

**Deliverable 3.4**

**Report: Test suite of ISO-conformant morph-syntactic and syntactic annotation**

Project reference number	e-Content-22236-LIRICS
Project acronym	LIRICS
Project full title	Linguistic Infrastructure for Interoperable Resource and Systems
Project contact point	Laurent Romary, INRIA-Loria 615, rue du jardin botanique BP101. 54602 Villers lès Nancy (France) romary@loria.fr
Project web site	http://lirics.loria.fr
EC project officer	Erwin Valentini
Document title	Test suite of ISO-conformant morph-syntactic and syntactic annotation
Deliverable ID	3.4
Document type	Report
Dissemination level	Confidential
Contractual date of delivery	M30
Actual date of delivery	30.08.2007
Status & version	Final
Work package, task & deliverable responsible	DFKI
Author(s) & affiliation(s)	Thierry Declerck, Tania Avgustionva and Valia Kordoni (DFKI), Adam Funk and Kalina Bontcheva (Sheffield)
Additional contributor(s)	
Keywords	Syntax, Annotation, Standards, Tree-Banks

**Document evolution**

Version	date	version	date
0.1	30.6.2007		
1.0	30.8.2007		
2.0	15.10.2007		

**Introduction**

One of the aims of LIRICS is the development of test suites, i.e. a set of resources in the form of practical examples associated to the international standards under development or finalized by the project, in order to test the applicability and usability of the proposed concepts. The objectives of developing test suites in conjunction with the delivery of a standard are to provide a guide for those who wish to apply them to their resources and, above all, to test their viability in NLP implementations and systems.

Test suites should accompany the standards, ensure both wide dissemination and demonstration, and support their implementation and capability of propagation during and after project life cycle. Finally, the development of test suites will allow implementers to combine a given standard proposal in the form of a meta-model with the relevant Data Categories taken from the Data Category Registry. They can thus be used as examples of the application of data categories themselves, for decorating the annotation meta-models with tagsets. Test suites also act as a reference to the best practices in the representation of those phenomena.

Since the meta-models for morpho-syntactic (MAF) and syntactic (SynAF) annotations are still in a development phase, we present in this document the actual state of text suites, which are not yet proposing a full integration of MAF and SynAF annotation.

We are proposing here test suites for MAF and SynAF for six languages: Bulgarian, English and French on the one side, and German, Italian and Spanish on the other side. This distinction among the two groups of languages is for the time being motivated by two aspects:

- University of Sheffield was in charge of the three first mentioned languages and DFKI for the other 3 languages.
- Sheffield took the very first proposal for syntactic annotation we had in WP3, and DFKI in the last months tried to anticipate the prefinal version of SynAF, which we expected to be submitted at the end of the year 2007.

The two comments above stress again that we are not delivering a final set of test suites, but it is nevertheless useful to do this exercise, since in doing so we tested the capability of two systems to provide for an annotation format generated by an institution (ISO) external to the research institutions themselves. The authors of this deliverable will for sure update the test suites according to the next developments of the standards, until they have reached their final state.

There is another difference between the test suites for SynAF provided by Sheffield and DFKI. Where Sheffield concentrated on existing validated corpora, DFKI implemented the generation of test suites immediately in the processing tools, so that there are certain mistakes in the annotation yet. Another difference: the corpus of Sheffield didn't have annotation for dependency, and in the actual version of the test suites for SynAF, information about dependency is missing in the three languages Bulgarian, English and French.

Here again, when the standards MAF and SynAF will about to be published, we will have in both cases a dedicated set of test suites. In the next version of the test suites we will also consider the implementation of the multi-layer linguistic annotation scheme defined in the more generic Linguistic Annotation Framework (LAF), being developed within ISO TC 37/SC4, and we will integrate MAF and SynAF in this very final version.

A last comment: in the MAF test suites, the reader can see how the relation is established with the data categories. This information is missing in the test suites for syntactic annotation, since the data categories tool hosted till now by Loria is being re-implemented at MPI, and the candidate data categories for SynAF could not be integrated till now.

We attached all the files containing the test suite, to improve readability of the XML code, using your favorite browser.

