

SynAF, ISO NWI

Thierry Declerck

DFKI

SynAF

- SynAF (Syntactic Annotation Framework) has been adopted by ISO as a NWI. For reference:
 - Candidate project number: 24615
 - Proposed project abbreviation: SynAF
 - Proposed project leader: Thierry Declerck, DIN
 - Proposed WG: ISO/TC 37/SC 4/WG 2 Representation schemes
- SynAF will be based on MAF (Morpho-Syntactic Annotation Framework) and will propose a base for future standardisation of (linguistic) semantic annotation.

Topic of SynAF

- SynAF will define a core standard for syntactic annotation of linguistic data based on the most recent practices in the field. In particular, it will design a syntactic metamodel that presents the necessary flexibility to cover both constituencies and dependencies in such annotations
- SynAF will be abstracting already existing data:
 - a) legacy data coming from existing Treebanks such as the Penn treebank and
 - b) existing grammars describing syntactic structures for various languages.
- SynAF should handle syntactic ambiguities, allow for flat and deep annotation (define shallow and deep annotation, address the topic of various layer of annotation. Should SynAF include information about operation and derivations?)

Basis for SynAF

- Corpus (Linguistic) Annotation Frameworks that combine syntactic constituency and syntactic dependency
 - Tiger for Germany
 - ITSS for Italian
 - And the same also for other family of languages (Asian etc.)
- Grammar Resources
 - Parsing output syntactic structures for various languages (HPSG, LS-GRAM Project, LFG parallel grammars, shallow grammars etc.)

Example from the Tiger corpus

- In the following example (next slide) , the feature *word* is declared as a feature of terminal nodes (T) and the feature *cat* as a feature of nonterminal nodes (NT). If a feature is used in both terminal and nonterminal nodes (e.g. *case*), its domain is called FREC. Element content of a feature value declaration is interpreted as an explanation of the feature value. Potential edge labels are declared in an <edgelabel> element, secondary edges in an <secedgelabel> element.

Example 1 from Tiger: Word

```
<head> ... <annotation>
<feature name="word" domain="T"/>
<feature name="pos" domain="T">
  <value name="ART">determiner</value>
  <value name="ADV">adverb</value>
  <value
    name="KOKOM">conjunction</value>
  <value name="NN">noun</value>
  <value name="PIAT">indefinite attributive
  pronoun</value> <value
  name="VVFIN">finite verb</value>
</feature> <feature name="morph"
  domain="T"> <value
  name="Def.Fem.Nom.Sg"/> <value
  name="Fem.Nom.Sg.*"/> <value
  name="Masc.Akk.Pl.*"/> <value
  name="3.Sg.Pres.Ind"/> <value name="--
">not bound</value> </feature>
```

- ```
<feature name="cat" domain="NT"> <value
 name="AP">adjektive phrase</value>
 <value name="AVP">adverbial
 phrase</value> <value name="NP">noun
 phrase</value> <value
 name="S">sentence</value> </feature>
<edgelabel> <value
 name="CC">comparative
 complement</value> <value
 name="CM">comparative
 conjunction</value> <value
 name="HD">head</value> <value
 name="MO">modifier</value> <value
 name="NK">noun kernel modifier</value>
 <value name="OA">accusative
 object</value> <value
 name="SB">subject</value> </edgelabel>
</annotation> </head>
```

