LIRICS

Deliverable 3.2A

Interim Report: WD of morpho-syntatic annotation standard for CD ballot

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	Interim Report: WD of morpho-syntatic annotation standard for CD ballot
Deliverable ID	3.2A
Document type	Report
Dissemination level	Confidential
Contractual date of delivery	M18
Actual date of delivery	21.06.2006
Status & version	Final
Work package, task & deliverable responsible	DFKI
Author(s) & affiliation(s)	Thierry Declerck, Mirjam Kessler, Ulrich Krieger and Bernd Kiefer (DFKI)
Additional contributor(s)	Gil Francopoulo (LORIA), Eric de la Clergerie (INRIA, Editor of the MAF document), and many experts from national standardisation bodies and ISO.
Keywords	Morpho-Syntax, Annotation, Standards, Tree-Banks

Document evolution

Version	date	version	date
1.0	1st Mai 2006	1..3 (final)	18 July 2006
1.1	15th Juni 2006		
1.2	30th Juni 2006		

Introduction

We present in this deliverable the version of MAF (Morpho-Syntactic Annotation Framework) that has been submitted for an ISO CD ballot. This document, reproduced here in section 2, is not an ISO International Standard, and is not even the exact document that was submitted as a CD, but has been edited for Word Text Processing by the WP3 leader, excluding all the ISO formal aspects. It is distributed **exclusively** to the LIRICS partners and to the European Commission as an intermediate report. This deliverable has to be considered as **strictly confidential** and for consortium internal purposes only.

Various members of the LIRICS have been involved as well in commenting former versions of the CD proposal of MAF reproduced in section 2. Also various members of national standardization bodies have been involved in this activity. We cannot name all of them here, but we would like to thank especially the members of LIRICS for their input on the former versions of MAF. And many thanks to Gil Francopoulo (LORIA), who as coordinator of LIRICS was very effective in coordinating the exchange of information and the discussion between the editors of MAF (actually Eric de la Clergerie, INRIA) and the relevant WP of LIRICS.

A last editorial note: The deliverable has some days delay with respect to the foreseen date (month 18). This is due to the fact that the ballot took place recently and that reactions from national bodies were about to be available, which we wanted to partially include in this document, for informing the LIRICS partners about those reactions and comments. The general tenor was positive, concerning the model, but negative concerning some linguistic issues, which were considered as not being enough elaborated and there was the request to revise a certain number of definitions.
So a new version of MAF has to be produced in the next months. We list at the end of the document a selection of comments, in order to inform the partners of LIRICS. Some of the comments are redundant, since they are coming from various standardization bodies. The comments are given here without mentioning from which standardization body they originate. Also those comments are **strictly confidential** to the LIRICS partners, the Project Officer and the Reviewers.

Another reason for delaying a bit the deliverable is the fact that ISO TC37/SC4 organized a meeting in Paris on the 5th and 6th of July 2006, for discussing the results of the ballot and for taking decisions. Thierry Declerck was representing LIRICS at this meeting, where also the editor of MAF, Eric de la Clergerie (INRIA) and Prof. Kiyong Lee as the convenor of the WG2 of ISO TC37/SC4 were among the participants.

In the next version of this deliverable, D.3.2.b, we will report on the updates proposed to MAF for reaching the DIS level. Additionally to the Meta-Model, for morpho-syntactic annotation, we will also provide for the actual list of morpho-syntactic data categories as developed within LIRCS, in close collaboration with the editorial board of MAF.

Content:

1 Morpho-syntactic Annotation Framework (MAF), as submitted to ISO for CD ballot

As mentioned in section 1, the reproduction of the MAF document below is selective and concentrates on the content aspects of MAF. Also the official MAF document was not generated in "Word" in so a lot of editing has been necessary. We would like to remind that this section of this deliverable is to be considered as **confidential** by all the members of LIRICS, as well as by the Project Officer and the reviewers of the project.

2 Foreword

ISO (The International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and nongovernmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3. Draft I International Standard 24611 was prepared by Technical Committee ISO/TC 37, Terminology and Other Language Resources, Subcommittee SC 4, Language Resource Management. All the Annexes are for information only.

3 Scope

In Natural Language Resource Management, the morpho-syntactic annotation phase assigns to each document segment (either text or speech) one or more tags providing morphosyntactic information about the part of speech (noun, adjective, verb, etc.), morphological and grammatical features (such as number, gender, person, mood, verbal tense, . . .) and possibly other specific linguistic properties. The morpho-syntactic annotations attached to a segment do not refer to other segments or annotations, even if the choice of an annotation may depend on the surrounding context.

4 Normative references

- ISO 8879: 1986 (SGML) as extended by TC2 (ISO/IEC JTC 1/SC 34 N029: 1998-12-06) to allow for XML
- ISO 19757-2, Document Schema Definition Language, part 2, to allow for RELAX NG specifications. RELAX NG is a schema language for XML, standing for Regular Language for XML for Next Generation, and simplifies and extends the features of DTDs, Document Type Definitions.
- ISO 12620 on Data Category Registry (DCR)
- ISO 24610-1 on Feature Structure Representation (FSR)
- ISO 24610-2 on Feature System Declaration (FSD)
- ISO 24612 on Linguistic Annotation Framework (LAF)
- Text Encoding Initiative (TEI) – Chapters to be defined

5 Terms and definitions

For the purpose of ISO 24611, the following terms and definitions apply.

Associative relation: relation by which a linguistic unit is associated with other units. It is a virtual association, which does not requires their effective presence and differs from a paradigmatic relation in that the latter only refers to linguistic units associated by substitutability.

Closed data category: data category whose content is constrained by a list of permissible values, which comprise its conceptual domain.

NOTE: A typical closed data category might be /grammatical number/, which can have as its content the values: /singular/, /plural/ or /dual/.

Conceptual domain: finite list of simple data categories that may be the values of a complex data category

Data category: result of the specification of a given data field or the content of a closed data field.
NOTE: A data category is to be used as an elementary descriptor in a linguistic structure or an annotation scheme. Examples are: /term/, /definition/, /part of speech/ and /grammatical gender/. Data categories for the management of lexical resources and terminology are comparable to data element concepts in ISO/IEC 11179-3:2003.

Directed acyclic graph (DAG): graph with directed edges and no cycle

Discourse:

Feature specification: the assignment of a value to a feature. In MAF, a feature shall denote a morpho-syntactic feature of a linguistic unit, such as the mood or tense of a verb.

Feature structure: a set of feature specifications, used in MAF to express morpho-syntactic content.

*Finite state automata (*FSA): finite set of transitions from state to state, with an initial state and a final one. See also DAG.

Form: any sequence of letters, pictograms and numerals used to write or pronounce a word

Grammatical category: See also part of speech

Inflection: modification or marking of a so that it reflects grammatical (i.e. relational) information, such as grammatical gender, tense, person, etc.

Inflection paradigm: a table illustrating the forms of an inflected word.

Inflected form: form that a word can take when used in a sentence or a phrase. NOTE: An inflected form of a word is associated with a combination of morphological features, such as grammatical number or case.

Lattice: term often used in the NLP community to denote (with some slight confusion with the notion of algebraic lattice), a directed acyclic graph with an initial node and a final node. See also DAG and FSA.

Lemma – lemmatized form: class of inflected forms differing only by inflectional morphology. A lemma is usually referred to by one of these forms, arbitrarily chosen (e.g., infinitive for French verbs).

Lexeme: lexical morpheme to be distinguished from a grammatical morpheme by the fact that it belongs to an open list and that it bears an autonomous signification.

Lexicon: resource comprising a collection of inflected forms or lemmas for a given language

Morpheme: smallest linguistic unit bearing a signification in a discourse and that cannot be divided into smaller meaningful units. A morpheme is either grammatical (grammeme) or lexical (lexeme).

Morphological feature/morpho-syntactic feature: category induced from the inflected form of a word. NOTE: ISO 12620 provides a comprehensive list of values for European languages. An example of a morphological feature is: /grammatical gender/.

Morphology of a word/morpho-syntax of a word: description comprising the lemmatized form or forms of a word, plus additional information on its /part of speech/ data categories, possibly its inflectional paradigm or paradigms, and possibly its explicitly listed inflected forms.
NOTE: The term morpho-syntax is often used in place of morphology as it describes such features as number, gender, case etc. which are essential for syntactic agreement.

Morpho-syntactic tag: to an associative relation corresponds a feature, for which the related entities share the same value. The morpho-syntactic tag lists some of these features (part-of-speech, grammatical category, etc.).

Multi-word expression (MWE): an expression composed of an ordered group of words that has properties that are not predictable from the properties of the individual words or of their normal mode of combination.
NOTE: The group of words making up an MWE can be continuous or discontinuous. EXAMPLE: "father in law" or "to be over the moon" that mean something different from what they appear to mean.

Natural language processing (NLP): the field of study covering knowledge and techniques which allow computerized processing of linguistic data. This field combines a variety of skills including linguistics, mathematical logic, statistics, and algorithms.

Open data category: data category whose content cannot be fully *enumerated* due to the organic nature of language.

EXAMPLE: Typical open data categories might include /term/, /lemma/.

Part of speech: category assigned to a word based on its grammatical and semantic properties. See also grammatical category.

NOTE: ISO 12620 provides a comprehensive list of values for European languages. Examples of such values are: /noun/ and /verb/.
Syntagmatic relation: relation by which linguistic units in a discourse are associated.

Token: non-empty contiguous discourse sequence identified as such by a morpho-phonological analysis or an automatic processing of the discourse. This can involve the recognition of a regular or algebraic language (matching of the separators), or a lexicological analysis (recognition of roots, morphological derivation and inflection, etc.).

Tokenization: the process of identifying tokens

Word-form/morpho-syntactic unit: contiguous or non-contiguous entity from a speech or text sequence identified as such in an associative relation. This identification is the basis of morpho-syntactic tagging (part-of-speech, grammatical category, agreement feature, etc.). Morpho-syntactic units may have no acoustic or graphic realization, or correspond to one or more tokens.

Romanization: transliteration from a non-Latin script into a Latin script

Script: set of graphic characters used for the written form of one or more languages (ISO/IEC 10646-1, 4.14)

Simple data category: data category that may be the possible content of a closed data category, but that cannot itself be further sub-divided

EXAMPLE: /masculine/, /feminine/, and /neuter/ are possible simple data categories associated with the conceptual domain of the closed data category /grammatical gender/ as it is associated with the German language.

Transcription: form resulting from a coherent method of writing down speech sounds

Transliteration: form resulting from the conversion of one writing system into another

Word: in the context of a given language, is a description composed of at least a part of speech and a lemmatized form.

NOTE: The description can include more morphological information and/or syntactic and semantic information. A word is either a single word or a multi-word expression.

Word class: See also part of speech.

6 Key standards used by MAF

6.1 ISO 12620 Data Category Registry (DCR)

The designers of any specific MAF tagset shall use data categories from the ISO 12620 DCR. The DCR is a set of data category specifications defined by ISO 12620 and maintained as a global resource by ISO TC 37 in compliance with ISO/IEC 11179-3:2003. Tagset creators can define a set of new data categories to cover data category concepts that are needed and that are not currently available in the DCR. The tagset creators shall be responsible for negotiating the addition of the new data categories to the DCR. This supplemental set of data categories shall be represented and managed in conformance with ISO 12620.

6.2 ISO 24610 Feature Structures (FSR and FSD)

Morpho-syntactic content shall be expressed using the ISO 24610 part 1 document on Feature Structure Representation and validated using the future ISO 24610 part 2 companion document on Feature System Declaration.

6.3 OLAC Metadata

Metadata for MAF shall be expressed following the recommendations and categories proposed by the Open Language Archives Community (OLAC), as described in the latest version of OLAC Metadata Standard (http://www.language-archives.org/OLAC/metadata.html). The OLAC Metadata Set includes the Dublin Core Metadata Set with qualifiers.

