



From <tiger2/> to ISOTiger – Community Driven Developments for Syntax Annotation in SynAF



Sonja Bosch, University of South Africa (UNISA) – Kerstin Eckart, University of Stuttgart – Gertrud Faaß, University of Hildesheim – Ulrich Heid, University of Hildesheim – Kiyong Lee, Korea University – Antonio Pareja-Lora, Universidad Complutense de Madrid – Laurette Pretorius, University of South Africa (UNISA) – Laurent Romary, Humboldt-Universität zu Berlin – Andreas Witt, Institut für Deutsche Sprache – Amir Zeldes, Georgetown University – Florian Zipser, Humboldt-Universität zu Berlin
E-mail: kerstin.eckart@ims.uni-stuttgart.de, laurent.romary@inria.fr

Architecture of corpus standards

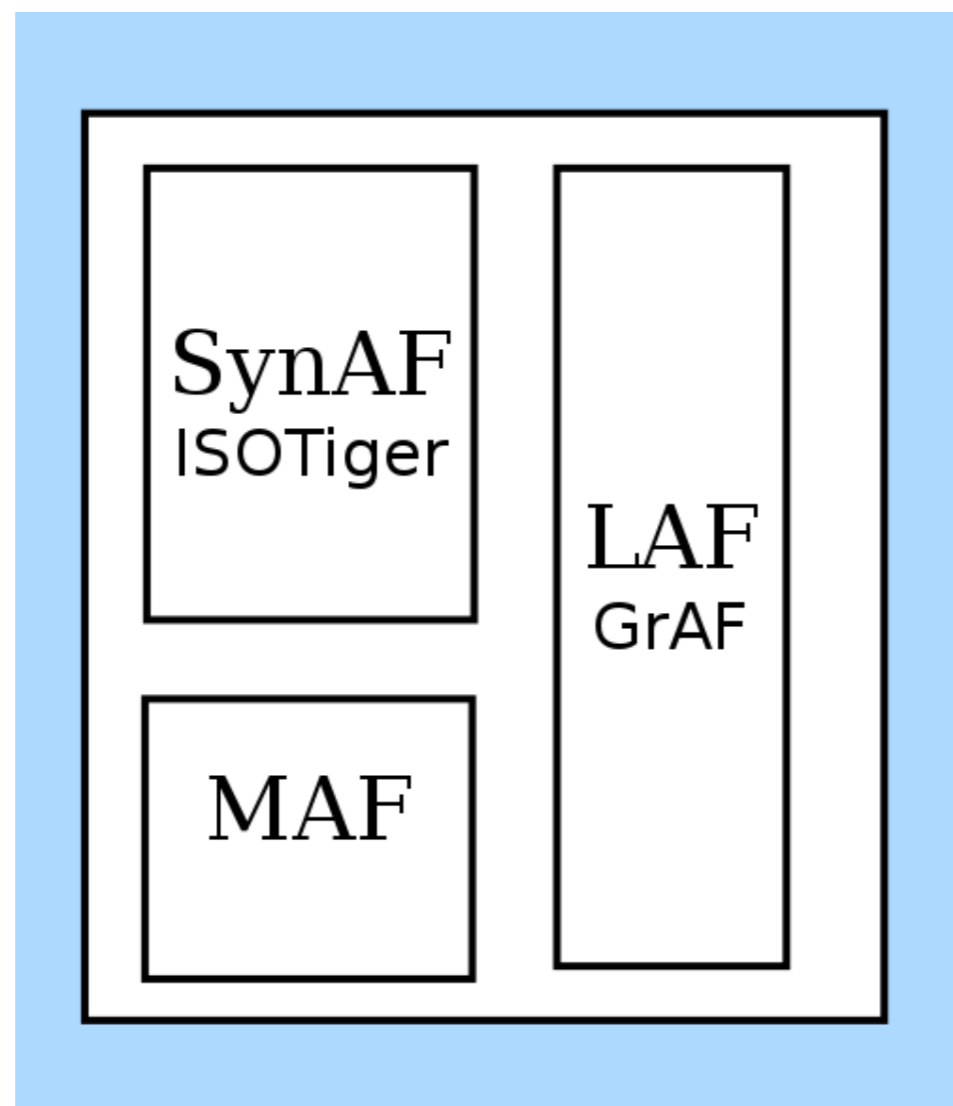
Specifications and serializations (XML)

• SynAF: [ISO 24615]

- Grammatical features
- Phrase structures
- Dependency structures
- XML: ISOTiger

• MAF: [ISO 24611]

- Tokenizing
- Word forms: single, multi, fused
- Inflection
- XML: in appendix



• LAF: [ISO 24612]

