Annotation schemas

HAP/LAP. Corpus Linguistics.

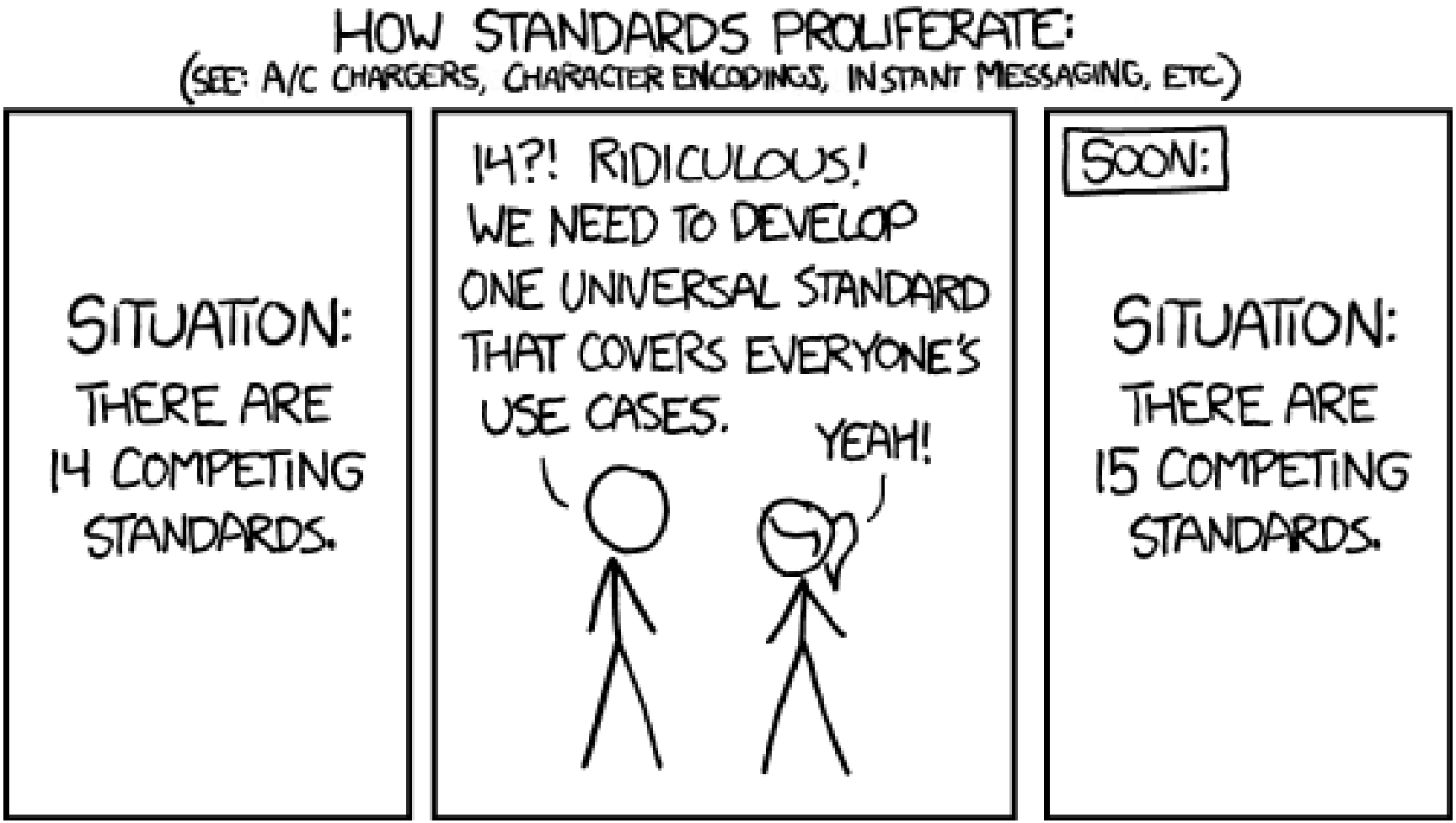
Linguistic Annotation formats

- Language-neutral annotation of text, concepts, facts, ...
- Multilingual
- Interoperability across linguistic processors
  - Annotation format is the basis for integration
- Flexibility and extensibility

But

Many annotation schemas and each one aims to be a standard

Standards...