6.4 Unified Modeling Language (UML)

MAF complies with the specifications and modeling principles of UML as defined by OMG. MAF uses the subset of UML that is described in Annex E.12

7 General characteristics of MAF

7.1 Overview

In the Linguistic Community, morpho-syntactic annotations provide an important layer of linguistic information in a document, even if they do not cover the full range of possible linguistic annotations. Other kinds of annotation on references, discourse, prosody, or parsing may complete morpho-syntactic annotations.

Syntax and semantics can not be avoided in the definition of parts of speech and of grammatical categories. For instance, pronouns and substantives intrinsically carry a reference to some entity; the tense or the aspect of verbs indicate the temporal deixis; the person, modality and other grammatical categories indicate the enunciation context, ...

Therefore, it is not easy to provide an exact and precise definition of what morpho-syntactic annotations cover because they are strongly related to many other linguistic properties of a given language in a given context.

Nevertheless, the present proposal tries to delimit minimal and maximal sequences in documents (either text or speech) that can be identified as morpho-syntactic units and tries to categorize the linguistic properties that may be used to mark these units, within some larger syntagmatic context. Minimal units can not be broken into sub-parts that could be identified by similar morpho-syntactic criteria, but may however still be broken into smaller units with morphological or phonological properties. Morpho-syntactic units can be nested to form maximal units (such as compound words) that act as elementary units for other level of linguistic analysis, particularly parsing. The exact boundary between morpho-syntax and parsing is sometimes difficult to define.

7.2 MAF Meta-Model

Figure 1 presents a simplified view of the proposed meta-model for morpho-syntactic annotations, while Figure 2 presents a more formal view based on UML. An annotated document is formed by a raw original document and a set of annotations. The annotations are carried by word forms covering zero, one or more segments or tokens of the original document. A word form may reference a lexicon entry and provides information about its underlying lemma and inflected form. The morpho-syntactic content attached to a word form is expressed by feature structures following the guidelines of one or more tagsets. The terminology or set of categories used in tagsets are described w.r.t. registered data categories. Because of structural ambiguities, both tokens and word forms are organized into one or more flows, materialized by lattices, or more formally by Directed Acyclic Graphs [DAGs]. The current proposal addresses the representation of segments (through tokens), word forms, morpho-syntactic content, tagsets, and ambiguity. A MAF model is instantiated from the MAF meta-model through the selection of a set of data categories.

7.3 Segmenting with tokens

Morpho-syntactic annotations are carried by segments, called tokens, present in the document flow, but this does not imply that the resulting segmentation corresponds to a sequence of adjacent segments partitioning the original document. It is particularly important to distinguish the morpho-syntactic units from their realizations. Some parts of a document may

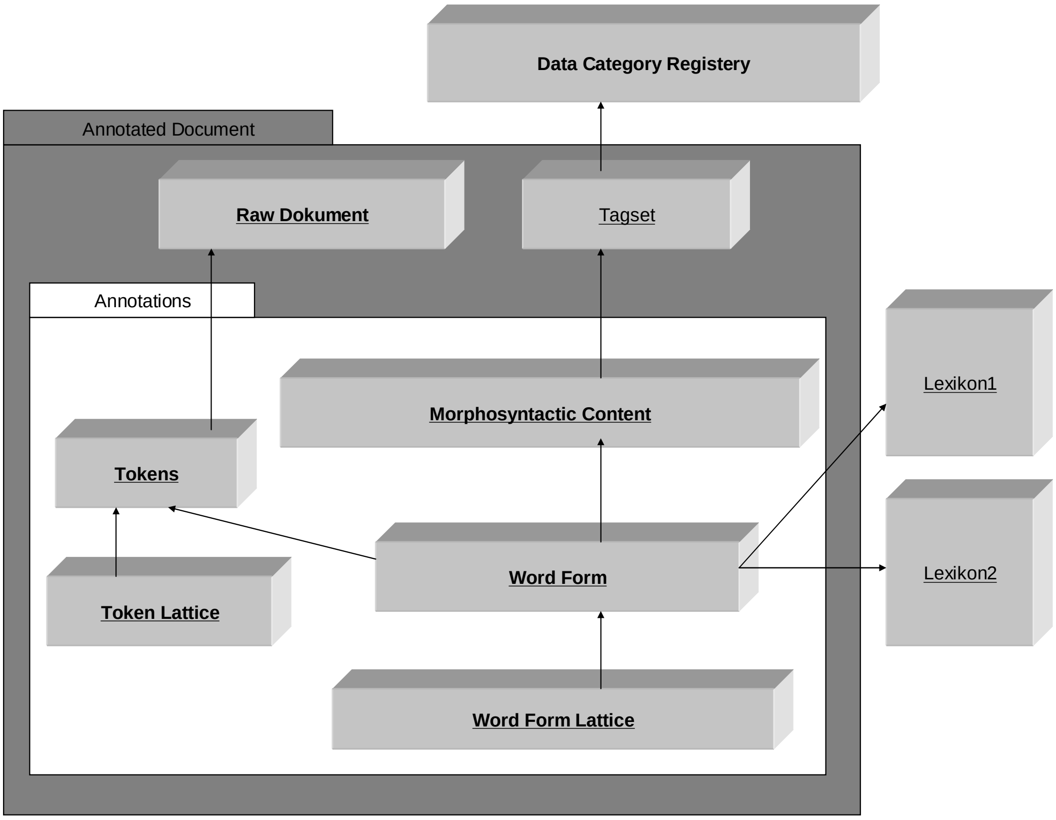

Figure 1: Simplified view of MAF meta-model

carry no annotations (typographic marks, didascalies, markup elements, . . .); other parts may not exactly correspond to their segmented form (abbreviations, brachygraphies, typographic errors and variations, typographic and morphological contractions, . . .). Also, a morpho-syntactic unit may not correspond to a segment identified by typographic marks (such as white spaces or hyphens), for instance for German compound words, speech transcription, or Sanskrit writing.

The element token is used to represent these segments of the original document that, roughly speaking, follow typographical, morphological, or phonological boundaries. The current proposal does not define the linguistic properties of tokens. In different languages, a token may be identified through typographic properties (white-space, hyphens, characters,. . .) and/or morphological properties (radical, affix, morpheme, . . .). The description of the morphological, phonological or lexicological structures that may define a token is not covered by the current proposal.

Other typographical marks used to format pages or to separate words and paragraphs, as well as encoding information, do not belong to morpho-syntactic annotations and are also not covered by this proposal, but rather by TEI.

7.4 Standoff notation

The element token provides an independence from the original document by providing a way to reference intervals in documents. The attributes from and to are used to define such intervals. The content of these attributes depends on some chosen addressing schema to denote non ambiguous document positions and depends on the nature of the original document.

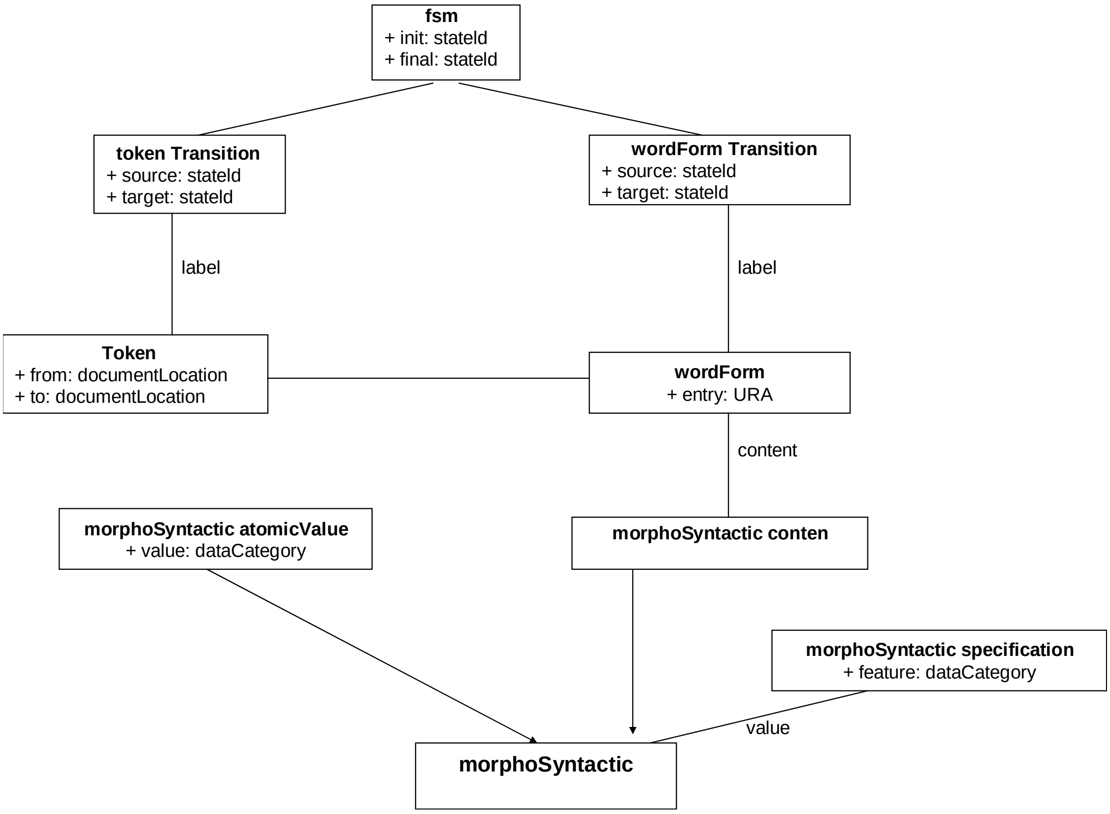

Figure 2: UML view of MAF meta-model

7.5 Embedding notation

It is not always necessary to separate the original document from its annotations. For simple cases, textual content may be directly embedded within token.

```
<token id=" t1 ">The</token>
<token id=" t2 ">victim</token>
<token id=" t3 ">’ s</token>
<token id=" t4 ">friends</token>
<token id=" t5 ">told</token>
<token id=" t6 ">police</token>
<token id=" t7 ">that</token>
<token id=" t8 ">Krueger</token>
<token id=" t9 ">drove</token>
<token id=" t10 ">into</token>
<token id=" t11 ">the</token>
<token id=" t12 ">quarry</token>
<token id=" t13 ">and</token>
<token id=" t14 ">never</token>
<token id=" t15 ">surfaced</token>
<token id=" t16 ">.</token>
```

The embedding notation will be used for most of the provided examples for MAF but it should be noted that the use of this notation is not recommended. A first reason is that the morpho-syntactic annotations may conflict with other annotations. A second reason is that the content of the textual material separating the textual content embedded within token is not precisely defined (white-space, newlines, no space, hyphen, . . .), except by relying on attribute join.

7.6 Informative attributes

Tokens address segments of the original document but also provide a level of possible abstraction w.r.t. this document, for instance w.r.t. graphical or phonological variations that are not linguistically pertinent. The non mandatory attributes form, transcription, transliteration may be used to perform this abstraction, providing, for instance, the phonetic transcription of a speech segment, the roman transliteration of some Cyrillic word, the expansion of an abbreviation, the correction of a typographical error, or the choice of a normalized form in presence of variations:

```
<token form=" etcetera " id=" t1 ">etc .</token>
<token form=" tzar " id=" t2 ">csar</token>
<token form=" tzar " id=" t3 ">tsar</token>
<token form=" 23/02/03 " id=" t4 ">February, 23 rd 2003</token>
<token form=" etcetera " phonetic="/ etsettr@/" from=" 1251"     to=" 1253" id=" t5 "/>
<token phonetic="/ platto /" id=" t6 ">plateau</ token>
```

The abstraction provided by the attribute form is also adequate to handle the phenomena of contraction and agglutination where two tokens may cover the same segment of the original document for distinct values (see Section 6.4.2).