**Content:**

<b>1</b>	<b>TEST SUITES FOR MAF</b>	<b>5</b>
1.1	INTRODUCTION	5
1.2	BULGARIAN	5
1.3	ENGLISH	5
1.4	FRENCH	5
1.5	GERMAN	5
1.6	ITALIAN	5
1.7	SPANISH	5
<b>2</b>	<b>TEST SUITES FOR SYNAF</b>	<b>6</b>
2.1	INTRODUCTION	6
2.2	BULGARIAN	6
2.3	ENGLISH	6
2.4	FRENCH	6
2.5	GERMAN	6
2.6	ITALIAN	6
2.7	SPANISH	6

## 1 Test Suites for MAF

### 1.1 Introduction

In the following XML files (to be visualized with a browser supporting XML), the reader can find two main blocks of information: the one within the "tagset" element, the list of tokens themselves (the result of the MAF segmentation), a list of feature value (fv) descriptions and the libraries containing them, and then the morpho-syntactic information properly encoded as feature structure matrixes (therefore the naming of the XML element "fsm").

The information encoded in the "tagset" element reflects the local tagset of the annotation engine (or the annotation schema used by manual annotators), and how it relates to the data categories available for MAF, within the "dcs" element ("dcs" standing for "data category selection"). An example:

```
<dcs local="posspron" registered="urn:dcr:morphosyntax:possessivePronoun" rel="eq"/>
```

Here the reader can see how the tag "posspron" is set in equality relation with the registered data category "possessivePronoun".

In the German MAF test suite, you will see the use of another local tag set, in fact a collection of digits that are also put in relation with the MAF data categories, using the "dcs" element:

```
<dcs local="2" registered="urn:dcr:morphosyntax:verb" rel="eq"/>
```

Following the DCS listing you may find some information about the libraries, where the Feature Values (fv) for the morpho-syntactic information is encoded. In this the reader can see some of the types associated with feature structures.

The tokens are listed, being the results of the segmentation work, and bearing no further information. A token is a unit having a starting point and an ending point. In the XML element "token", there is a feature called "joint", which just expresses the possible "glue" between the actual token and a preceding or following token.

The list type of information to be found in the annotation, is the morpho-syntactic information attached to the token. The tags used can be part of a local tagset, referring back then to the data category, within the "dcs" elements (see just above).

### 1.2 Bulgarian

See attached file [maf-bg-stojan](#)

### 1.3 English

See attached file [maf-en-poe](#)

### 1.4 French

See attached files [maf-fr-baudelaire](#) and [maf-fr-verne](#)

### 1.5 German

See attached file [de-MAF-out](#)

### 1.6 Italian

See attached file [it-MAF-out](#)

### 1.7 Spanish

See attached file [es-MAF-out](#)

## 2 Test Suites for SynAF

### 2.1 Introduction

We describe in this introduction mainly the annotation files provided for the languages German, Italian and Spanish, since in those cases we have information about syntactic dependency. Also, as we stressed above already, the second group of languages is for the time being annotated with a first version of SynAF, and this annotation will be updated for the

The annotation files start with the element "head", which contains a listing of the terminal nodes (T), such as they are described for the time being in the provisional data categories for SynAF. We specify in the very first lines of the XML code, that the part-of-speech (POS) and morphological information (basically the wordForm annotation resulting from MAF), are the domain of annotation for the Terminal Nodes.

Following we introduce the labels for the non-terminal nodes. The set of terminal and non-terminal nodes are delivering what we consider to be the syntactic information relevant for describing the syntactic constituency part of SynAF.

Following the introduction of the values associated with labels for the nodes, we list the labels for the edges, which are introduced in SynAF (but also in LAF) for describing dependency relations (like head-modifier relation, or Subject-Predicate relations). This is closing the "head" part of the annotation file.

In the "body" part of the annotation, the actual SynAF meta-model can be seen as a XML encoded graph representation. We have two levels: the non-terminal nodes and the terminal nodes. At the terminal level, we have the information about the POS and the Lemma, inherited from MAF<sup>1</sup>.

Within the non-terminal annotations, we include the information about the syntactic nodes (constituency) and the grammatical relations (dependency), giving a label for them and specifying their range (the starting points and the ending points for constituents) and in case of the dependency relations, the point in the node annotation which is getting a dependency label).

### 2.2 Bulgarian

See attached file [synaf-bg-stojan](#)

### 2.3 English

See attached file [synaf-en-poe](#)

### 2.4 French

See attached files [synaf-fr-baudelaire](#) and [synaf-fr-verne](#)

### 2.5 German

See attached file [de-SynAF-out](#)

### 2.6 Italian

See attached file [it-SynAF-out](#)

### 2.7 Spanish

See attached file [es-SynAF-out](#)

<sup>1</sup> Whereas due to a bug, the POS information is displayed for the time being only in the annotation files for Bulgarian, English and French.