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Standard formats

- 2 levels of standarization:
  - Conceptual interoperability: using standarized tagset for linguistic analysis
  - Structural interoperability: annotation schema

Outline

Conceptual interoperability

- Use standard tags to represent linguistic information
- A tagset for each linguistic level
- So that tools understand each other

Conceptual interoperability

- Penn Treebank tagset (`https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`)
  - Widely used for English
  - Many tools (FreeLing, Stanza) use these tags natively
  - Many corpora annotated using Penn Treebank
- PAROLE and EAGLES: morphosyntactic tags for 12 European languages
  - Not widely used
- Universal dependencies (`http://universaldependencies.org/`)
  - POS, morphology and syntax (dependency)
  - Used to annotate treebanks
  - More than 40 languages!

Conceptual interoperability

- Data Category Registry (ISOcat)
  - Many linguistic level
  - Not widely adopted
- GOLD ontology
  - Descriptive linguistics
  - Adheres to LOD
- OLiA: Ontologies of Linguistic Annotation (`http://www.acoli.informatik.uni-frankfurt.de/resources/olia/`)
  - Integrate more than 30 models of 65 languages
  - Uses description logics

Outline

1. Conceptual interoperability

2. Structural interoperability
   - Ad-hoc Schemas
   - TEI
   - NAF
   - Tipster
   - UIMA CAS
   - NIF
   - FoLiA
   - AWA

Structural interoperability

- Use unified structures to annotate linguistic information
- Deal with phenomena such as ambiguity.

Outline

1. Conceptual interoperability
2. Structural interoperability
   - Ad-hoc Schemas
   - TEI
   - NAF
   - Tipster
   - UIMA CAS
   - NIF
   - FoLiA
   - AWA

CoNLL

- Tabulated, ad-hoc format
- inline annotation, no ambiguity
- Widely used

```
U.N.        NNP   I-NP   I-ORG
official    NN    I-NP   O
Ekeus       NNP   I-NP   I-PER
heads       VBZ   I-VP   O
for         IN    I-PP   O
Baghdad     NNP   I-NP   I-LOC
.           .     O      O
```

Outline

1. Conceptual interoperability
2. Structural interoperability
   - Ad-hoc Schemas
   - TEI
   - NAF
   - Tipster
   - UIMA CAS
   - NIF
   - FoLiA
   - AWA

TEI guidelines

- There are many ways to encode linguistic information.
  - the TEI guidelines provide recommendations for markup.
  - `http://www.tei-c.org/Guidelines/P5/`
  - TEI-conformant stamp
    `http://en.wikipedia.org/wiki/Text_Encoding_Initiative`
- XML based.
  - Clarity.
  - Simplicity.
  - Formally rigorous.
  - Recognized as an international standard.

TEI guidelines

The text consists of elements:

- Almost any textual unit: word, sentence, paragraph, ...
- Uses XML marks to explicitly represent the structure
  - *This is a paragraph* .
  - <s> *This is a sentence* </s>.
- It can also represent structured information (feature structures)

TEI guidelines: main parts

- TEI divided into *modules*
- In principle, you can combine any module combination
  - but some modules are obligatory
- Two important parts: header and text
- TEI header: metadata about the document
  - author, title, publication date.
  - version.
  - information about markup.
- Text
  - Structure, segmentation (paragraphs, sentences, etc).
  - Linguistic annotations.

TEI modules

analysis	Analysis and Interpretation
certainty	Certainty and Uncertainty
core	Common Core
corpus	Metadata for Language Corpora
dictionaries	Print Dictionaries
drama	Performance Texts
figures	Tables, Formulae, Figures
gaiji	Character and Glyph Documentation
header	Common Metadata
iso-fs	Feature Structures
linking	Linking, Segmentation, and Alignment
msdescription	Manuscript Description
namesdates	Names, Dates, People, and Places
nets	Graphs, Networks, and Trees
spoken	Transcribed Speech
tagdocs	Documentation Elements
tei	TEI Infrastructure
textcrit	Text Criticism
textstructure	Default Text Structure
transcr	Transcription of Primary Sources
verse	Verse