7.7 Completing the embedding token notation

As above mentioned, the embedding token notation is less precise than the standoff one, in particular to explicit the contiguity and the overlapping of tokens (which are obvious to check using the document positions in the case of the standoff notation).

7.7.1 Joining tokens

The embedding notation for tokens is completed by the attribute join used to specify how a token is joined with its sibling tokens. By default, two sibling tokens are considered to be separated by whatever separator is standard for the document language (for instance, space separated for many languages). By using the attribute join, it is possible to indicate that a token is contiguous with its left or right sibling or with both.

```
<!–– it  is  said . . . ––>
<token id=" t1 ">L’</ token>
<token id=" t2 " join=" left ">on</ token>
<token id=" t3 ">dit</ token>
```

It should be noted that a token may enclose material usually considered as separator, such as spaces, newline, dash, apostrophe, . . . , even if these tokens do not anchor linguistic units at the level of word forms.

```
<!–– it  is  said . . . ––>
<token id=" t1 ">L</token>
<token id=" t2 " join="both">’</token>
<token id=" t3 ">on</token>
<token id=" t4 ">dit</token>
```

Another example, in Modern Greek, is provided by the idiomatic gajäexpression “kaloka- ” (good and brave) that may be segmented in three agglutinated segments “ä ”,“ki”, and “agjä” and represented by:

```
<token form="kalä
" id=" t0 ">kalo</ token>
<token form="i" id=" t1 " join=" left ">k</ token>
<token form="agjä
" id=" t2 " join=" left ">agjä
</ token>
```

7.7.2 Overlapping tokens

Two tokens may overlap, for instance to denote an agglutinated or contracted form (for instance, in French, “des” may be seen as a contraction for “de les” [of the]), or to denote multi-locutor documents with overlapping discourses. In these cases, a token may not mark just the realization of a typographical or vocal sequence, but expresses a deeper linguistic reality pertinent for segmenting a document. It is however still possible not to mention overlapping at the level of tokens and to postpone the issue at the level of linguistic units , i.e. word forms.

The value overlap for the token attribute join may be used to denote overlapping at the level of embedding tokens. For instance, the following example illustrates the contraction of an abbreviation with a punctuation mark for “etc.”, for the standoff and embedding notations for element token:

a Standoff notation

```
<token form=" etcetera " id=" t1 " from="p1" to="p3"/>
<token form="#dot#" id=" t2 " from="p1" to="p3"/>
```

b Embedding notation

```
<token form=" et cetera " id=" t1 ">e t c .</token>
```

```
<token form="#dot#" id=" t2 " join=" over lap "/>
```

7.8 Formal description: token

```
\ token -
        element token {
                attribute id { xsd : ID }? ,
                token . information ,
                (
                        (
                        attribute from { DocumentLocation } ,
                        attribute  to { DocumentLocation }
                        )
                | ## DTD => ,
                        (
                        [ a : defaultVa lue = "no" ]
                        attribute join { "no" | " left " | " right " | "both" |
                        "over lap " }
                                ? ,
                        text
                        )
                )
        }
token.information &= attribute form { string }?
token.information &= attribute phonetic { string }?
token.information &= attribute transcription { string }?
token.information &= attribute transcription { string }?
```

8 Word Forms as linguistic units

The segments identified by token elements are used to anchor word forms, that may generally be associated, through attribute entry, to a lexical entry in a lexicon. Words forms are also characterized by a part of speech as well as morphological and grammatical properties expressed by feature structures (see Section 8.1). Immediate information about the lemma and inflected forms may also be attached with the attributes lemma and form. In particular, the attribute form is useful when the inflected form attached to the word form does not coincide with the content attached to the covered tokens, because, for instance, of spelling corrections.

A token may be associated to more than one word form and, conversely, a word form may cover more than one token.

For instance, in French, the morphological agglutination of auquel ("of which") may have several representations, depending on the granularity of the tokenization:

coarse granularity The character sequence *auquel* is not decomposed and is covered by a single token, with two word forms covering this segment.

```
        token id=" t0 ">auquel</ token>
```

```
wordForm lemma="à" tag="pos.prep" tokens=" t0 "/>
wordForm lemma=" lequel " tag="pos.pronrel " tokens=" t0 "/>
```

fine granularity The tokenizer identifies two agglutinated parts materialized by two tokens, each of them anchoring a word form:

```
<token form="a" id=" t0 ">auquel</ token>
<token form=" lequel  " id=" t1 " j o i n=" overlap "/>
<wordForm lemma="à" tag="pos.prep" tokens=" t0 "/>
<wordForm lemma=" lequel " tag="pos.pronrel " tokens=" t1 "/>
```

The choice of a level of granularity can be motivated by the usage or by the available tools for a given language.
As mentioned before, there are no mandatory linguistic properties for defining the tokens, which can, for instance, be automatically recognized by regular languages. On the other hand, a word form, that may cover zero, one or more tokens, should represent a linguistic unit carrying morpho-syntactic information.
The current proposal does not discuss the linguistic choices that define these linguistic units but provides enough flexibility to annotate them. The choice may be motivated by lexical or morphological properties based on context and language (depending on the nature and function of words).

8.1 Token attachment

8.1.1 One token; one word form

The simplest case of relationship between tokens and word forms is when a word form covers a single token.

```
<token id=" t0 ">apple</ token>
<wordForm lemma=" apple " tokens=" t0 "/>
```

8.1.2 Several contiguous tokens; one word form

However, the current proposal allows the handling of more complex cases, as the identification of compound words covering several adjacent tokens:

```
<token id=" t0 ">prime</ token>
<token id=" t1 ">minister</ token>
<wordForm lemma=" prime_minister" tokens=" t0 t1 "/>
```

8.1.3 Several discontinuous tokens; one word form

A sequence of non contiguous tokens may also be attached to a word form, for instance to handle cases where some material is inserted inside the components of a word form:

```
<token id=" t1 ">afin</token>
<token id=" t2 ">justement</token>
<token id=" t3 ">de</token>
<wordForm lemma=" afin_de " tokens=" t1 t3 "/>
<wordForm lemma=" justement " tokens=" t2 "/>
```

This kind of phenomena may also occur for verbs with detached particles, for instance in English or German. The English infinitive verbal form “to <verb>” may also fit in this scheme.

```
<token id=" t1 ">to</ token>
<token id=" t2 ">eventually</ token>
<token id=" t3 ">decide</ token>
<wordForm lemma=" to_decide " tokens=" t1 t3 "/>
<wordForm lemma=" eventually " tokens=" t2 "/>
```

In order to identify discontinuous word-form while preserving some information about the position of each component in the flow of word forms, one may use word forms covering the same sequence tokens and referring to the same entry (but possibly sub-entries).

```
<token id=" t1 ">to</ token>
<token id=" t2 ">eventually</ token>
<token id=" t3 ">decide</ token>
<wordForm entry=" urn : lexicon : en : decide : to " tokens="        t1 t3 "/>
<wordForm entry=" urn : lexicon : en : eventually " tokens="        t2 "/>
<wordForm entry=" urn : lexicon : en : decide : main"
tokens=" t1 t3 "/>
```

8.1.4 Zero token; one word form

Another case that may arise is when one wishes to insert a word form which is not realized in the original document, and is, therefore, associated with an empty sequence of tokens, e.g., some pronouns in Spanish or the hypothesis of traces.

```
<token id=" t1 ">Jean</ token>
<token id=" t2 ">propose</ token>
<token id=" t3 ">de</ token>
<token id=" t4 ">partir</ token>
```

```
<wordForm lemma="Jean " tokens=" t1 "/>
<wordForm lemma=" proposer " tokens=" t2 "/>
<wordForm lemma="de" tokens=" t3 "/>
<wordForm lemma="PRO" tokens=""/>
<wordForm lemma=" partir " tokens=" t4 "/>
```

Even if a word form covers no tokens, it still has a relative position w.r.t. the other word forms. It is this relative position which is pertinent for further processing, rather than some absolute document position.

8.1.5 One token; several word forms

Finally, several word forms may be attached to a same token, as illustrated by the following examples.

```
<!–– Give it to me ––>
token form="damelo" id=" t1 ">Damelo</ token>
<wordForm lemma="da" tokens=" t1 "/> <!–– (Donne ) ––>
<wordForm lemma="me" tokens=" t1 "/> <!–– (le ) ––>
<wordForm lemma=" lo " tokens=" t1 "/> <!–– (moi ) ––>
```

```
<!–– of which ––>
<token id=" t0 ">auquel</ token>
<wordForm lemma="à" tag="pos.prep" tokens=" t0 "/>
<wordForm lemma=" lequel " tag="pos.pronrel " tokens=" t0 "/>
```

8.2 Referring lexicon entries

A word form is a linguistic unit carrying morpho-syntactic properties. Generally, a linguistic unit may be characterized by a label corresponding to an entry if some lexicon. This identification is materialized by the attribute entry, whose content should express a reference (an URN) to the lexicon entry.

```
<token id=" t1 ">Prime</ token>
<token id=" t2 ">minister</ token>
<wordForm entry=" urn : lexicon : en : prime_minister"
tokens=" t1 t2 "/>
```

The notion of “ lexicon entry” is outside the scope of MAF. A reference to a lexicon entry is therefore not precisely defined but, in first approximation, should correspond to an URN (Uniform Resource Name). It should be noted that one may wish to reference lexicons “sub-entries” for polysemous entries or for compound forms.

```
<token id=" t1 ">to</ token>
```

```
<token id=" t2 ">eventua l ly</ token>
<token id=" t3 ">devide</ token>
<wordForm entry="urn:lexicon:en:decide:to" tokens=" t1 t3 "/>
<wordForm entry="urn:lexicon:en:eventually" tokens=" t2 "/>
<wordForm entry="urn:lexicon:en:decide:ain"
tokens=" t1 t3 "/>
```

A token or a sequence of tokens may sometimes be identified as forming a word form because of various properties but can not associated to some lexicon entry, either because no lexicon is available or because the word form corresponds to a named entity (a proper name, a date, an address, . . .) or to a neologism. In that case, the content of attribute entry may be left empty. The other informative attributes lemma and form may still be used to provide more information about the word form.

```
<token id=" t0 ">October</ token>
<token id=" t1 ">,</ token>
<token id=" t2 ">23 rd</ token>
<token id=" t3 ">2005</ token>
<wordForm lemma="DATE" form=" 2005/10/23 " tokens=" t0 t1 t2 t3 "/>
```

For such unknown words, it is however suggested that they can be collected into a document specific lexicon, in order for the unknown words to refer entries in this lexicon.

8.3 Compound word forms

The structure of compound forms (including multi-word expressions) may be expressed using nested word forms, therefore providing information about the subparts even when none is available for the whole, for instance for neologisms:

```
<!—— birthday gift wrapping paper ——>
<token form=" Geburtstag " id=" t1 " join=" right ">Geburtstags</ token>
<token form=" Geschenk " id=" t2 " join  =" right ">Geschenk</    token>
<token form=" Papier " id=" t3 ">papier</ token>
<wordForm tokens=" t1 t2 t3 ">
<wordForm entry=" urn:lexicon:de:geburstag " lemma=" geburstag "
tokens=" t1 "/>
<wordForm entry=" urn:lexicon:de:geburstag " lemma=" geschenk"
tokens=" t2 "/>
<wordForm entry=" urn:lexicon:de:papier " lemma=" papier " tokens="     t3
"/>
</wordForm>
```

Note: Precising the derivational morphology of a compound word is outside the scope of MAF. Still, the addition of a deriv attribute on embedded word forms is being investigated, for instance to mention the head of a compound form.

