SynAF: Towards a Standard for Syntactic Annotation

Thierry Declerck

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken
declerck@dfki.de

Abstract

In the paper we present the actual state of development of an international standard for syntactic annotation, called SynAF. This standard is being prepared by the Technical Committee ISO/TC 37 (Terminology and Other Language Resources), Subcommittee SC 4 (Language Resource Management), in collaboration with the European eContent Project “LIRICS” (Linguistic Infrastructure for Interoperable Resources and Systems).

1. Background

There have been in the past no thorough standardization activities in the domain of syntactic annotation, despite the numerous projects (see Abeillé & al, 2003) that have designed ways to implement linguistic TreeBanks, i.e. syntactically annotated corpora. For several years the Penn Treebank initiatives have served as a de facto standard, but more recent work (e.g. the Negra/Tiger initiatives[1] in Germany or the ISST initiative in Italy[2]) has shown that a more coherent framework could be designed to account for both (hierarchical) constituency and dependency phenomena in syntactic annotation.

Within the eContent LIRICS project, a group of international experts has started the ISO process, called SynAF (Syntactic Annotation Framework), whereas SynAF has already been accepted at the ISO Level as a New Work Item (ISO TC37-4 N204 New_work_item_proposal_SynAF).

The eContent project LIRICS (see lirics.loria.fr) is about the development of a **L**inguistic **I**nf**r**astructure for **I**nteroperable Resour**c**es and **S**ystems. More specifically the project is concerned with the enforcement or the development of ISO procedures for standards in the domain of language resources, including NLP lexica, morpho-syntactic and syntactic annotations as well as semantic content. This work is done with the purpose of enabling and ensuring the interoperability and reuse of existing and new language resources.

In this paper we will emphasize on the actual state of development of the SynAF initiative. We were starting from scratch within the ISO procedure for the establishment of a standard, supported by the international NLP community, but are in the position in building on the ISO MAF (morpho-syntactic annotation framework) initiative, which is quite advanced (see Clément & de la Clergerie, 2005).

2. Scope of SynAF

SynAF has a goal to define both a meta-model for syntactic annotations and top provide for a set of so-called data categories.

SynAF will build on the ISO MAF proposal (WD 24611, see also Clément & de la Clergerie, 2005). MAF (Morpho-Syntactic Framework) is dealing with the morpho-syntactic annotation of specific segments of textual documents. The morpho-syntactic annotation framework is about *part of speech* (noun, adjective, verb, etc.), *morphological* and *grammatical* features (such as number, gender, person, mood, verbal tense).

SynAF is about the annotation of the syntactic constituency of such (groups of) morpho-syntactically annotated fragments and the syntactic relations existing between those (groups of) morpho-syntactically annotated fragments. We consider that the sentence will define the boundaries of the fragments of textual documents to which SynAF will apply.

SynAF is dealing with the description of a meta-model for syntactic annotation, which means that SynAF will describe elementary linguistic (in fact syntactic) abstractions that support the construction and the interoperability of (syntactic) annotations and resources, as well as the procedure for the creation of data categories for syntactic annotation. SynAF will thus not propose a tagset for syntactic annotation, but is dedicated to proposing a (possibly hierarchical) list of data categories, which is much easier to update and extend, and which will represent a point of reference for particular tagsets used for the syntactic annotation of various languages, also in the context of various application scenarios.

Syntactic annotation has at least two functions in language processing:

(1) To represent linguistic constituencies, like Noun Phrases (NP), describing a structured sequence of morpho-syntactically annotated items[3], where we consider also constituents built from non-contiguous elements, and

(2) To represent dependency relations, like head-modifier relation[4]. The dependency information can exist between morpho-syntactically annotated items within a phrase (an adjective is the modifier of the head noun within an NP) or describe a specific relation between syntactic constituents at the clausal and sentential level

[1] See for futher information: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/

[2] See Montemagni (2003).

[3] Following this view, we would not deal with constituents like empty elements or traces generated by movements at the constituency level.

[4] Including also relations between same categories, like the head-head relation between nouns in appostions or nominal coordinations.

(i.e. an NP being the "subject" of the main verb of a clause or sentence).

In the first case we speak of an *internal dependency* and in the second case we speak of an *external dependency*. The dependency relation can also be stated including empty elements (like the pro-drop property in romance languages[5])

SynAF will be concerned thus with a meta-model that covers both dimensions of syntactic *constituency* and *dependency*, and SynAF will propose a multi-layered annotation framework that allows the combined and interrelated annotation of language data along both lines of consideration. Also the data-categories to be proposed to ISO standardization will be about the basic annotation concerning both dimensions.

Possible applications that might benefit from this standardization activity are information extraction, knowledge extraction from text and machine translation. Since LIRICS is also dedicating investigation work for (linguistic) semantic annotation, we assume that SynAF will be helping in defining a proper interface between syntactic and semantic annotation. At the end of the project, special attention will be given to linking linguistic annotations and semantic annotation as designed in the context of the Semantic Web initiatives.