TEI header: metadata

```
<TEI>
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>
<!-- title of the resource -->
   </title>
  </titleStmt>
  <publicationStmt>
   <p>
<!-- Information about distribution of the resource -->
   </p>
  </publicationStmt>
  <sourceDesc>
   <p>
<!-- Information about source from which the resource derives -->
   </p>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

TEI text: default structure

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <text>
    <front> <!-- front matter of copy text, if any, goes here -->
    </front>
    <body>
      <div type="part" n="1">
        <div type="chapter" n="1">
          <!-- text of part 1, chapter 1 -->
        </div>
        <div type="chapter" n="2">
          <!-- text of part 1, chapter 2 -->
        </div>
      </div>
      <div type="part" n="2">
        <div n="1" type="chapter">
          <!-- text of part 2, chapter 1 -->
        </div>
        <div n="2" type="chapter">
          <!-- text of part 2, chapter 2 -->
        </div>
      </div>
    </body>
    <back> <!-- back matter of copy text, if any, goes here --> </back>
  </text>
</TEI>
```

TEI text: dictionary

```
<entry>
  <form>
    <orth>careen</orth>
    <hyph>ca|reen</hyph>
    <pron>k@"ri:n</pron>
  </form>
  <gramGrp>
    <pos>vt</pos>
    <pos>vi</pos>
  </gramGrp>
  <sense n="1">
    <gramGrp>
      <subc>VP6A</subc>
    </gramGrp>
    <def>turn (a ship) on one side for cleaning, repairing, etc.</def>
  </sense>
  <sense n="2">
    <gramGrp>
      <subc>VP6A</subc>
      <subc>VP2A</subc>
    </gramGrp>
    <def>(cause to) tilt, lean over to one side.</def>
  </sense>
</entry>
```

TEI text: dictionary

```
<entry>
  <form><orth>xukatu, xuka, xukatzen</orth></form>
  <GramGrp><subc>du</subc><pos>ad.</pos></GramGrp>
  <usg type="time">1627</usg>
  <usg type="geo">Ipar.</usg>
  <sense n="1">
    <def>Lehortu, bustitasuna edo hezetasuna kendu.</def>
    <eg><q>Sukaldeko ontziak xukatu. Esku-gibelaz kopeta xukatu. Jesus
    maitearen begitartea oihal zuriz xukatzera. Eta apezak behar ditu
    xukatu aitama eta senar zaurtuen negarrak. Bakailao puska egosiak
    oihal batean xuka itzazu.</q></eg>
  </sense>
  <sense n="2">
    <def>Idortu; agortu.</def>
    <eg><q>Nola lur xukatu eta idortu bat euriaren beharrean.</q></eg>
    <sense n="n1">
      <eg><q>Bere odol guztia edaten, xukatzen. Juduek xukatzen dute
      Frantziaren aberastasuna.</q></eg>
    </sense>
  </sense>
</entry>
```

TEI text: linguistic annotations

```
<text xml:id="A01" decls="A">
 <body>
  <p>
   <s n="1">
    <w type="AT">The</w>
    <w type="NP" subtype="TL">Fulton</w>
    <w type="NN" subtype="TL">County</w>
    <w type="JJ" subtype="TL">Grand</w>
    <w type="NN" subtype="TL">Jury</w>
    <w type="VBD">said</w>
    <w type="NR">Friday</w>
    <w type="AT">an</w>
    <w type="NN">investigation</w>
    <w type="IN">of</w>
    <w type="NPg">Atlanta's</w>
    <w type="JJ">recent</w>
    <w type="NN">primary</w>
    <w type="NN">election</w>
    <w type="VBD">produced</w>
    ...
 </s></p></body>
</text>
```

Corpus in TEI format

http://wiki.tei-c.org/index.php/Samples

Outline

1. Conceptual interoperability
2. Structural interoperability
   - Ad-hoc Schemas
   - TEI
   - NAF
   - Tipster
   - UIMA CAS
   - NIF
   - FoLiA
   - AWA

NAF: NLP annotation format

- Designed within the NewsReader project
- Comes from KAF (Kyoto Annotation Format)
- Compatible with main standards
  - LAF, Ide *et al.*, 2003
  - GATE (Cunningham *et al.*, 1996)
  - UIMA, (Ferrucci and Lally, 2004)
  - ...
- Represents linguistic annotations.
- Stand-off, multi-layered annotation format.
- Based on XML.
- Allows parallel processing.
- Can be exported to RDF triplets.