8.4 Formal description: wordForm

```
wordForm =
      element wordForm {
            wordForm. identification ,
            wordForm. tokens ,
            wordForm* ,
            wordForm. content ?
      }
  wordForm. tokens =
      (attribute tokens { xsd: IDREFS }
      | ## DTD => ,
      \ token *
      )
wordForm.identification &= attribute entry { xsd: anyURI } ?
wordForm.identification &= attribute lemma { string } ?
wordForm.identification &= attribute form { string } ?
```

9 Morpho-syntactic content

This section explains how to attach morpho-syntactic content to word forms and how to define reusable tagsets to provide compact notations through tags and to control the validity of these contents.

The previous section explains how to enrich a document with morpho-syntactic annotations. However, it does not define the content of these annotations. What set of features and feature values should we use to express this content (within element wordForm) and with which meaning.

Such a set is usually referred as a tagset specifying the content of possible annotations. However, the diversity of approaches and languages makes almost impossible the proposition of an unique tagset. More modestly or pragmatically, the current proposal seeks to provide mechanisms to define tagsets by relying on a Data Category Registry (DCR) and Feature Structures Representations (FSR).

An annotated document will therefore be completed by either adding or referring to a tagset.

9.1 Using feature structures

A word form may be completed by a morpho-syntactic content defining its linguistic nature and its grammatical function in its current context. This content is expressed using Feature Structures, following the recommendation of ISO 24610 Part 1 document on “Feature Structure Representation” [FSR]. In first approximation, a feature structure may attach one or several (possibly complex) values to linguistic properties (i.e., noun to part of speech, present to tense, indicative to mood, . . .).

[1] <!—— nice ——>

```
<token id=" t0 ">belle</ token>
<wordForm entry=" urn:lexicon:fr:beau " lemma="beau" tokens=" t0 ">
    <fs>
        <f name="pos "><symbol value=" adjective "/></f>
        <f name=" adj_type"><symbol value=" qualifier "/></f>
        <f name=" gender "><symbol value=" feminine "/></f>
        <f name="number"><symbol value=" singular "/></f>
    </fs>
</wordForm>
```

The feature structure content attached to a word form may also provides additional information of interest about a word form.

9.2 Compact morpho-syntactic tags

FSR proposal provides ways for the compact representation of feature structures, by relying on libraries naming feature values and feature specifications (a feature specification being a pair formed by a feature and a value). These names may be used in wordForm attribute tag to get compact tags, following a standard practice in the NLP community.

```
<token id=" t0 ">belle</token>
 <wordForm tokens=" t0 "
entry=" urn:lexicon:fr:beau "
tag=" pos.adj.adj type.qual.gender.fem.num.sing "/>
```

The content of attribute tag should be similar to the content of attribute feats defined in FSR, namely a space-separated sequence of feature specification identifiers. The libraries naming recurrent values and feature specifications are part of the tagset(s) coming with the annotated document.

9.2.1 FSR libraries

The generic way provided by FSR to use libraries is illustrated by the following example, with the attribute feats of element fs:

```
<!—— A  feature  value library ——>
<fvLib n="French morpho values ">
    <symbol xml : id="noun" va lue="noun"/>
    <symbol xml : id=" sing " va lue=" singular "/>
    <symbol xml : id=" plu" va lue=" plural "/>
    <symbol xml : id="masc" va lue="masculine "/>
    <symbol xml : id="fem" va lue=" feminine "/>
</fvLib>
    <!—— A  feature  specification  library ——>
<fLib>
    <f xml : id=" pos . n " name=" pos " fVal=" noun "/>
```

```
        <f xml : id=" num. s " name=" number " fVal=" s ing "/>
        <f xml : id=" num. p " name=" number " fVal=" plu"/>
        <f xml : id=" gen . f " name=" gender " fVal=" fem "/>
        <f xml : id=" gen .m " name=" gender " fVal=" masc "/>
</fLib>
```

With such a library, following FSR rules, one may write:

```
  <wordForm lemma=" prime_minister" tokens=" t1 ">
        <fs feats ="pos . nnum. sg.gen.f "/>
  </wordForm>
```

or, equivalently, by using attribute tag, one may write:

```
<wordForm tokens=" t1 t2 "
                lemma=" prime_minister"
tag ="pos.nnum.sg.gen.f "/>
```

Disjunctive values are allowed by FSR and may also be simplified, following the same mechanism:

```
<!–– A  feature value library ––>
  <tagset>
        <fvLib>
                <vAlt xml : id=" first.third ">
                        <symbol value=" first "/>
                        <symbol value=" third "/>
                </vAlt>
                        <symbol xml : id=" verb" va lue=" verb"/>
                        <symbol xml : id=" sing " va lue=" singular "/>
        </fvLib>
<!–– A featur specificationlibrary ––>
        <fLib>
                <f xml : id=" pers.13 " name=" pers " fVal=" first.third ">
                </f>
                <f xml : id=" pos.v" name=" pos " fVal=" verb "/>
                <f xml : id=" num.s " name=" number " fVal=" sing "/>
        </fLib>
    </tagset>
<!–– Annotated document ––>
    <token id=" t0 ">po r t e</ token>
<wordForm tokens=" t0 "
                        entry=" urn:lexicon:fr:porter "
                        tag="pos.v.ers .13num.s"/>
```

9.3 Designing tagsets

The features, values, and possibly feature types used to specify morpho-syntactic content are not just labels but carry linguistic meanings, or, in other words, semantic content. To avoid misinterpretations, the semantic content attached to a feature, a value or a type should be clearly defined. The combination of features, values and types should also be controlled in order to avoid linguistically invalid combinations, such as using /neuter/ as a value for/gender/ in French, or using a feature /tense/ for nouns in most languages.

MAF does not try to define the semantic content of an unique complete set of such features, values, and types. It would be an almost impossible task given the diversity of languages, and it would be equally impossible to assign to each component a meaning agreed on by the whole community.

Instead, it is proposed that an annotated document should be completed by including or referring one or more tagsets.

The first objective of a tagset is to list the terminology used to annotate a document as a set of data categories whose meanings is precisely defined in a Data Category Registry, following the recommendation of ISO 12620 proposal on “Data Category Registry”. The process may be seen as selecting a subset of morpho-syntactic data categories (Data Category Selection – DCS).

```
<tagset>
        <dcs local=" genre " registered=" dcs:morphosyntax :
        gender:fr " rel=" eq "/>
        <dcs local="fem" registered=" dcs:morphosyntax:gender:fr:feminine "
        rel=" eq "/>
  </tagset>
```

The correspondence with a registered data category may not be perfect. The *rel* may be used to specify which relationship exists between the local and registered data categories. For instance, one may introduce a local data category /advneg/ as being subsumed by a more general registered data category /adverb/.

```
<dcs local=" advneg " registered=" dcs:morphosyntax:pos:adverb" rel=" subs
"/>
<dcs local=" strange " rel="none"/>
```

It is also possible (but not advised) to introduce a local data category bearing no relationship with any registered data category.

```
        <dcs local=" title ">
        <description> A part of speech used to denote honorific  titles like
        Pr.or S.A.S .
        </ description>
        </ dcs>
```

The second objective of a tagset is to specify the set of valid feature structures based on the selected data categories. It will be achieved by relying on the proposed ISO 24610 Part 2 on “Feature System Declaration” [FSD]. The third objective of a tagset is to name the most common morpho-syntactic structures through the use of FSR libraries, as seen in Section 8.2.1.

9.4 Formal description: tagset

```
wordForm. content -
        ( attribute tag { xsd : IDREFS }
        | ## DTD => ,
                fs
        )
fs |= notAl lowed # defined in iso−fs−standalone.rnc
tagset=
        element tagset{
                ( attribute ref { xsd : anyURI }
                | ## DTD => ,
                        ( dcs *& fsd * & tagset . lib * )
                )
        }
 dcs =
        element dcs {
                attribute local { xsd :NCName } ,
                (attribute registered { xsd : anyURI } ,
                        attribute rel { " eq" | " subs " | " gen" } ) ? ,
        element description { text }*
        }
  fsd |= notAllowed # defined in future iso−fsd.rnc
  tagset . lib |= fvLib
  tagset . lib |= fLib
  fLib |= notAllowed # defined in iso−fs−standalone.rnc
        fvLib |= notAllowed # defined in iso−fs−standalone.rnc
```

The *dcs* corresponds to a Data Category Selection part whose exact content is still to be defined.

The *fsd* corresponds to a Feature Structure Declaration part whose normalization is yet to be done.

10 Handling ambiguities

Ambiguities naturally arise when handling natural language, and especially for automatically produced annotations. Ambiguities may occur at various levels and, therefore, MAF proposes several alternatives to cope with ambiguities as simply as possible.

10.1 Word form Content Ambiguities

The proposal on Feature Structure Representation provides several ways to represent ambiguities, for instance at the level of feature values. These mechanisms may be used to handle the ambiguities occurring within the morpho-syntactic content of a word-form.
For instance, the French inflected verb form “mange” (to eat) is ambiguous between the 1st and 3rd persons, and this ambiguity can be captured by the vAlt element present in FSR:

```
<token id=" t0 ">mange</ token>
<wordForm tokens=" t0 " entry=" urn : lexicon : fr : manger ">
        <fs>
                <f name=" pos "><symbol value=" verb "/></ f>
                <f name=" aux "><symbol value=" avoir "/></ f>
                <f name=" mood "><symbol value=" indicative "/></ f>
                <f name=" tense "><symbol value=" present "/></ f>
                <f name=" person">
                        <vAlt>
                                <symbol value=" first "/>
                                <symbol value=" third "/>
                        </ vAlt>
                </f>
                <f name=" number "><symbol value=" singular "/></f>
        </fs>
</wordForm>
```

A compact tag notation can still be used by registering most frequent cases of ambiguities in FSR libraries (Section 8.2.1).

```
<token id=" t0 ">mange</ token>
<wordForm tokens=" t0 "
              entry=" urn:lexicon:fr:manger "
              tag=" pos.v aux.avoir mood.itense.ppers.13num.s "/>
```

10.2 Lexical Ambiguities

Ambiguities between different lexical entries for a same sequence of tokens can be handled by the element wfAlt:

```
<token id=" t0 ">porte</ token>
<wfAlt>
        <wordForm tokens=" t0 " entry=" lexicon : porte " tag="pos.n . . . "/>
```

```
    <wordForm tokens=" t0 " entry=" lexicon : porter " tag="pos.v . . . "/>
</wfAlt>
```

10.3 Structural Ambiguities

10.3.1 Structural ambiguities over word forms

A general and very generic answer is to describe the possible readings as paths through an Directed Acyclic Graph (DAG) whose edges are labeled by a word form. Such DAGs forms a sub-part of Finite State Automata and also cover the notion of word lattice used in parsing and speech recognition communities. They are powerful enough to represent ambiguities between several decompositions into compound forms. They can also be used to denote simpler cases of lexical ambiguities.