- Graph-based representation of primary data and annotations
- General exchange format
- XML: GrAF

Separation of representation and content

- Terminals in SynAF are equivalent to MAF word forms.
- Annotation vocabulary should be defined by means of a data category registry (DCR), cf. ISOcat. [ISO 12620]

The SynAF metamodel

• Syntactic Nodes

- Terminal nodes for morpho-syntactically annotated word forms or for empty elements T node
- Non-terminal nodes: can be annotated with syntactic categories from the phrasal, clausal and sentential level NT node

• Syntactic Edges between both terminal and non-terminal nodes supports representation of both constituency and dependency analyses

• Annotations

- Can be applied to nodes and edges alike
- Should make use of a DCR according to [ISO 12620]

An XML serialization for SynAF

- Based on **TIGER-XML**, an existing and utilized format, [König et al. 2003] rather than 'inventing' a completely new format.
- Modified wrt TIGER-XML, to represent dependency structures, to relate annotations to a DCR, to allow for different node and edge types, etc.

TIGER-XML

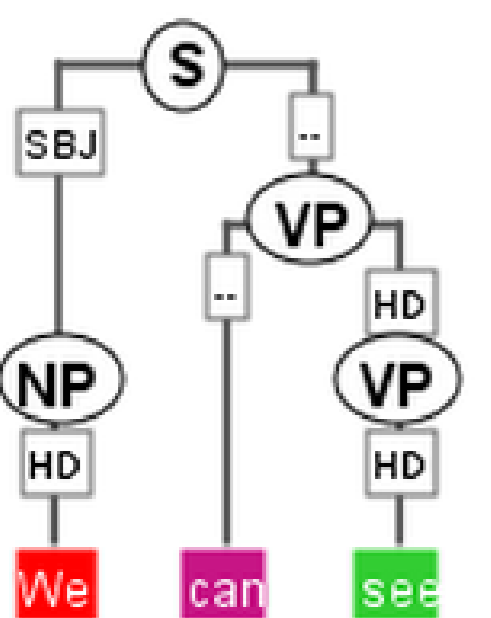
<tiger2/> [Bosch et al. 2012, Romary et al., to appear]

```
<!-- ... -->
<annotation>
  <feature name="word" domain="T"/>
  <feature name="cat" domain="NT"/>
  <value name="NP">nominal phrase</value>
  <value name="S">sentence or clause</value>
  <value name="VP">verbal phrase</value>
</feature>
<edglabel>
  <value name="SBJ">subject</value>
  <value name="HD">head</value>
  <value name="SBJ">subject</value>
  <value name="--">/>
</edglabel>
<secedglabel/>
</annotation>
</head>
```

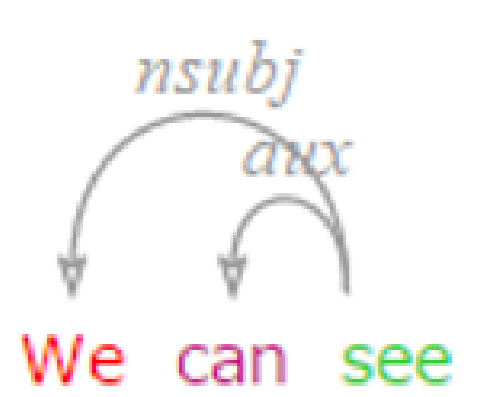
```
<!-- ... -->
<annotation>
  <feature name="cat" domain="nt"
    dcr:datcat="http://www.isocat.org/datcat/DC-1506">
    <value name="NP" dcr:datcat="http://www.isocat.org/datcat/DC-2256"/>
    <value name="S" dcr:datcat="http://www.isocat.org/datcat/DC-2295"/>
    <value name="VP" dcr:datcat="http://www.isocat.org/datcat/DC-2255"/>
  </feature>
  <feature name="label" domain="edge" type="prim"
    dcr:datcat="http://www.isocat.org/datcat/DC-5596">
    <value name="HD" dcr:datcat="http://www.isocat.org/datcat/DC-2306"/>
    <value name="SBJ" dcr:datcat="http://www.isocat.org/datcat/DC-2261"/>
    <value name="--"/>
  </feature>
  <feature name="label" domain="edge" type="dep"
    dcr:datcat="http://www.isocat.org/datcat/DC-2304">
    <value name="aux" dcr:datcat="http://www.isocat.org/datcat/DC-2262"/>
    <value name="nsubj">nominal subject</value>
  </feature>
</annotation>
</head>
```

```
<flib n="constituency features">
  <f xml:id="catNP" name="cat">
    <symbol value="NP"
      dcr:datcat="http://www.isocat.org/datcat/DC-2256"/>
  </f>
  <f xml:id="catS" name="cat">
    <symbol value="S"
      dcr:datcat="http://www.isocat.org/datcat/DC-2295"/>
  </f>
  <f xml:id="catVP" name="cat">
    <symbol value="VP"
      dcr:datcat="http://www.isocat.org/datcat/DC-2255"/>
  </f>
  <f xml:id="labelHD" name="cat">
    <symbol value="HD"
      dcr:datcat="http://www.isocat.org/datcat/DC-2306"/>
  </f>
</flib>
<flib n="dependency features">
  <!-- ... -->
</flib>
```