NLP Annotation Format (NAF)

- Covers many linguistic levels:
  - Header, including:
    - Document metadata (creation time, author, etc)
    - LP processors which created the NAF.
  - Raw layer.
  - Tokenization, Segmentation.
  - Morphosyntax (POS tagging).
  - Syntax (dependencies and constituents).
  - Named Entities Recognition.
  - Word Sense Disambiguation, Named Entity Disambiguation.
  - Co-reference resolution.
  - Semantic Role Labeling.
  - Time expressions.
  - Factuality.

NAF layers

```
<NAF version="v3" xml:lang="en">
  <nafHeader>...</nafHeader> <!-- header -->
  <raw>...</raw>             <!-- raw text -->
  <text>...</text>           <!-- tokens -->
  <terms>...</terms>         <!-- lemmas, POS, externalRefs -->
  <deps>...</deps>           <!-- dependency syntax -->
  <constituency>...</>       <!-- constituent syntax -->
  <entities>...</entities>   <!-- Named entities -->
  <coreferences> ..</>       <!-- coreferences (nominal) -->
  <srl>...</srl>             <!-- Semantic Rol Labeling -->
  <timeExpressions>...</>    <!-- Time expressions (timex3) -->
  <temporalRelations>...</>  <!-- Temporal relations -->
  <causalRRelations>...</>   <!-- Causal relations -->
  <factualitylayer>...</>    <!-- Factuality layer -->
</NAF>
```

- NAF is multi-layered
  - each layer represents a linguistic analysis level (more or less)
- Layers refer to lower levels

NAF header

```
<NAF version="v3" xml:lang="en">
  <nafHeader>
    <fileDesc author="Michael Green" creationtime="2014-09-15"
              filename="obama" filetype="HTML"
              title="President Barack Obama puts brave face ..."/>
    <public publicId="62527d3ffeeb6966980c6e886a5265ec334e54e0"
            uri="..."/>
    <linguisticProcessors layer="text">
      <lp name="ixa-pipe-tok-en" version="1.5.0"
           beginTimestamp="2014-09-15T09:48:38+0200"
           endTimestamp="2014-09-15T09:48:38+0200" />
    </linguisticProcessors>
    <linguisticProcessors layer="terms">
      <lp name="ixa-pipe-pos-en" version="1.0.0"
           beginTimestamp="2014-09-15T09:48:39+0200"
           endTimestamp="2014-09-15T09:48:39+0200"/>
    </linguisticProcessors>
    ...
  </nafHeader>
```

NAF: Raw layer

- Store the actual text from the document.
- Use the CDATA section

```
<?xml version="1.0" encoding="UTF-8"?>
<NAF ...>
<raw><![CDATA[
President Barack Obama puts brave face on schoolgirl's Beyonce snub.

A Washington DC schoolgirl has prompted laughter from the US President
after admitting that she had hoped the special guest visitor to her
school was going to be Beyonce.

Barack Obama seemed understanding, admitting that his daughters would
prefer a visit from the singing superstar. First Lady Michelle Obama
also agreed that she would "rather see Beyonce."

The president and first lady were taking part in an event at the
Inspired Teaching charter school, filling schoolbags with toys for
homeless children.]]></raw></NAF>
```

NAF: Tokenization

```
<NAF version="v3" xml:lang="en">
<raw>President Barack Obama puts brave face on schoolgirl's Beyonce ...</raw>
<text>
  <wf id="w1" sent="1" para="1" offset="0" length="9">President</wf>
  <wf id="w2" sent="1" para="1" offset="10" length="6">Barack</wf>
  <wf id="w3" sent="1" para="1" offset="17" length="5">Obama</wf>
  <wf id="w4" sent="1" para="1" offset="23" length="4">puts</wf>
  <wf id="w5" sent="1" para="1" offset="28" length="5">brave</wf>
  <wf id="w6" sent="1" para="1" offset="34" length="4">face</wf>
  <wf id="w7" sent="1" para="1" offset="39" length="2">on</wf>
  <wf id="w8" sent="1" para="1" offset="42" length="10">schoolgirl</wf>
  <wf id="w9" sent="1" para="1" offset="52" length="2">'s</wf>
  <wf id="w10" sent="1" para="1" offset="55" length="7">Beyonce</wf>
  ...
</text>
```