For instance, the French textual sequence “ fer à cheval ” (horse shoe) can still be decomposed into several readings (“[horse shoe]”, “[iron] [on horse]”, “[iron] [of] [horse]”), giving the following DAG:

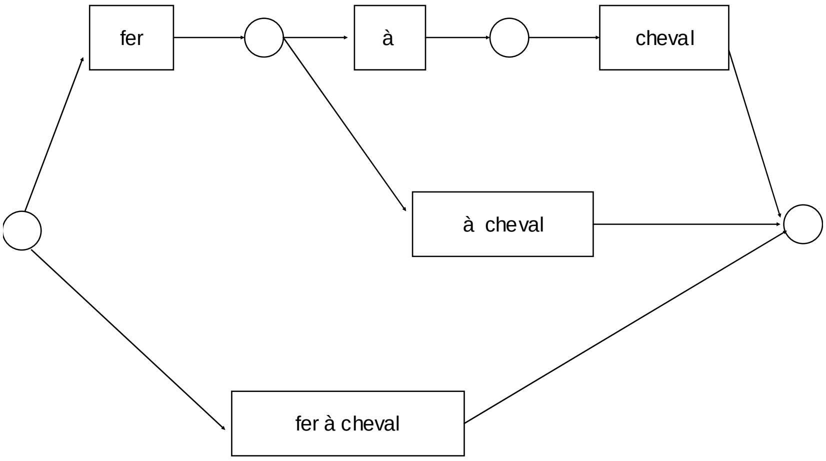

Figure 3: DAG for “fer a cheval“

```
<token id=" t1 ">fer</ token>
<token id=" t2 ">à</ token>
<token id=" t3 ">cheval</ token>
<fsm init="S0" final="S3">
    <transition source="S0" target ="S3">
        <wordForm tokens=" t1 t2 t3 "
            entry=" urn : lex : fr : fer_%E0_cheval"
            lemma=" fer_à_cheval"/>
    </transition>
    <transition source="S0" target ="S1">
        <wordForm entry=" urn : lex : fr : fer " tokens=" t1 "/>
    </transition>
    <transition source="S1" target ="S2">
            <wordForm tokens=" t2 "
                    entry=" urn : lex : fr :%E0" lemma="à"/>
    </transition >
    <transition source="S2" target ="S3">
            <wordForm tokens=" t3 " entry=" urn : lex : fr : cheval "/>
    </transition>
    <transition source="S1" target ="S3">
            <wordForm tokens=" t2 t3 "
                    entry=" urn : lex : fr :%E0_cheval" lemma=" à_cheval"/>
    </transition>
  </fsm
```

The linguistic units "fer à cheval", "fer", "à", "cheval", and "à cheval" correspond to minimal syntagmatic units that can be annotated. Additional information could be added to edges such as probabilities.

10.3.2 Structural ambiguities over tokens

Structural ambiguities may also arise over sequences of tokens, resulting from ambiguities in the tokenization of the annotated document, e.g. speech documents.
Structural ambiguities over tokens are represented by transitions labeled by tokens. The attributes tinit and tfinal on elements fsm are used to state the initial and final states for the token paths.

The two levels of structural ambiguities are represented by two lattices that form a kind of chart. It is not mandatory but advised that the two lattices share their states, whenever possible. A validity condition has to be expressed between the two levels of structural ambiguity: the tokens covered by word forms along a word form path belong to some token path.

10.4 Simplified structuring variants

10.4.1 Non ambiguous linear representation

When there is no ambiguity, MAF allows to replace the global lattice notation by a much simpler linear notation where the token, wordForm and wfAlt elements are implicitly chained following their appearance order, as illustrated by the following example:

```
<token id=" t1 ">fer</ token>
<token id=" t2 ">à</ token>
<token id=" t3 ">cheval</ token>
<wordForm entry=" urn : lex : fr : fer " tokens=" t1 "/>
<wordForm entry=" urn : lex : fr :%E0" tokens=" t2 "/>
<wordForm entry=" urn : lex : fr : cheval " tokens=" t3 "/
```

10.4.2 Mixed linear and lattice representation

Ambiguities are generally localized and it is tempting to also localize the use of the lattice notation only where it is needed. MAF allows to insert local lattice fsm in a linear flow of token, wordForm and wfAlt elements.

```
<token id=" t0 ">afin</ token>
<token id=" t1 ">de</ token>
<fsm init=" s0 " final=" s2 ">
    <transition source=" s0 " target =" s2 ">
        <wordForm tokens=" t0 t1 "
            entry=" urn : lex : fr : afin de " tag=" pos.prep "/>
    </transition>
    <transition source=" s0 " target =" s1 ">
        <wordForm tokens=" t0 "
            entry=" urn : lex : fr : afin " tag=" pos.prep "/>
    </transition>
    <transition source=" s1 " target =" s2 ">
        <wordForm tokens=" t1 "
            entry=" urn : lex : fr : de" tag=" pos.prep "/>
    </transition>
</fsm>
<token id=" t2 ">grandir</ token>
        <wordForm entry=" urn : lex : fr : grandir " tag=" pos.verb . . . "
    tokens=" t2 "/>
<token id=" t3 ">,</ token>
        <wordForm entry=" lexicon : , " tag=" pos.ponct " tokens=" t3 "/>
<token id=" t4 ">il</ token>
        <wordForm entry=" urn : lex : fr : il " tag=" pos.pronoun . . . "
        tokens=" t4 "/>
<token id=" t5 ">manger</ token>
        <wordForm tokens=" t5 "
            entry=" urn : lex : fr : manger " tag=" pos.verb . . . "/>
<token id=" t6 ">des</ token>
        <wordForm tokens=" t6 "
            entry=" urn : lex : fr : une " form=" des " tag=" pos .
            det.num@pl . . . "/>
<token id=" t7 ">pommes</ token>
<token id=" t8 ">de</ token>
<token id=" t9 ">terre</ token>
<fsm init=" s8 " final=" s11 ">
    <transition source=" s8 " target =" s11 ">
        <wordForm tokens=" t7 t8 t9 "
            entry=" urn : lex : fr : pomme_de_terre " tag=" pos.noun . .
        . "/>
    </transition>
    <transition source=" s8 " target =" s9 ">
        <wordForm tokens=" t7 "
            entry=" urn : lex : fr : pomme " tag=" pos.noun . . . "/>
    </transition>
    <transition source=" s9 " target=" s10 ">
        <wordForm tokens=" t8 "
            entry=" urn : lex : fr : de " tag=" pos.prep "/>
    </transition>
    <transition source=" s10 " target=" s11 ">
        <wordForm tokens=" t9 "
            entry=" urn : lex : fr : terre " tag=" pos.noun . . . "/>
</transition>
</fsm>
```

10.5 Expanding the simplified variants

The simplified variants are allowed because they may always be expanded into a global lattice, by applying the steps sketched in the following sub-sections.

10.5.1 Separating tokens and word forms

All tokens embedded within a word form may be extracted and moved just before the word form (and before an enclosing wfAlt) , not changing the relative order between tokens.

```
<wordForm entry=" urn : lex : fr :manger " tag=" pos.verb . . . ">
    <token id=" t6 ">des</ token>
</wordForm
```

becomes

```
<token id=" t6 ">des</ token>
<wordForm entry=" urn : lex : fr :manger " tag=" pos.verb . . . " tokens=" t6 "/
```

Note: There is no clear semantic to handle tokens embedded in word forms, themselves embedded in transitions. This case should be avoided.

10.5.2 Wrapping into local lattices

Tokens and word forms outside transitions are embedded into local lattices, wfAlt elements being considered as word forms.

```
<token id=" t4 ">i l</ token>
<wordForm ent ry=" urn : lex : fr : il " tag="pos . pronoun . . . " tokens=" t4
 "/>
<token id=" t5 ">mange</ token>
<wordForm entry=" urn : lex : fr :manger " tag="pos.verb . . . " tokens=" t5
```

```
 "/>
<token id=" t6 ">des</ token>
```

becomes

```
<fsm tinit=" s0 " tfinal=" s1 " init=" s0 " final=" s0 ">
        <transition source=" s0 " target=" s1 ">
                <token id=" t4 ">il</token>
        </transition
</fsm>
 <fsm init=" s0 " final=" s1 " tinit=" s0 " tfinal=" s0 ">
        <transition source=" s0 " target=" s1 ">
                <wordForm entry=" urn : lex : fr : il " tag="pos.pronoun . . . "
                tokens=" t4 "/>
        </transition >
  </fsm>
<fsm tinit=" s0 " tfinal =" s1 " init=" s0 " final =" s0 ">
        <transition source=" s0 " target=" s1 ">
                <token id=" t5 ">manger</ token>
        </transition >
</fsm>
  <fsm init=" s0 " final=" s1 " tinit=" s0 " tfinal=" s0 ">
        <transition source=" s0 " target=" s1 ">
                <wordForm ent ry=" urn : lex : fr :manger " tag="pos.verb . . . "
        tokens=" t5 "/>
        </transition>
  </fsm>
```

Lattice states are local to each lattice.

10.5.3 Merging local lattices

Two adjacent lattices may be merged by renaming the intermediary states in order to avoid name clashes and in such a way that the word form (resp. token) final state of the first lattice equals the word form (resp. token) initial state of the second lattice. Whenever possible, it is recommended, when merging, to rename the lattice states in such a way that the final (resp. final) states for tokens and word form coincide. The previous example becomes:

```
  <fsm tinit=" s0 " tfinal =" s1 " init=" s0 " final=" s1 ">
        <transition source=" s0 " target =" s1 ">
                <token id=" t4 ">i l</token>
        </transition>
        <transition source=" s0 " target =" s1 ">
                <wordForm entry=" urn : lex : fr : i +il " tag="pos.pronoun . . . "
        tokens=" t4 "/ >
        </transition>
</fsm>
```

```
<fsm tinit=" s0 " tfinal =" s1 " init=" s0 " final =" s1 ">
    <transition source=" s0 " target=" s1 ">
        <token id=" t5 ">mange</token>
    </transition>
    <transition sour c e=" s0 " target=" s1 ">
        <wordForm ent ry=" urn : lex : fr : manger " tag="pos . verb . . . "
    tokens=" t5 "/>
    </transition>
</fsm>
```

and then

```
<fsm tinit=" s0 " tfinal=" s2 " init=" s0 " final=" s2 ">
    <transition source=" s0 " target=" s1 ">
        <token id=" t4 ">i l</ token>
    </transition >
    <transition source=" s0 " target=" s1 ">
        <wordForm entry=" urn : lex : fr : il " tag="pos.pronoun . . . "
    tokens=" t4 "/>
    </transition >
    <transition source=" s1 " target=" s2 ">
        <token id=" t5 ">mange</ token>
    </transition >
    <transition source=" s1 " target=" s2 ">
        <wordForm entry=" urn : lex : fr : manger " tag="pos.verb . . . "
    tokens=" t5 “/>
    </transition >
</fsm>
```

10.5.4 Removing wfAlt

A transition over a lexical ambiguity, materialized by a **wfAlt** element, may be expanded into two equivalent simpler transitions.

```
<transition source=" s0 " target=" s1 ">
    <wfAlt>
        <wordForm tokens=" t0 " entry=" lexicon : porte " tag="pos.noun 4
        . . . "/>
        <wordForm tokens=" t0 " entry=" lexicon : porter " tag="pos.verb 6
        . . . "/>
    </wfAlt>
    </transition>
```

becomes

```
  <fsm init=" s0 " final=" s1 ">
  <transition source=" s0 " target=" s1 ">
        <wordForm tokens=" t0 " entry=" urn : lex : fr : porte " tag="pos.noun . . 4.
"/>
  </transition>
   <transition source=" s0 " target=" s1 ">
        <wordForm tokens=" t0 " entry=" urn : lex : fr : porter " tag="pos.verb . . 8
        . "/>
   </transition>
   /fsm>
```

The ordering of transitions inside lattices is not pertinent. On the other hand, the ordering of word forms and tokens outside lattices is pertinent. The relative ordering of local lattices is also pertinent.