As a starting point for SynAF we have been looking at numerous projects that have been carried out to implement TreeBanks, i.e. syntactically annotated corpora (see Abeillé 2003 and Declerck & al. 2006 for further references). This included work on many languages, like Czech, English, French, German, Italian, Japanese, Turkish etc. We also had an extended look at former European initiatives proposing guidelines for morpho-syntactic annotation for a large variety of languages, like EAGLES and Multext-East[6].

We found some approaches (e.g. the Negra/Tiger initiatives in Germany, or the ISST, Italian Semantic-Syntactic Treebank, framework for Italian) proposing coherent frameworks accounting for both (hierarchical) constituency and dependency phenomena in syntactic representation. We consider for the time being those 2 initiatives as the starting point for SynAF, which will abstract over the particular annotation strategies and tagsets proposed. In the next sections, we summarize those initiatives.

3. The Tiger Annotation Scheme

The Tiger annotation framework foresees 2 types of annotation: for constituency (represented than by a *node* label in the annotation framework) and for dependency (represented as an *edge* label in the annotation framework). This annotation strategy has reached in the meantime a kind of consensus within the corpus linguistics. We consider this to be a good basis for starting our standardization work in SynAF.

Below, two examples of the Tiger annotation framework are given. The first one shows the overall annotation strategy. There one can see that the feature *word* is declared as a feature of terminal nodes (T) and the feature *cat* as a feature of non-terminal nodes (NT), this reflecting the hierarchy of constituents. Potential edge labels are declared in an <edgelabel> element. The various “cat” values within the NT nodes of Tiger will build in SynAF the starting point for a list of data categories for constituency annotation. Within the edge label below, we can see a small list of dependency labels which will also offer a starting point (together with additional labels proposed in works dedicated to other languages) for data categories for dependency annotation.

```
<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
concjunction</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>
```

[5] This point has been particularily stressed by the authors of the ISST framework, showing here an advantage of the two-level approach, where the dependency information do not have to map entirely to the constituencey approach. In this sense, both levels of annotation have a certain independency in relation to each other.

[6] See http://www.ilc.cnr.it/EAGLES96/home.html and http://nl.ijs.si/ME respectively.

The second example below shows in a concrete example (the sentence: “Die Tagung hat mehr Teilnehmer als je zuvor” (*the conference has more participants as ever before*). the data model of Tiger and the way the annotation layers are integrated. As the reader can see, the data model of Tiger is based on *syntax graphs*, i.e. directed acyclic graphs with a single root node, whereas such graphs cannot be encoded by embedding XML elements. As a solution, all terminal and non-terminal nodes (constituency) are listed in their order of appearances n the text. ID features are co-indexing the nodes. For examples referring back from a constituent node (NT) to the annotation of words (node T) building this constituent. Edges (dependencies) are then explicitly encoded as elements, with an indexing feature pointing to the non-terminal or terminal element involved in the dependency relation to be represented.[7]

```
<body>

<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>
```

In the example above, “s” stays for *sentence*, “nt” for “non-terminal” and “t” for “terminal”. We do not go here into the details of the tagset used, and in future versions of SynAF, we will replace as far as possible the tags used in our examples with data categories proposed in both MAF and SynAF.

4. The ISST Annotation Scheme

The approach followed in the ISST (Italian Syntax Semantic Treebank) framework, is similar to the one proposed in Tiger, in the sense that annotation a multi-layered syntactic annotation strategy is proposed: One level for constituency and one level for dependency[8], with a pointing mechanism for referring from the second level to the first one. Differences can be seen in the terminology used (ISST uses the word “functional” for dependency”) and in the file organization of the XML annotations. And for sure the tagsets used are different. This is also the point where the proposition of data-categories in SynAF can help in ensuring interoperability of annotation in different syntactically annotated corpora for different languages.

In the following we show two examples of the ISST syntactic annotation, applied to the sentence: “Presentato un libro bianco del Governo Major “ (*A white book of the Governo Major has been presented*). The first example shows the constituency annotation and the second one the related dependency annotation.

```
<frase id="0" morfofile="sole.morph026"
rs="Presentato un libro bianco del Governo Major .">
  <nodo tipo="F3">
    <nodo tipo="SV3" id="0">
      <foglia lemma="presentare" href="mw_001"/>
      <nodo tipo="COMPT" id="1">
        <nodo tipo="SN" id="2">
        <foglia lemma="un" href="mw_002"/>
        <foglia lemma="libro" href="mw_003"/>
          <nodo tipo="SA" id="3">
            <foglia lemma="bianco" href="mw_004"/>
          </nodo>
          <nodo tipo="SPD" id="4">
            <foglia lemma="di" href="mw_005"/>
               <nodo tipo="SN" id="5">
              <foglia lemma="governo"
href="mw_006"/>
                 <nodo tipo="SN" id="6">
```

[7] It should be stressed here that the use of “edges” is also foreseen for representing dislocated constituents. Thus the “node” is not the only way of representing constituency, since constituency is not in all cases and all languages a strictly herarchical phenomenon.