NAF: Term layer

- Represent lexical units (terms)
  - May group tokens ("New York")
- Include many information for the terms:
  - Lemmatization
  - POS tagging
  - External references (senses)

```
<NAF version="v3" xml:lang="en">
  <terms>
    <term id="t1" lemma="President" morphofeat="NNP" pos="R">
      <span><target id="w1"/></span>
    </term>
    <term id="t2" lemma="Barack" morphofeat="NNP" pos="R">
      <span><target id="w2"/></span>
    </term>
    <term id="t3" lemma="Obama" morphofeat="NNP" pos="R">
      <span><target id="w3"/></span>
    </term>
    <term id="t4" lemma="put" morphofeat="V" pos="VBZ">
      <span><target id="w4"/></span>
    </term>
  </terms>
  </NAF>
```

NAF: Term layer

- Link terms with senses (WSD)

```
<!--puts-->
<term id="t4" type="open" lemma="put" pos="V" morphofeat="VBZ">
  <span>
    <target id="w4"/>
  </span>
  <externalReferences>
  <externalRef resource="wn30g.bin64" reference="ili-30-01494310-v"
               confidence="0.419871"/>
  <externalRef resource="wn30g.bin64" reference="ili-30-01493741-v"
               confidence="0.333348"/>
  <externalRef resource="wn30g.bin64" reference="ili-30-00981276-v"
                confidence="0.0983404"/>
  ...
</term>
```

NAF: Named Entities

```
<entities>
  <entity id="e1" type="PERSON">
    <references><!--Barack Obama-->
      <span><target id="t2"/><target id="t3"/></span>
    </references>
    <externalReferences>
      <externalRef reference="http://dbpedia.org/resource/Barack_Obama"
                   confidence="0.99999934"/>
    </externalReferences>
  </entity>
  <entity id="e2" type="PERSON">
    <references><!--Beyonce-->
      <span><target id="t10"/></span>
    </references>
    <externalReferences>
      <externalRef reference="http://dbpedia.org/resource/Beyoncé_Knowles"
                   confidence="1.0"/>
    </externalReferences>
  </entity>
```

NAF: Named Entities (cont.)

```
<entity id="e3" type="LOCATION">
  <references><!--Washington-->
    <span><target id="t14"/></span>
  </references>
  <externalReferences>
    <externalRef reference="http://dbpedia.org/resource/Washington,_D.C."
                 confidence="0.9635455"/>
    <externalRef reference="http://dbpedia.org/resource/Washington_(state)"
                 confidence="0.030298341"/>
  </externalReferences>
</entity>
```

NAF: Coreferences

- Clusters of terms which share the same referent.

```
<coref id="co1">
  <!--President Barack Obama-->
  <span>
    <target id="t1"/><target id="t2"/><target id="t3"/>
  </span>
  <!--US President-->
  <span>
    <target id="t22"/><target id="t23"/>
  </span>
  <!--Barack Obama-->
  <span>
    <target id="t43"/><target id="t44"/>
  </span>
  <!--his-->
  <span>
    <target id="t50"/>
  </span>
  <!--The president-->
  <span>
    <target id="t76"/><target id="t77"/>
  </span>
</coref>
```

NAF: Semantic role labeling

- Detects predicated and associated arguments.
- Link predicates with external resources
  - PropBank
  - FrameNet

NAF: Semantic role labeling

```
<!--t4 puts : A0[t1 President] A1[t5 brave] A2[t7 on]-->
<predicate id="pr1">
  <!--puts-->
  <span><target id="t4"/></span>
  <externalReferences>
     <externalRef resource="PropBank" reference="put.01"/>
  </externalReferences>
  <role id="rl1" semRole="A0">
    <!--President Barack Obama-->
    <span><target id="t1"/><target id="t2"/><target id="t3" head="yes"/>
    </span>
  </role>
  <role id="rl2" semRole="A1">
    <!--brave face-->
    <span><target id="t5"/><target id="t6" head="yes"/>
    </span>
  </role>
  <role id="rl3" semRole="A2">
    <!--on schoolgirl 's Beyonce snub-->
    <span><target id="t7" head="yes"/>
    ...
    </span>
  </role>
</predicate>
```

NAF: Time expressions

- Follows TimeML `timex3` tag
- Represents normalized values of time expressions
  - Points in time ("now")
  - Time ranges ("January")
  - Durations ("two hours")
  - Quantifications ("every hour")
  - ...

```
<timeExpressions>
  <timex3 id="tmx0" type="DATE" value="2014-09-15"/>
</timeExpressions>
```