10.6 Formal description: wfAlt and fsm

```
maf.flow = (\token | wordForm | wordForm.alt | fsm )+
fsm =
        element fsm {
                ( attribute init { fsm.state } ,
                 attribute final { fsm.state } ) ? ,
                (attribute tinit { fsm.state } ,
                 attribute tfinal { fsm.state } ) ? ,
                transition+
        }
fsm.state = xsd: Name
                transition =
        element transition {
                attribute source { fsm.state } ,
                attribute target { fsm.state } ,
                ( \token|wordForm|wordFormalt )
        }
  wordForm.alt =
        element wfAlt { wordForm+ }
```

11 Header and metadata

The global maf element is introduced as a root element to encapsulate morpho-syntactic annotations and carries global metadata relative to the annotated documents. Two MAF specific metadata categories are introduced for the token standoff notation, namely the document and addressing attributes. The addressing attribute indicates the addressing schema used to refer positions in the annotated document. A full list of such schema will be provided in ISO 24612 proposal “Linguistic Annotation

Framework” (LAF). The following fragment illustrates the use of these attributes for a video document:

```
<maf document=" interview.mpeg" addressing="mpeg7">
    <token id=" t0 "
            from="T00:01:16:4484F30000"
            to="T00:01:16:14494F30000"
            transcription=" mister "/>
<wordForm tokens="t0" lemma="mister "> . . . </wordForm>
. . .
</maf>
```

The other non-mandatory metadata are handled following the recommendations of the OLAC Metadata Standard and should therefore be included in an olac:olac element.

```
 <maf document=" http://abu.cnam.fr /cgi−bin/donner_abu? tdm80j2 "
            addressing=" char_offset ">
    <olac: olac xmlns : olac="http://www.language−archives.org
    /OLAC/1.0/ "
      xmlns="http://purl.org/dc/elements/1.1/ "
      xmlns :xsi="http://www.w3.org/2001/XMLSchema−instance "
      xsi: schemaLocation="http://www.language−archives.org/OLAC/1.0/
            http://www.language−archives.org/OLAC/1.0/olac xsd">
    <creator>MySuperMorphoTool</creator>
    <created>2005/09/30</created>
    <hasVersion>1 . 1</hasVersion>
    <identifier>TDM80MAF. 1 . 1</identifier>
    <replaces>TDM80MAF.1.0</replaces>
    <requires>http://abu.cnam.fr/cgi−bin/donner_abu? tdm80j2</requires>
    <language xsi: type=" olac: language " olac: code=" fr ">French</ language>
    <publisher>MyInstitution</publisher>
    <title xml: lang=" fr ">Le Tour du Monde en 80 Joursversion
MAF</ title>
    <abstract xml : lang=" en"> A set of MAF annotations for Jules Vernes
    famous nove l
    <abstract>
    <rightHolder>MyInstitution</rightHolder>
    <license>LGPL−LR</license>
</olac:olac>
    . . .
</maf>
```

11.1 Formal description

```
start =
    element maf {
            ( maf.document ,
               maf.addressing ) ? ,
              tagset ? ,
              maf.metadata ? ,
              maf.flow
    }
 maf.document = attribute document { xsd: anyURI }
 maf.addressing = attribute addressing { xsd: NMTOKEN }
maf.metadata |= notAllowed # to be imported from OLAC
```

The complete list of addressing schema allowed by MAF will be inherited from ISO 24612 document on Linguistic Annotation Framework (LAF). A possible list of such schema could include:

- TEI ptrs,

- XML Xpointers,

- character offsets (depending on the original document encoding)

- MPEG7 multimedia addressing (MediaTimePointType)

A (informative) RELAX NG compact schema

Note: The following RELAX NG compact schema may be found online at http://atoll.inria.fr/~clerger/MAF/maf.rnc

```
# $Id:maf.rnc , v 1 . 1 2005/09/06 08 : 38 : 02 clerger Exp $

default namespace = ""
namespace a = " http://relaxng.org/ns/compatibility/annotations/1.0 "

## Preliminary Relax NG schema for MAF −− Morpho−syntactic
Annotation Framework
## Eric de la Clergerie <Eric De_La_Clergerie@inria.fr>

## The following is for Feature Structures
include " iso−fs−standalone.rnc "

start =
    element maf {
            ( maf.document ,
            maf.addressing ) ? ,
            tagset ? ,
            maf.metadata ? ,
            maf.flow
    }
maf.document = attribute document { xsd:anyURI }
## To be defined in LAF
maf.addressing = attribute addressing { xsd :NMTOKEN }
## Global Metadata: to be completed
maf.metadata |= notAllowed        # to be imported from OLAC
\ token =
    element token {
            attribute id { xsd: ID }? ,
            token information ,
    (
            (
            attribute from { DocumentLocation } ,
            attribute to { DocumentLocation }
            )
    | ## DTD => ,
            (
    [ a :defaultValue = "no" ]
    attribute join { " no "|" left "| " right " | "both" | " overlap " }
    ? ,
    text
            )
    )
}
token.information &= attribute form { string }?
token.information &= attribute phonetic { string }?
token.information &= attribute transcription { string }?
token.information &= attribute transliteration { string }?
wordForm =
    element wordForm {
wordForm.identification ,
wordForm.tokens ,
wordForm* ,
wordForm.content ?
    }
wordForm.tokens =
    (attribute tokens { xsd : IDREFS }
    | ## DTD => ,
    \ token *
    )
wordForm.identification &= attribute ent ry { xsd : anyURI } ?
wordForm.identification &= attribute lemma { string } ?
wordForm.identification &= attribute form { string } ?
maf.flow = (\token| wordForm | wordForm.alt | fsm )+
fsm =
    element fsm {
            ( attribute init { fsm.state } ,
              attribute final { fsm.state ) ? ,
            ( attribute tinit { fsm.state } ,
              attribute tfinal { fsm.state } ) ? ,
            transition +
    }
fsm.state = xsd :Name
transition =
    element transition {
            attribute source { fsm.state } ,
            attribute target { fsm.state } ,
            (\ token | wordForm | wordForm.alt )
    }
wordForm.alt =
    element wfAlt { wordForm+ }
wordForm. content =
    (attribute tag { xsd : IDREFS }
    | ## DTD => ,
            fs
    )
fs |= notAllowed   # defined in iso−fs−standalone.rnc
tagset =
    element tagset {
            (attribute r e f { xsd: anyURI }
            | ## DTD => ,
                    ( dcs * & fsd * & tagset.l i b * )
            )
    }
dcs =
    element dcs {
              attribute local { xsd :NCName } ,
            ( attribute registered { xsd : anyURI } ,
             attribute rel { " eq" | " subs " | " gen" } ) ? ,
    element description { t e x t }*
    }
  fsd |= notAl lowed         # defined in future iso−f sd . rnc
  tagset . l ib |= fvLib
  tagset . l ib |= fLib
  fLib |= notAl lowed        # defined in iso−f s−standalone.rnc
  fvLib |= notAl lowed       # defined in iso−fs−standalone.rnc
  DocumentLocation = xsd :NMTOKEN # defined in LAF
```

A.1 Validating MAF documents

For validating MAF document, the first step is to convert the RELAX NG compact schema into an XML RELAX NG schema (for instance using trang). Such a XML RELAX NG schema may be found at http://atoll.inria.fr/~clerger/MAF/maf.rng Then, the validation may be performed, for instance, using xmllint (from libxml2).

xmllint --relaxng maf.rng mafdoc.xml

It should be noted that some semantics constraints of MAF are not checked by the RELAX NG schema, in particular the constraint between the word form and token paths expressed in Section 9.3.2.

B (informative) DTD

The following DTD is only be an approximation of the RELAX NG schema. Note: The current DTD is outdated w.r.t. the RELAX NG schema.

```
<?xml encoding="UTF−8"?>

<!−− $ Id:maf.rnc , v1.1 2005/09/06 08:38:02 clerger Exp $ −−>

<!−−
Preliminary Relax NG schema for MAF −− Morpho−syntactic Annotation Framework
Eric de la Clergerie <Eric . De_La_Clergerie@inria.f r>
−−>

<!ELEMENT maf ( tagset ? , ( fsm| token | wordForm| wfAlt )+)>
<!ATTLIST maf
    xmlns CDATA #FIXED ' '>

<!ATTLIST maf
    document CDATA #IMPLIED
    addres sing NMTOKEN #IMPLIED>

<!ENTITY % DocumentLocation "NMTOKEN">

<!ELEMENT token (#PCDATA)>
<!ATTLIST token
    xmlns CDATA #FIXED ' '
    id ID #IMPLIED
    form NMTOKEN #IMPLIED
    transcription NMTOKEN #IMPLIED
    transliteration NMTOKEN #IMPLIED
    from %DocumentLocation ; #REQUIRED
    to %DocumentLocation ; #REQUIRED
    join ( no|left|right|both|overlap) ' no '>

<!ELEMENT wordForm ( fs,token * ,wordForm*)>
<!ATTLIST wordForm
    xmlns CDATA #FIXED ' '
    entry CDATA #IMPLIED
    lemma CDATA #IMPLIED
    form CDATA #IMPLIED
    tag CDATA #REQUIRED
    tokens IDREFS #REQUIRED>

<!ELEMENT fsm (transition)+>
<!ATTLIST fsm
    xmlns CDATA #FIXED ' '
```

```
        init IDREF #IMPLIED
        final IDREF #IMPLIED
        tinit IDREF #IMPLIED
         tfinal IDREF #IMPLIED>

<!ELEMENT transition ( wordForm|wfAlt |token )>
<!ATTLIST transition
        xmlns CDATA #FIXED ’ ’
        source IDREF #REQUIRED
        target IDREF #REQUIRED>

<!ELEMENT wfAlt ( wordForm)+>
<!ATTLIST wfAlt
        xmlns CDATA #FIXED ’ ’>

<!ELEMENT tagset ( dcs|fsd|fvLib|fLib ) *>
<!ATTLIST tagset
        xmlns CDATA #FIXED ’ ’
        ref CDATA #REQUIRED>

<!ELEMENT dcs (description) *>
<!ATTLIST dcs
        xmlns CDATA #FIXED ’ ’
        loca l NMTOKEN #REQUIRED
        registered CDATA #IMPLIED
        rel ( eq|subs|gen ) #IMPLIED>

<!ELEMENT description (#PCDATA)>
<!ATTLIST description
        xmlns CDATA #FIXED’ ’>

<!–– should be completed by DTD’s for Feature Structures ––>

<!ELEMENT fsd EMPTY>
<!ATTLIST fsd
xmlns CDATA #FIXED ’ ’>
```

C (informative) Illustrative examples

C.1 Tagsets

C.2 Demonstrator

A preliminary demonstrator covering most of the functionalities provided by MAF may be tried on line at http://atoll.inria.fr/mafdemo.

D (illustrative) Morpho-syntactic Data Categories

This annexe lists and documents the morpho-syntactic data categories used in the MAF examples. A repository of data categories, including morpho-syntactic data categories, may be found at http://syntax.inist.fr/.

/grammatical gender/ with conceptual values /feminine/, /masculine/
/grammatical number/ with conceptual values /singular/, /plural/
/grammatical pos/ with conceptual values /noun/, /verb/, /preposition/, /determiner/, conceptadverb
/grammatical mood/ with conceptual values /indicative/, /subjunctive/
/grammatical tense/ with conceptual values /present/
/grammatical person/ with conceptual values /first/, /second/, /third

E (informative) UML notions used within MAF

E.1 Introduction

MAF complies with the specifications and modeling principles of UML as defined by OMG [32]. UML is well defined and broadly used in the industry. MAF uses a subset of UML that is relevant for linguistic description.
The following notions are used:

- The notion of class
- The notion of relationship
- The notion of instance
- The notion of package

E.2 The notion of class

A class is a named descriptor for a set of objects that share the same attribute s and relationships. Classes are described within a class model.

E.3 The notion of attribute

An attribute is the description of a named element of a specified type in a class; each object of a class separately holds a value of the type.

E.4 The notion of relationship

A relation is a connection between classes. This includes association and generalization. Relations are described within a class model.

