parser suite: FoLiA output (input in later stage).
- **ucto** – tokeniser: FoLiA input and output.

Converters

- DCOI \longleftrightarrow FoLiA
- FoLiA \longrightarrow CSV (limited)

Conclusion

- **Uniformity:** generic framework with simple paradigm, XML based
- **Expressiveness:** Ability to encode many kinds of linguistic annotation, including structural annotation, alternatives, and corrections
- **Extensibility:** easy to add new annotations with the same paradigm
- A variety of tools and converters already available!

URLs

- <http://ilk.uvt.nl/folia>
- <http://github.com/proycon/folia>



Questions?

Supported Annotations (1/2)

FoLiA supports the following linguistic annotations:

- Part-of-Speech tags (with features)
- Lemmatisation
- Domain tagging
- Lexical semantic sense annotation (used in DutchSemCor)
- Named Entities / Multi-word units (used in SoNaR)
- Syntactic Parses
- Dependency Relations

Supported Annotations (2/2)

FoLiA supports the following linguistic annotations:

- Chunking
- Corrections (used in valkuil.net)
- Morphology
- Event/Time annotation
- Phonetic annotation

FoLiA: Format for Linguistic Annotation

Alternative Token Annotations

Annotations of the same type, but different sets need *not* be alternatives.

```
<w xml:id="example.p.1.s.1.w.2">
  <t>luid</t>
  <pos set="brown" class="jj" />
  <pos set="cgn" class="adj" />
</w>
```

There can be only one of the same set though, this is illegal and requires usage of alternatives instead:

```
<w xml:id="example.p.1.s.1.w.2">
  <t>luid</t>
  <pos set="cgn" class="adj" />
  <pos set="cgn" class="adv" />
</w>
```

Alternative Token Annotations

Encodes mutually exclusive alternative annotations. Any annotations that are not alternatives are considered “selected”.

```
<w xml:id="example.p.1.s.1.w.2">
  <t>bank</t>
  <sense set="wordnet3.0" class="bank%1:17:01:"
    annotator="Maarten_van_Gompel" annotatortype="manual"
    confidence="0.8">
    sloping ground near water</sense>
  <alt xml:id="example.p.1.s.1.w.2.alt.1">
    <sense set="wordnet3.0" class="bank%1:14:01:"
      annotator="WSDsystem" annotatortype="auto"
      confidence="0.6">
      financial institution</sense>
    </alt>
</w>
```

Alternative Token Annotations

All token annotations grouped as one alternative are considered dependent. Multiple alternatives are always independent:

```
<w xml:id="example.p.1.s.1.w.2">
  <t>vlieg</t>
  <pos class="N" />
  <lemma class="vlieg" />
  <alt xml:id="example.p.1.s.1.w.2.alt.1">
    <pos class="V" />
    <lemma class="vliegen" />
  </alt>
</w>
```