NAF: Temporal relations

- Among:
  - two temporal expressions.
  - a temporal expression and an event.

```
<temporalRelations>
  <!--BEFORE(pr4, tmx0)-->
  <tlink id="tlink0" from="pr4" to="tmx0"
         fromType="event" toType="timex" relType="BEFORE"/>
  <!--BEFORE(pr13, tmx0)-->
  <tlink id="tlink1" from="pr13" to="tmx0"
         fromType="event" toType="timex" relType="BEFORE"/>
  <!--IS_INCLUDED(pr1, tmx0)-->
  <tlink id="tlink3" from="pr1" to="tmx0"
         fromType="event" toType="timex" relType="IS_INCLUDED"/>
</temporalRelations>
```

Outline

1. Conceptual interoperability
2. Structural interoperability
   - Ad-hoc Schemas
   - TEI
   - NAF
   - Tipster
   - UIMA CAS
   - NIF
   - FoLiA
   - AWA

Tipster

- First *stand-off* format
- Annotations are separated from text (using offsets)

Text				
Cyndi savored the soup.				
Annotations				
Id	Type	Start	End	Attributes
1	token	0	5	pos=NP
2	token	6	13	pos=VBD
3	token	14	17	pos=DT
4	token	18	22	pos=NN
5	token	22	23	
6	name	0	5	name_type=person
7	sentence	0	23	constituents=[1],[2],[3],[4],[5]
8	parse	0	5	symbol=NP constituents= [1]
9	parse	14	22	symbol=NP constituents=[3],[4]
10	parse	6	22	symbol=VP constituents=[2],[9]
11	parse	0	22	symbol=S constituents=[8],[10]

Tipster

- GATE format is based on tipster

```
<TextWithNodes><Node id="0"/>Seven <Node id="6"/>UK<Node id="8"/>
airlines including <Node id="28"/>...</TextWithNodes>
<AnnotationSet Name="Key" >
  <Annotation Id="43" Type="Organization" StartNode="1367" EndNode="1381">
    <Feature>
      <Name className="java.lang.String">rule1</Name>
      <Value className="java.lang.String">OrgXBase</Value>
    </Feature>
</AnnotationSet>
<AnnotationSet Name="Original markups" >
  <Annotation Id="79" Type="p" StartNode="938" EndNode="1127">
  </Annotation>
  <Annotation Id="112" Type="TEXT" StartNode="0" EndNode="2707">
  </Annotation>
</AnnotationSet>
```

Outline

1. Conceptual interoperability
2. Structural interoperability
   - Ad-hoc Schemas
   - TEI
   - NAF
   - Tipster
   - UIMA CAS
   - NIF
   - FoLiA
   - AWA

UIMA CAS

- Annotation format of IBM watson
- stand-off
- Uses feature structures to encode linguistic information
- Does not specify which features/tags to use
- Scalable to distributed architectures

UIMA CAS

```
<xmi:XMI >
  <cas:NULL xmi:id="0"/>
  <cas:Sofa xmi:id="1" sofaNum="105" mimeType="text"
       sofaString="President Barack Obama puts brave face on schoolgirl's" .. <</>
  <ixatypes:tok xmi:id="106" sofa="105" begin="0" end="9"/>
  <ixatypes:tok xmi:id="107" sofa="105" begin="10" end="16"/>
  <ixatypes:tok xmi:id="108" sofa="105" begin="17" end="22"/>
...
  <ixatypes:lexUnit xmi:id="207" sofa="105" begin="0" end="9"
                    lemma="president" posTag="N" morphofeat="NN"/>
  <ixatypes:lexUnit xmi:id="208" sofa="105" begin="10" end="16"
                    lemma="barack" posTag="R" morphofeat="NNP"/>
  <ixatypes:lexUnit xmi:id="209" sofa="105" begin="17" end="22"
                    lemma="obama" posTag="R" morphofeat="NNP"/>
...</XMI>
```

Outline

1. Conceptual interoperability
2. Structural interoperability
   - Ad-hoc Schemas
   - TEI
   - NAF
   - Tipster
   - UIMA CAS
   - NIF
   - FoLiA
   - AWA

NIF (NLP Interchange Format)