E.5 The notion of association

An association is a relationship between two specified classes that describes connections among their objects. The extension of the association is a collection of such links. Associations are the glue that holds together the model: without associations, there is only a set of isolated classes. An association holds two ends. Each end has "a multiplicity" and an ordering qualifier. The multiplicity is the specification of the range of allowable cardinality values that a collection may assume. The multiplicity range is an integer interval with its minimum and maximum values. An ordering qualifier specifies whether the connection forms a set (an unordered collection) or a list (an ordered collection).

E.6 The notion of aggregation

An aggregation is a form of association that specifies a whole-part relationship between an aggregate (a whole) and a constituent part. It is not permissible for both ends to be aggregates.

E.7 The notion of generalization

A generalization relationship is a directed relationship between two classes. On e class is called the parent or the super-class, and the other is called the child or the sub-class. The parent is the description of a set of objects with common properties over all children. The child is a description of a subset of those objects that have the properties of the pa rent but that also have additional properties peculiar to the child. A parent may have more than on e child and a child may have more than one parent. Generalization is a transitive and anti-symmetrical relationship. No directed generalization cycles are allowed. A child inherits the attributes and associations of its parent.

E.8 The notion of instance

An instance is an object that conforms to a class. Instances are not described within a class model but within an instance model (sometimes called an object model).

E.9 The notion of package

A package is a grouping of classes and relations. Usually there is a single root package that owns the entire model for a system. A package may contain nested packages. Packages may have dependencies to other packages.

E.10 Graphical notations

Each notion has a graphical notation that is precisely defined as follows:

References

[1] Gilles Adda, Joseph Mariani, Patrick Paroubek, Martin Rajman, and Josette Lecomte. L'action grace d'évaluation de l'assignation de parties du discours pour le français. Langues, 2-2:119-130, 1999.

[2] Consortium Genelex. Projet Eureka Genelex - Rapport sur la couche Syntaxique - Rapport sur la couche morphologique, 1993.

[3] Nelson Francis and Henry Ku˘cera. Manual of Information to accompany a Standard Corpus of Present-day Edited American English, for use with Digital Computers. Brown University, Providence, Rhode Island, Revised 1989.

[4] Eva Hajicova, Jarmila Panevova, and Petr Sgall. Language ressources need annotations to make them reusable: the prague dependency treebank. In Proceedings First Conference on Linguistic Resources, pages 713-718, Granada, 1998.

[5] Nancy Ide, Jean Véronis, and Greg Priest-Dorman. Corpus encoding standard. Technical report, EAGLES/MULTEX, 1996.

[6] Timo Järvinen. Annotating 200 millions words: the bank of english project. In Proceedings 15th COLING, pages 565-568, Kyoto, 1994.

[7] Timo Järvinen. Bank of English and beyond, chapter Treebanks (éd. Anne Abeillé). Kluwer Academic Publishers, 2000.

[8] Patrick Paroubek and Martin Rajman. Etiquetage morpho-syntaxique, volume Ingénierie des langues, chapter in Ingénierie des Langues (éd Jean-Marie Pierrel). HERMES-Science, Paris, 2000.

[9] Antonio Sanfilippo. EAGLES Subcategorization Standards, 1996. http://www.icl.pi.cnr.it/EAGLES96/syntax/syntax.html.

[10] Kevin Sinclair. Preliminary recommendations on corpus typology. Technical report, EAGLES, 1996.

[11] Jean Véronis and Liliane Khouri. étiquetage grammatical multilingue : le projet multex. TAL, 36, 1995.

[12] Ursula von Rekowski. Elm-fr : Specifications for french morphosyntax, lexicon specification and classification guidelines. EAGLES document, 1996.

Comments

As mentioned in the introduction we list here some of the comments made to the MAF proposal. As also mentioned, the “upgrade” of MAF to the DIS level has not be approved by the majority of national standardization bodies, since a relevant number of editorial issues has been raised, touching mainly the clarification of linguistic concepts. But we can keep in mind that the general MAF meta-model for morpho-syntactic annotation has been received positively.
In the following the abbreviations “ed” stays for “editorial”, “te” for “technical” and “ge” for “general”.

1	2	(3)	4	5	(6)	(7)
MB[1]	**Clause No./ Subclause No./ Annex** (e.g. 3.1)	**Paragraph/ Figure/Table/Note** (e.g. Table 1)	**Type of com-ment**[2]	**Comment (justification for change) by the MB**	**Proposed change by the MB**	**Secretariat observations** on each comment submitted

SIS	3	3.7	ED	The form of the definition doesn’t follow the requirements specified in ISO 10241, clause 5.25	Move the sentence “In MAF, a feature shall denote …” to NOTE.	
SIS	3	3.23	ED	The form of the definition doesn’t follow the requirements specified in ISO 10241, clause 5.25	Move the sentence “This field combines…” to NOTE	
SIS	3	3.26	TE	The definition doesn’t follow the basic principles to be applied to the drafting of definitions ISO 10241, clause 5.25 b)	Rewrite the definition	
SIS	3	3.27 & 3.36	TE	The definitions of part of speech and word form an external circle (if one substitutes *word* in the definition of *part of speech by* the definition given in 3.36). The form of the definition in 3.36 is not consistent with the requirements specified in ISO 10241, clause 5.25.	Define word without referring to part of speech	
SIS	3	3.28	ED	The form of the definition doesn’t follow the requirements specified in ISO 10241, clause 5.25	Move the sentence “This identification is the basis …” to NOTE	

1	2	(3)	4	5	(6)	(7)
MB[1]	**Clause No./ Subclause No./ Annex** (e.g. 3.1)	**Paragraph/ Figure/Table/Note** (e.g. Table 1)	**Type of com-ment**[2]	**Comment (justification for change) by the MB**	**Proposed change by the MB**	**Secretariat observations** on each comment submitted
	clause 1, scope			annotation phase': why seen as a process? Which advantage has process view? Mention corpora etc., to contextualize the statements.		
	3			- The definitions should be reworked.		
	3.1			`which does not require their effective presence'. Not clear what is meant.		
	3.2			would it make sense to relate the `closed data category' to the notion of `enumerable values' ?		
	3.3			the notion of `complex data category' is not defined in the standard, but made reference to		
	3.6:			no definition provided for `discourse' `lemmatized form', as lemma and lemmatized form are by no means the same thing, except in the one exceptional case.		
	3.9			finite state automaton (singular) preferred over the plural form		
	3.12			`marking of a' --> `marking of a word form'		
	3.16			we would prefer if a formal definition were given. Even if you think that there is confusion in the literature, we expect that YOU have a clear definition. - p.9.: why is the `lemma' seen as a `class of inflected forms differing only in inflectional morphology'? This could be understood as if a set of forms were called, if appearing together, a `lemma'. This is at least not thestandard view in (computational) linguistics. Why not call `lemma' a ``label attached to inflected forms to show that such inflected form belongs to a paradigm of inflected forms of a given underlying lexical element'' ? Maybe then have a separate entry for `lemmatized form', as lemma and lemmatized form are by no means the same thing, except in the one exceptional case.		
	3.17			if you have the `lexeme', wouldn't you want to have a term for the `grammatical morpheme' ? Introducing the definition via a parenthesis in sec. 3.19 is bad definition style.		
	3.21			We don't understand why `morphology of a word' and`morphosyntax of a word' are presented as quasi-synonyms. This is against all current theories.		
	3.22			The fact of being ordered is relevant for a very small subset of all MWEs, namely pair formulae, such as Punch and Judy, clair et net, frank und frei, etc. For all other MWEs, the order of the elements is mostly determined by the constraints of syntax. Do you want a narrow view on MWEs?		
	3.24			remove `due to the organic nature of language'. This does not explain anything; you may talk about productivity, if you feel the need. Rather leave 3.24 without indication of a reason.		
	3.36			`single word expression': not defined.		
	5.1			the introduction is not linguistically sound in its current version. - reformulate first paragraph: you mix phenomena and processes for NLP:		

				`other kinds of annotations on references, discourse, prosody (so far: phenomena) or parsing (process) may complete morpho-syntactic annotations' - `in the definition of parts of speech and grammatical categories': for most theories, these two are the same thing. Maybe say: `in the definition of parts of speech and of the categories used for morphoysntactic annotation' - `of a given langugage in a given context' --> - `of a given langugage unit in a given context' --> (??) - `Minimal units cannot be broken into subparts that could be identifiedby similar morpho-syntactic criteria, but may however still be broken into smaller units with morphological or phonological properties'. This may be right in its basic idea. but it is an overt contradiction with the (completely erroneous) introduction of `morphology of a word' and `morphosyntax of a word' as quasi-synonyms in sec. 3.21, p.10.		
	5.2			Sec. 5.2.: We think that your excursus into DAGs is not relevant here.		
				- p.14.: `phonological or lexicological structure': ---> `or lexical' Lexicology is the field of semantics dealing with lexical meaning.		
	7.1.4			- p.20: Your notion of zero token as one word form is heavily dependent on syntactic theory. This should be said, because otherwise people get a false impression. The MORPHOsyntactic view is not clear, here. We think you code a certain approach to syntax. - p.21: Example `damelo'  If this is meant to be Italien, it should be written as `dammelo'; then, the question would be how to account for the (morphophonologically determined) duplication of -m-. Secondly, why is the imperative form `da' seen as a value of `lemma'? Thirdly, the form `me' does not appear in Italian. The free variant of it is obviously `mi'. It may be useful to address, in the MAF proposal, the issue of Italian `me/mi/mme' in corpora. If your example is meant to be Spanish, as in the song line Dámelo dámelo dame lo que quiero Sin excusas ni rodeos then the problem of the lemmatization of `dá' as opposed to `dar' remains. You might usefully note which langugae the examples are meant to cover.		
	7.2			p.21., sec. lexicons `subentries' ---> lexicon_ subentries		
	7.3			- p.22, Your example is an ad-hoc compound which is not found in corpora (not in 200 M words of DE news texts i have at hand). Your note: `precising the derivational morphology of a compound word...' This is either meant to mean `... the compounding morphology'		

				or it is obvious, as MAF is not meant to describe derivation and compounding. The idea to introduce a `deriv' attribute remains completely opaque to us. Note that most morphological theories distinguish between derivatino and compounding.		