We can see
tiger2 (constituents)



tiger2 (dependencies)



```
<body>
  <s id="s1">
    <graph root="s1_nt4">
      <terminals>
        <t id="s1_t1" word="we"/> <!-- we -->
        <t id="s1_t2" word="can"/> <!-- can -->
        <t id="s1_t3" word="see"/> <!-- see -->
      </terminals>
      <nonterminals>
        <nt id="s1_nt1" cat="NP">
          <edge label="HD" idref="s1_t1"/>
        </nt>
        <nt id="s1_nt2" cat="VP">
          <edge label="HD" idref="s1_t3"/>
        </nt>
        <nt id="s1_nt3" cat="VP">
          <edge label="--" idref="s1_nt2"/>
          <edge label="HD" idref="s1_t2"/>
        </nt>
        <nt id="s1_nt4" cat="S">
          <edge label="SBJ" idref="s1_nt1"/>
          <edge label="--" idref="s1_nt3"/>
        </nt>
      </nonterminals>
    </graph>
  </s>
</body>
```

```
<body>
  <s xml:id="s1">
    <graph xml:id="s1_g1">
      <terminals>
        <t xml:id="s1_t1" tiger2:corresp="m1.maf#wf1"/> <!-- we -->
        <t xml:id="s1_t2" tiger2:corresp="m1.maf#wf2"/> <!-- can -->
        <t xml:id="s1_t3" tiger2:corresp="m1.maf#wf3"/> <!-- see -->
        <edge tiger2:type="dep" label="nsubj" tiger2:target="#s1_t1"/>
        <edge tiger2:type="dep" label="aux" tiger2:target="#s1_t2"/>
      </terminals>
      <nonterminals>
        <nt xml:id="s1_nt1" cat="NP">
          <edge tiger2:type="prim" label="HD" tiger2:target="#s1_t1"/>
        </nt>
        <nt xml:id="s1_nt2" cat="VP">
          <edge tiger2:type="prim" label="HD" tiger2:target="#s1_t3"/>
        </nt>
        <nt xml:id="s1_nt3" cat="VP">
          <edge tiger2:type="prim" label="--" tiger2:target="#s1_nt2"/>
          <edge tiger2:type="prim" label="HD" tiger2:target="#s1_t2"/>
        </nt>
        <nt xml:id="s1_nt4" cat="S">
          <edge tiger2:type="prim" label="SBJ" tiger2:target="#s1_nt1"/>
          <edge tiger2:type="prim" label="--" tiger2:target="#s1_nt3"/>
        </nt>
      </nonterminals>
    </graph>
  </s>
</body>
```

Proposals for ISOTiger

```
<!-- LAF reference mechanism to primary data:
      |w|e| |c|a|n| |s|e|e|
      0 1 2 3 4 5 6 7 8 9 10 -->
<!-- Only possible in cases where span is equal to word form. -->
<terminals>
  <t xml:id="s1_t1" from="0" to="2"/> <!-- we -->
  <t xml:id="s1_t2" from="3" to="6"/> <!-- can -->
  <t xml:id="s1_t3" from="7" to="10"/> <!-- see -->
</terminals>
```

```
<nt xml:id="s1_nt1"> <!-- utilizing feature library -->
  <fs feats="#catNP"/>
  <edge xml:id="s1_e3" type="prim" target="#s1_t1">
    <fs feats="#labelHD"/>
  </edge>
</nt>
```

```
<nt xml:id="s1_nt1"> <!-- utilizing feature structure -->
  <f name="cat"
    dcr:datcat="http://www.isocat.org/datcat/DC-1506">
    <symbol value="NP"
      dcr:datcat="http://www.isocat.org/datcat/DC-2256"/>
  </f>
  <edge xml:id="s1_e3" type="prim" target="#s1_t1">
    <!-- ... --> </edge>
</nt>
```

<https://github.com/laurentromary/ISOTiger>