# Example 2 from Tiger: Sentence

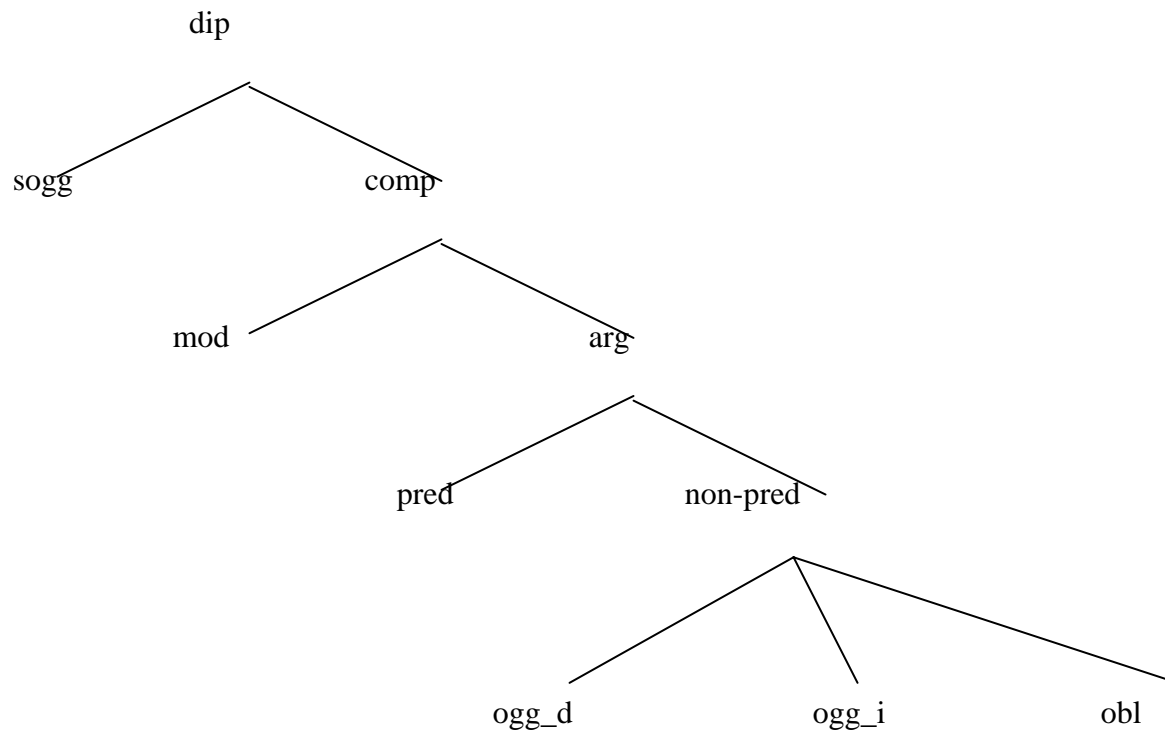
```
<s id="s5"> <graph root="s5_504">
 <terminals> <t id="s5_1" word="Die"
 pos="ART" morph="Def.Fem.Nom.Sg"/>
 <t id="s5_2" word="Tagung" pos="NN"
 morph="Fem.Nom.Sg.*"/> <t id="s5_3"
 word="hat" pos="VVFIN"
 morph="3.Sg.Pres.Ind"/> <t id="s5_4"
 word="mehr" pos="PIAT" morph="--"/> <t
 id="s5_5" word="Teilnehmer" pos="NN"
 morph="Masc.Akk.Pl.*"/> <t id="s5_6"
 word="als" pos="KOKOM" morph="--"/>
 <t id="s5_7" word="je" pos="ADV"
 morph="--"/> <t id="s5_8" word="zuvor"
 pos="ADV" morph="--"/> </terminals>
```

```
<nonterminals> <nt id="s5_500" cat="NP">
 <edge label="NK" idref="s5_1"/> <edge
 label="NK" idref="s5_2"/> </nt> <nt
 id="s5_501" cat="AVP"> <edge
 label="CM" idref="s5_6"/> <edge
 label="MO" idref="s5_7"/> <edge
 label="HD" idref="s5_8"/> </nt> <nt
 id="s5_502" cat="AP"> <edge label="HD"
 idref="s5_4"/> <edge label="CC"
 idref="s5_501"/> </nt> <nt id="s5_503"
 cat="NP"> <edge label="NK"
 idref="s5_502"/> <edge label="NK"
 idref="s5_5"/> </nt> <nt id="s5_504"
 cat="S"> <edge label="SB"
 idref="s5_500"/> <edge label="HD"
 idref="s5_3"/> <edge label="OA"
 idref="s5_503"/> </nt> </nonterminals>
</graph> </s>
```

# Benefits of SynAF

- Applications:
  - Information Extraction, Knowledge Acquisition, Translation would greatly benefit from standardized syntactic annotation
- Linguistics:
  - Interface to a future standard on semantic annotation

# The Hierarchy of Dependencies in the ITSS Corpus



# Issues for SynAF

- Level of complexity: deal only with the intersection of syntactic phenomena that are present in all (or most) languages vs. an almost complete list of phenomena describing language dependant phenomena in details.
- Closely related: monolingual description vs. multilingual descriptions. Cross-lingual aspects: for example including in the annotation information that supports translation?)
- Surface syntactic phenomena vs. „deep“ linguistic phenomena (including transformation, movement, lexical rules)
- Etc...

# Dates for SynAF

- NWIP: July 2005
- WD: 15 December 2005
- CD: 15 July 2006