- Annotations in RDF (Linked Data)
- Uses Olia for TAG definition

```
<http://freme-project.eu/#char=5,12>
        a                       nif:Word , nif:RFC5147String ;
        nif:anchorOf            "Clinton" ;
        nif:beginIndex          "5" ;
        nif:endIndex            "12" ;
        nif:nextWord            <http://freme-project.eu/#char=13,15> ;
        nif:previousWord        <http://freme-project.eu/#char=0,4> ;
        nif:referenceContext    <http://freme-project.eu/#char=0,217> ;
        nif:sentence            <http://freme-project.eu/#char=0,160> ;
        itsrdf:taIdentRef       <http://dbpedia.org/resource/Bill_Clinton> .

<http://freme-project.eu/#char=51,61>
        a                       nif:RFC5147String , nif:Word ;
        nif:anchorOf            "Department" ;
        nif:beginIndex          "51" ;
        nif:endIndex            "61" ;
        nif:nextWord            <http://freme-project.eu/#char=62,71> ;
        nif:previousWord        <http://freme-project.eu/#char=45,50> ;
        nif:referenceContext    <http://freme-project.eu/#char=0,217> ;
        nif:sentence            <http://freme-project.eu/#char=0,160> ;
        itsrdf:taIdentRef       <http://dbpedia.org/resource/Departments_of_France>
```

Outline

1. Conceptual interoperability
2. Structural interoperability
   - Ad-hoc Schemas
   - TEI
   - NAF
   - Tipster
   - UIMA CAS
   - NIF
   - FoLiA
   - AWA

FoLiA

- Mixes inline and stand-off annotations
- Annotates also the document structure

```
<FoLiA ... >
  <metadata type="native">
     <annotations> <token-annotation annotator="ilktok" annotatortype="auto" />
     ... </annotations>
  </metadata>
  <text xml:id="WR-P-E-J-0000000001.text"><lang class="nl"/>
    <div xml:id="WR-P-E-J-0000000001.div0.1" class="chapter">
      <p xml:id="WR-P-E-J-0000000001.p.1" class="firstparagraph">
        <s xml:id="WR-P-E-J-0000000001.p.1.s.1">
          <w xml:id="WR-P-E-J-0000000001.p.1.s.1.w.1">
            <t>Stemma</t>
            <pos class="N(eigen,ev,basis,zijd,stan)"/>
            <lemma class="Stemma"/>
          </w>
                  ...
          <entities>
            <entity class="ander_woord" set="mwu-set">
              <wref id="WR-P-E-J-0000000001.p.1.s.1.w.4" t="ander"/>
              <wref id="WR-P-E-J-0000000001.p.1.s.1.w.5" t="woord"/>
            </entity>
          </entities>
          <syntax>...</syntax>
          <dependencies>...</dependencies>
          <chunking>...</chunking>
          <timing>...</timing>
        </s></p></div></text></FoLiA>
```

Outline

1. Conceptual interoperability
2. Structural interoperability
   - Ad-hoc Schemas
   - TEI
   - NAF
   - Tipster
   - UIMA CAS
   - NIF
   - FoLiA
   - AWA

AWA: Annotation Web Architecture (Artola et al. 2007)

- General purpose format for representing information produced by different linguistic processors.
- Deals with morphology rich languages.
- Many implementations:
  - XML
  - Relational DB

AWA: multi-layered schema

- AWA is stand-off and multi-layer
- Annotations from one layer refer to previous layers
- Currently, it supports:
  - Tokenization
  - Word segmentation
  - Morpho-syntax
  - Lemmatization
  - Named Entities
  - Co-reference
  - Dependency parsing

AWA

- Three basic elements:
  - Anchors: The source of the annotation.
  - Linguistic Information: The analysis (at any level)
  - Links: The actual annotation.

AWA: anchors

- Textual anchors
  - word offsets.
  - XPath expressions.
- Linguistic annotations generated from previous layers
- *joins*: groups of Anchors

AWA: Linguistic Interpretations

- Structured information describing a linguistic analysis.
  - Segmentation
  - Lemmatization
  - ...
- Encoded using feature structures (FS).