1	2	(3)	4	5	(6)	(7)
MB[1]	**Clause No./ Subclause No./ Annex** (e.g. 3.1)	**Paragraph/ Figure/Table/Note** (e.g. Table 1)	**Type of com-ment**[2]	**Comment (justification for change) by the MB**	**Proposed change by the MB**	**Secretariat observations** on each comment submitted

Ca			ge	Which existing morpho-syntactic annotations were analyzed as a basis for developing this standard? It is important to know the industry coverage.	List the existing morpho-syntactic annotations that were evaluated, in an annex.	
Ca			ge	The document seems to lack sufficient interconnections with other SC4 documents/ standards such as Linguistic Annotation Framework, Feature Structure Representation, and Feature System Declaration. We need to add a clear picture of how these are tied together and work together.		
Ca			ed	Formatting. Use consistent approach for bold, italics, etc. Many inconsistencies here.		
Ca			ge	Terminology comments… Be consistent, precise, and cohesive in use of key linguistic terms such as: morphological vs morpho-syntactic; morphology vs morpho-syntax -- They are often used as synonyms but in linguistics they are not synonyms. Use just “morphological” when you are referring to purely morphological concepts. Do not append “syntax” unless there are real syntactic elements or properties involved. Why are you using “morpho-syntax” in place of “morphology” in most cases, as suggested in the Note at 3.21. Linguistics has a long history of referring to features such as number, gender, etc., as morphological. It is understood that they have a role in syntax. We think the use of the term “morpho-syntax” throughout the document merely adds confusion. The definition of lexicon is too broad. The authors might want to to link or distinguish the use of the term in this document with the usage done in other SC4 documents. semantics – use this term only when referring to the linguistic concept of semantics – the study of meanings of words. Do not use to refer to the meaning of tags or tagsets.		

				terminology – do not use to refer to the nomenclature of tags		
Ca	3		ge	There following terms should be added, with definitions: word form morpho-syntactic annotation Data Category Registry embedding notation standoff notation structural ambiguity lexical ambiguity different types of tokens – element token, joining token, sibling token annotation transition (if it can be defined with the linguistic-specific meaning used here) feature tagset lemmatize (verb) compound word (make sure it can easily be differentiated with multi-word unit)		
Ca	3		ge	The definitions of the following terms are poor and should be rewritten and clarified. associative relation - 2nd sentence is incomprehensible and the “their” reference is ambiguous. data category – isn’t a “closed data field” also a “given data field”? Therefore the “or” part makes no sense. What is a “closed data field”? Other SC4 documents only talk of “closed data categories” not “closed data fields” directed acyclic graph – Definition not understandable by itself. finite state automata – change to “finite set of transitions from an initial state to a final state” inflection – missing word before “so”. Remove “(i.e. relational)” lattice – Definitions should never contain constructions like “term often used…”. Remove everything up to “a directed acyclic graph…” lexeme – incomprehensible. The definition relies on a distinction with grammatical morpheme… but this term is not defined itself. “open list” and “automonous signification” do not explain the meaning at all. Add a “see also”		

				to grammatical morpheme if you do define it. “lexical morpheme” should be indicated as synonym. morpheme – remove “and that cannot be divided into smaller meaningful units” (repetition with “smallest linguistic unit”) morpho-syntactic feature – “category” is not specific enough to be a genus. Change “values” to “morpho-syntactic features” in the note. multi-word expression – change “that” to “, which” in the example. natural language processing – change “skills” to “disciplines” open data category – remove everything from “due” to the end. morpho-syntactic tag – the definition is not a proper definition. Suggested rewrite: “a tag that represents a feature value pair that is common to related entities in an associative relation. For example, a morpho-syntactic tag could represent the feature-value pair part-of-speech=”noun”. (Structurally improved definition but I am not sure it is the accurate meaning. For example, I wonder whether the part “that is common to related entities in an associative relation” is essential.) word – remove “in the context of a given language”. Also, is this really the meaning of “word”? … “a description composed of a part of speech and a lemmatized form”… It seems very odd. A word is not a description and it does not contain a part of speech (it has a POS) and a lemmatized form (the word can be linked to a lemma or brought back to a lemmatized form). Also, it seems inaccurate to state that a word is a “multi-word expression” when in fact a multi-word expression is comprised of several words. It is circular to even use this terminology (a **word** is a multi-**word** expression).		
Ca	3		ge	The definitions of the following terms contain nested definitions of other terms, or inline terms that should be defined. The nested term is in parentheses. associative relation (paradigmatic relation) morpheme (grammeme/grammatical morpheme)	Separate the nested term and definition to its own entry.	

Ca	3		ge	The use of “see also” in the following terms should be checked because “see also” should only be used for related terms and it appears to sometimes be used here for synonyms. grammatical category - if “part of speech” is a synonym, then use “see”. If it is not a synonym, then this entry is lacking a definition. Need to clearly state whether or not these are synonyms (this affects other parts of the standard, e.g. 5.1) and if they are not, clearly indicate the differences. Affects also 3.27 – part of speech word class	If the reference is a synonym, change “see also” to “see”. Note that also any term that has a “see” reference should NOT have a definition.	
Ca	3		ge	discourse – missing a definition	Add a definition	
Ca	3		ed	Bolding is used in definitions, we assume to indicate that the bolded word is also defined. But often this is not the case.	Verify that the bolded word is defined. If it is not, either add the word and definition or remove bolding.	
Ca	3		ed	The see also references should use the full form not the abbreviation. For example, change “See also DAG” to “See also directed acyclic graph”.		
Ca	3		ed	lemma – lematized form This is not a term. The term is “lemma” and a synonym is “lemmatized form”. Reformat. Also, the genus seems incorrect. How can a “lemmatized form” be a “class of inflected forms”? The entire definition is odd. For instance, “infinitive for French verbs” is not a “lemma”. “porter” is a lemma, but it is an instantiation of the infinitive of French verbs.	Suggested rewrite: lemma lemmatized form The non-inflected canonical form of all the inflected forms of a word class. For example, “porter” is the lemma of “portons” and “portaient”.	
Ca	4.2		ed	move “FSR” and “FSD” from the heading to the paragraph text		
Ca	4.4		ed	Change “OMG” to “Object Management Group (OMG)”		
Ca	5.1		ed	Put “Linguistic Community” to lower case. Can we simply refer to “nouns” instead of “substantives”? Avoid using the construction “…” Say instead “and so forth”. (this affects many other parts of the document) 2nd last sentence – change “level” to “levels” 3rd last sentence (starting with “Minimal units…”) is contradictory.		
Ca	5.2		ed	Avoid “w.r.t.” (this affects other parts of the doc). Use the full expansion. Do not use the term “terminology” to refer to the		

				nomenclature of tagsets. Change “Because of structural ambiguities” to “Because of the inherent structural ambiguities of natural language”.		
Ca	6		ed	Change “brachygraphies” to “short-hand” Change “TEI” to “Text Encoding Initiative (TEI)”		
Ca	6.1		ge	Add an example of a standoff notation		
Ca	6.2		ed	Change “A” to “The” in “A first reason” and “A second reason”		
Ca	6.2		ge	The last sentence is very difficult to understand.		
Ca	6.3		te	It seems that the “form” attribute is used for multiple purposes (abbreviation, variation, typo, date format, etc.) This seems to go against the granularity priniciple needed for meaningful processing. We would prefer unique attributes for unique form types.		
Ca	6.4		ed	“explicit” is an adjective not a verb. Suggested rewrite: “… in particular to represent…” Change “As above mentioned” to “As mentioned above”		
Ca	6.4.2		ed	“multi-locutor” is a literal translation from French which does not work well in English. I am also not sure how overlapping tokens would be used for overlapping discourses (or even, what an overlapping discourse is). An example would help demonstrate this.	Suggested rewrite: “… or to mark overlapping discourses in documents with verbal exchanges”. or a longer sentence that could explain the concept of overlapping.	
Ca	6.5		te	Explain the meanings of the join attribute values : left, right, both, overlap. (It seems that the example with “etc” should have the value “left” instead of “overlap” but without an explanation of these attribute values we cannot be sure).		
Ca	7.1.2		te	Does the underscore belong in the lemma value? (lemma=”prime_minister”).		
Ca	7.1.3		te	same comment as above		
Ca	7.1.3		te	Explain the “to” and “main” parts of the entry urn values. It appears odd that two different lexical entries, denoted by “to” and “main” in the urn pointers, would be used to record information about the verb “to decide”.		
Ca	7.1.4		te	Explain the lemma=”PRO”.		
Ca	7.1.5		te	There seems to be an error in the last 2 lines of the “give it to me” example. The “le” comment belongs with the last line and the “moi” comment belongs with the 2nd last line.	Switch comments	
Ca	7.2		ed	Change “an URN” to “a URN”. The underscore in the URN seems acceptable in a URN but not in a lemma value as		

				commented above. Last sentence – missing “to” before “entries”.		
Ca	7.2		te	How can the lemma of “October, 23rd 2005” be “DATE”? Please explain.		
Ca	7.2		te	Samme comment as for 7.1.3 regarding “to” and “main”		
Ca	7.3		ed	Is a multi-word expression a type of compound form? It is worded like this. Distinguish between “compound form”, “compound word” and “multi-word expression” in the glossary. And ajust the wording here to match the relationships.		
Ca	7.3		ed	Change “precising” to “specifying”.		
Ca	8		ed	Add “to” after “referred” Change “makes almost” to “makes it almost” Change “the proposition” to “to propose” Change “of an unique” to “of a unique”		
Ca	8.1		ed	The sentence after the example could be removed. (“The feature structure content…”). It is fairly obvious.		
Ca	8.2		ed	Remove “proposal” in the first sentence. Change “coming” to “provided” in the last sentence.		
Ca	8.3		te	Review the use of “semantic” in this context. The first sentence seems odd to say that the features that specify “morpho-syntactic content” also specify “semantic content”… if this was true then we could say they specify “morpho-syntactic-semantic” content. Do you mean that that the features, values and types represent data categories and as such have a definable meaning? This could be the result of confusion between semantics in the true linguistic sense and metadata properties. We should avoid using the term “semantics” in this context. Perhaps use expressions like “linguistic role” and “definable scope”.		
Ca	8.3	3rd and 4th paragraph	ed	Change “referring one” to “referring to one”. Change “terminology” to “labels” (3rd paragraph) Change “whose meanings is” to “whose meanings are” Remove “precisely”		
Ca	8.3		te	The author seems to use the rel attribute “none” or not specify a rel attribute at all to achieve the same goal. Why have 2 methods?		
Ca	8.4		te	Explain the rel attribute values – eq, subs and gen. Also, it seems that “none” is missing.		
Ca	9.1 and 9.2		te	Are you are saying that this level of annotation would be the first stage, before the meta-context is analyzed, and a 2nd stage would look at the context and then choose the relevant vAlt element to retain? The process to resolve the		

				ambiguity in the entire text analysis and annotation process should be explained here or a reference made to another document that covers the processing.		
Ca	9.3.1		ed	Change “A general and very generic answer is” to “It is possible”. Add “of tokens” after “readings” Change “forms” to “form” after “DAGs”		
Ca	9.3.1		ed	Explain the meaning of “edge” in DAGs. There does not seem to be any edges in Fig 3. Maybe this should go in the glossary. Or maybe “edge” isn’t the best term. (Should it be boxes?) Explain further the last sentence. Probabilities of what?		
Ca	9.3.2		ge	We need an example of a lattice. The text in bold is incomprehensible.		
Ca	9.4.1		te	wfAlt is not formally described until section 9.6 but is used here. Also, it says it shows in the example yet the example doesn’t contain it. Perhaps the section describing wfAlt should be moved forward and this example should be enhanced to include the element.		
Ca	9.4.1		te	The example is supposed to represent an annotation where no ambiguity exists. Yet the chosen expression “fer à cheval” demonstrates ambiguity (and in fact is used to do so in section 9.3.1), by virtue of the fact that the meaning is not “fer” + “à” + “cheval”. Example should be changed to a non-ambiguous one such as “maison en brique” (Note for argument sake how you represent “pomme de terre” on the next page).		
Ca	9.5.1		te	Same comment about wfAlt as for 9.4.1		
Ca	9.5.1		te	Do not use “word form” when you mean “<wordForm>”. This is confusing. A search on the entire document for “word form” and see if others should be replaced by “<wordForm>” otherwise it is very easy to misunderstand when you are talking about an MAF element and not a general idea of a word.		
Ca	9.5.1		te	Shouldn’t the tokens value of the last line in the example be “t5”?		
Ca	9.5.1		ed	What do you mean by “semantic” in the last sentence? Change to “method”?		
Ca	9.5.2		ed	Clarify the meaning of the last sentence. “Lattice states are local to each lattice.”		
Ca	10		te	The sample has a closing <abstract> tag which should be </abstract> (it is missing the slash)		
Ca	A.1		ed	add “an” before “MAF document”		
Ca	A.1		te	The strings trang, xmlint, libxm12 are without any context here. Please expand this section with a basic technical description for them or, otherwise, remove this detail and just point readers to the inria web site.		

Ca	B		ed	Remove “only be”		
Ca	E		ge	This section is not described in relation at all to MAF. In fact, it could be taken from another UML source. If so, we don’t think we should repeat general UML information here and it would be sufficient to keep E.1 and remove the rest. Refer the reader for the original source of the UML descriptions. Note the extra spaces that should be removed: attribute s On e on e pa rent E.10 is missing content.		