[8] ISST proposes also a third level of annotation, but this one is reserved for semantic annotation, which is not a topic of the present document.

```
                    <foglia lemma="major"
href="mw_007"/>
                  </nodo>
                </nodo>
              </nodo>
            </nodo>
          </nodo>
        </nodo>
       <foglia lemma="." href="mw_008"/>
     </nodo>
</frase>
```

As can be seen above the ISST annotation strategy for constituency proposes a flat tree, similar on this to Tiger[9]. Nodes in a tree are also used for the representation of (most of) the constituency information. We would like to keep this as a point for the SynAF meta-model for syntactic annotation: all contiguous syntactic information can (possibly) be encoded using an embedded XML tree representation.

In the following XML representation, one can see the dependency information associated to the same sentence as above. An important feature of ISST is that it annotates word with dependency information, and not the syntactic constituents. We will have to see how to accommodate this with that approaches (like Tiger), which associate dependency mostly to constituents.

```
<frase id="0" morfofile="sole.morph026"
rs="Presentato un libro bianco del Governo Major .">
  <partec partec_id="partec_000"
lemma="presentare" modo="part_pass"
href="mw_001"/>
  <partec partec_id="partec_001" lemma="libro"
definitezza="-" href="mw_003"/>
  <partec partec_id="partec_002" lemma="bianco"
href="mw_004"/>
  <partec partec_id="partec_003" lemma="governo"
definitezza="+" introdep="di" href="mw_006"/>
  <partec partec_id="partec_004" lemma="major"
href="mw_007"/>
  <relfunz relazione_funzionale="mod"
partec1_id="partec_001" partec2_id="partec_002"
relfunz_id="r_000"/>
  <relfunz relazione_funzionale="mod"
partec1_id="partec_001" partec2_id="partec_003"
relfunz_id="r_001"/>
  <relfunz relazione_funzionale="mod"
partec1_id="partec_003" partec2_id="partec_004"
relfunz_id="r_002"/>
  <relfunz relazione_funzionale="mod"
partec1_id="partec_001" partec2_id="partec_000"
relfunz_id="r_003"/>
</frase>
```

With respect to the dependency annotation in ISST, there is furthermore a proposal for a hierarchy of dependencies, which we might take into consideration for the SynAF data-model. It is unclear to us if we will include a hierarchy for dependencies in the SynAF meta-model, but the ISST proposal has the merit of making explicit that there are different types of dependencies, as we mentioned earlier in this document (internal vs. external dependencies). On the other hand this particular hierarchy might be too language dependent.

5. Conclusions

We tend ourselves towards the Tiger and ISST frameworks as a good base for the standardization work on syntactic annotation, since the combination of constituency and dependency is probably able to cover more languages than just one of the annotation type. Nevertheless for every single language, a specific annotation scheme will have to fix which percentage of the linguistic phenomena are best described using constituency or dependency annotation. And also it is not clear yet, if dependency annotation applies to words or to constituents? Or even if they are various levels of graph annotation, as Tiger at least is suggesting: graph annotation there can also apply to dislocated constituents.

6. Acknowledgements

The standardization work presented in this paper has been partially done within the eContent project LIRICS, under the contract e-Content-22236-LIRICS. I would like to thank Gil Francopoulo (INRIA-Loria, Paris and Nancy), Monica Monachini (CNR-ILS, Pisa), Mirjam Kessler (DFKI, Saarbrücken) and Carla Parra Escartin (UPF, Barcelona). I would also like to thank various members of national standardization bodies for their comments on former versions of SynAF.

7. References

Clément, Lionel & de la Clergerie, Eric. (2005). ‘MAF’ (2005): a morphosyntactic annotation framework’. Proceedings. of the 2nd Language & Technology Conference (LT'05), Poznan (Poland), 90-94.

Abeillé A., S. Hansen-Schirra & H. Uszkoreit (eds.), 2003. Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03).

Abeillé, Anne. TREEBANKS. Building and Using Parsed Corpora (2003). Kluwer Academyic Publishers. 2003, Cordrecht, The Netherlands.

Declerck T., Kessler K., Krieger U., Monachini M., Bel N., Escartin C.P. LIRICS Deliverable 3.1 “Evaluation of initiatives for morpho-syntactic and syntactic annotation” (2006), available at lirics.loria.fr.

Montemagni S., Barsotti F. Battista, M., Calzolari N. Lenci A., Corazzari O., Zampolli A., Fanciulli F., Massetani M., Basili R., Raffaelli M., Pazienza, T. Saracino D., Zanzotto F., Pianesi F., Mana N. and Delmonte R. (2003). *Building the Italian Syntactic-Semantic Treebank*. In: Anne Abeillé (2003): *Building and Using syntactically annotated corpora*, Kluwer, Dordrecht, 189-210.

EAGLES: http://www.ilc.cnr.it/EAGLES96/home.html

Multext-East project : http://nl.ijs.si/ME

TIGER project: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/

[9] In Tiger, purely consituency relation between discontinuous elements is represented using “edges” instead of nodes.