AWA: Linguistic Interpretations

```
<fs id="S-A-CNOUN-3" type="Segmentation">
 <f name="Form"><str>goikoekin</str></f>
 <f name="Lemma-Morph" org="list">
  <fs type="LemmaSeg">
   <f name="Lex">
    <fs type="Key">
     <f name="Head"><str>goi</str></f>
     <f name="HomId"><nbr value="2"/></f>
    </fs>
   </f>
   <f name="Feats">
    <fs type="FeatList">
     <f name="POS"><sym value="NOUN"/></f>
     <f name="SUBC"><sym value="COMMON"/></f>
     <f name="ANIM"><minus/></f>
    </fs>
   </f>
   <f name="Twol"><str>goi</str></f>
  </fs>
  <fs type="MorphSeg">
   <f name="Lex">
    <fs type="Key">
     <f name="Head"><str>0</str></f>
     <f name="HomId"><nbr value="5"/></f>
    </fs>
   </f>
   <f name="Feats">
    <fs type="FeatList">
     <f name="POS"><sym value="DEC"/></f>
     <f name="NUM"><sym value="S"/></f>
     <f name="DEF"><sym value="M"/></f>
    </fs>
   </f>
   <f name="Twol"><str>0</str></f>
  </fs>
  <fs type="MorphSeg">
   <f name="Lex">
    <fs type="Key">
     <f name="Head"><str>ko</str></f>
     <f name="HomId"><nbr value="2"/></f>
```

```
   <f name="Feats">
    <fs type="FeatList">
     <f name="POS"><sym value="DEC"/></f>
     <f name="CASE"><sym value="GEL"/></f>
     <f name="FSL" org="list">
      <sym value="@&lt;NCOMPL"/>
      <sym value="@NCOMPL&gt;"/>
     </f>
    </fs>
   </f>
   <f name="Twol"><str>ko</str></f>
  </fs>
  <fs type="LemmaSeg">
   <f name="Lex">
    <fs type="Key">
     <f name="Head"><str>0</str></f>
     <f name="HomId"><nbr value="14"/></f>
    </fs>
   </f>
   <f name="Feats">
    <fs type="FeatList">
     <f name="POS"><sym value="ELL"/></f>
    </fs>
   </f>
   <f name="Twol"><str>0</str></f>
  </fs>
  <fs type="MorphSeg">
   <f name="Lex">
    <fs type="Key">
     <f name="Head"><str>ekin</str></f>
     <f name="HomId"><nbr value="3"/></f>
    </fs>
   </f>
   <f name="Feats">
    <fs type="FeatList">
     <f name="POS"><sym value="DEC"/></f>
     <f name="CASE"><sym value="ASSOC"/></f>
     <f name="NUM"><sym value="P"/></f>
     <f name="DEF"><sym value="M"/></f>
    </fs>
```

AWA: Links

- Represent the actual annotation.
- Associates one anchor with one linguistic interpretation.
- Ambiguity:
  - More than one link associated with the same anchor.
  - Links contain an attribute with confidence value.

Interpretational anchors

- Because anchors are interpretation, there is a risk for exponential grow of links.
- Interpretational anchors.
- Consider:

```
publikoak :: lSfI4 :: publiko.A.ABS.P.@OBJ
publikoak :: lSfI7 :: publiko.A.ERG.S.@SUBJ
publikoak :: lSfI8 :: publiko.N.ABS.P.@OBJ
publikoak :: lSfI11 :: publiko.N.ERG.S.@SUBJ
```

- We want to link those interpretations to two WordNet senses:

  `euswn-publiko.n.1` (noun)
  `euswn-publiko.a.1` (adjective)

Interpretational anchors

- We want to attach a **adjective** sense to interpretations `lSfi4` and `lSfi7`
- We want to attach a **noun** sense to interpretations `lSfi8` and `lSfi11`
- We group those interpretations by means of interpretational anchors (`lSfiSet1, lSfiSet2`), and link these to the appropriate senses.

```
publikoak :: lSfI4 :: publiko.A.ABS.P.@OBJ
publikoak :: lSfI7 :: publiko.A.ERG.S.@SUBJ
publikoak :: lSfI8 :: publiko.N.ABS.P.@OBJ
publikoak :: lSfI11 :: publiko.N.ERG.S.@SUBJ

//interpretational anchors:
lSfISet1 = {lSfI8, lSfI11}
lSfISet2 = {lSfI4, lSfI7}

lSfISet1 :: wsdI1, 0.83 :: euswn-publiko.n.1
lSfISet2 :: wsdI2, 0.74 :: euswn-publiko.a.1
```