Linguistic resource management – Meta-model for morpho-syntactic annotation

ISO TC37 SC4 Language Resource Management
WG 1–1 Linguistic Annotation Framework
Morphosyntactic annotation working draft
Lionel Clément
INRIA Rocquencourt –Atoll project

April 7, 2003

Introduction

Morpho-syntactic annotation involves the identification of word classes and properties over a continuous stream of word tokens. The annotation may refer to the segmentation of the input stream into word tokens, but may also involve grouping together sequences of tokens or identifying sub-token units, depending on the language under consideration and, in particular, the definitions of the linguistic units as applied to this language.

The description of word classes may include one or several features such as part of speech, lemma, flexional features etc., which is again dependent on the language being analyzed.

1 Normative references

To be completed:

- XML – Extensible Mark-up Language

  `http://www.w3.org/XML/`

- XML Schema

  `http://www.w3.org/XML/Schema`

- RDF – Resource Description Framework

  `http://www.w3.org/RDF/`

- XSL – Extensible Stylesheet Language

  `http://www.w3.org/Style/XSL/`

- XML Query

  `http://www.w3.org/XML/Query`

- MULTEXT

  `http://www.lpl.univ-aix.fr/projects/multext/`

- EAGLES

- CES

  `http://www.cs.vassar.edu/CES/`

- Negr@

  `http://www.coli.uni-sb.de/sfb378/negra-corpus/`

2 Basic concepts

The morpho-syntactic annotation may refer to different segmentation units:

- word segmentation
- token segmentation
- compound segmentation
- chunks

These units are decorated with tags and feature structures which describe their morpho-syntactic properties.

In order to represent disjunctions between continuous streams of tokens of different lengths, we propose to represent a Directed Acyclic Graph (DAG) as a Finite State Machine. This DAG is a possible segmentation for a continuous stream and other segmentations can be added for the same stream.

A Finite Automaton is defined as a set of states, including one initial state and one or more final states, and a set of labeled transitions between two states.

The user may add information on states and transitions, thus we propose to create elements for both and we may envisage that this graph is decorated with weights for a stochastic description.

Tags are compound objects with an internal structure, that may specify morpho-syntactic informations, as for instance a Part Of Speech, a grammatical category (gender, mood, etc.), and so on.

The internal structure should refer to feature structures either explicitly through **fs** notations (that should rely on ongoing normalizations efforts) or through references to **fs** from a library (for a more compact notation closer to those generally used).

Tag component (names of POS, names of properties such as gender, names of property values such as masc.) should in principle be registered.

If not registered, tag components should be linked to a registered tag component. Several kind of linking may be investigated:

- sub-typing: one may define some tag component as a subtype of some registered tag component

  question: do we need multiple inheritance?

- FS equation/subsumption: one may establish a relation between FS based on registered components and user defined FS.

  e.g.

  ```
  <fs>
     <f name="pos">
        <sym value="numeral"/>
     </f>
  </fs>
  ```

  User defined tag component should come with a short explanation and linguistic samples (for human understanding).

2.1 Segmentation units

Due to several levels we want to represent, we propose different elements and attributes for different segmentation units. We propose some general criteria:

- The text is represented only once as the PCDATA in the more embedded element.

- A hierarchical representation is proposed for tags in order to mark several levels of precision.

- The source text (graphical, phonetic or phonetic re-transcription) is associated with variations.

- The morpho-syntactic annotation is added to other kinds of annotations.

2.2 Token

We assume that the only kind of data to annotate is a sequence of linguistic units. These units may already be annotated with XML tags concerning a lower level of description (e.g., phonological annotation, morphological annotation, speech re-transcription annotations, etc.)

We propose an element **token** which can be defined as a segment of the source data, without any formatting information (blanks, paragraph jumps, etc.)

The delimitation of a token can be given inside the source data, in which are preserved typos, abbreviations and any kind of errors, as well as XML annotations as explained before. It can also be given as spans of already defined constituents of the data.

Example of the latter case:

```
<token idrefs="34 35"/>
<token idrefs="35 36"/>
<token idrefs="36 37"/>
<token idrefs="37 38"/>
```

Because the contents does not always correspond to the string recognized as a word (or a part of word), we add an attribute **form** which contains the stream to analyze. The contents of the element **token** is the effective realization of the attribute **form**.

For example, the French graphical realization encountered in a text *Apparaitre* corresponds to the form *apparaître*. In this case the graphical realization and the form are the same except for the capital and the diacritic accent.

The French graphical realization *Auquel* ("to which") corresponds to a pair of words (because this graphical stream is polycategorial) *à* ("to"), *lequel* ("which").

We propose two visions for the representation of these phenomena.

In the first one, the graphical realization *Auquel* is split into two tokens, one with *Auquel* as contents and one with an empty contents. In the other one, we have only one token associated with two words, in the sense of the DAG described hereafter.

Another example of this phenomenon is the string *etc.* If it is the last word of a sentence (which would normally be finished by a punctuation dot), the dot following *etc* is a marker of the abbreviation or the merging of such a dot with a punctuation dot. Therefore, depending on the context, the string *etc.* corresponds to a token which is the abbreviation of *et cetera*, or the merging of this abbreviation and the punctuation dot. Both solutions proposed for *auquel* can be used here if *etc.* ends the sentence.

Example of the first solution for *etc.* ending the sentence:

```
<token form="etc" id="0">etc.</token>
<token form="DOT" id="1"/>
```

In order to mark other graphical realizations (e.g. German glued polycategorial words), we propose to add a special attribute to the **token** element, that takes the values **glued (to followed token)**, **end of sentence**, **end of paragraph**, **beginning of sentence**, etc.

The German word *Geburtstagsgeschenkpapier* ("birthday gift paper") corresponds to several words (depending to this analysis in tree or four words). For this reason, the graphical string is split in tree or four tokens with an attribute which indicates that the compound noun is agglutinated.

```
<token form="Geburtstags" id="0" type="glued">Geburtstags</token>
<token form="Geschenk" id="1" type="glued">geschenk</token>
<token form="Papier" id="2">papier</token>
```

For example, *aujourd'hui* ("today" in French) is split in two tokens in case the splitter is only based on regular expressions. Both tokens will always be combined as a unique word.

```
<token form="aujourd" id="0">aujourd'</token>
<token form="hui" id="1">hui</token>
```

2.3 Word

A word is a linguistic unit whose definition is not considered here. It corresponds to an element **w** without any content but an attribute idref which refers to one or several tokens. One or more words can refer to the same token.

An attribute **entry** is proposed in order to annotate the exact lexicographic entry. The entry attribute may or may not correspond to the realization. For example, *Auquel* ("to which") is a compound word which can be seen (as previously explained) as formed by two tokens :

```
<token form="à" id="0"/>
<token form="lequel" id="1">Auquel</token>
<w entry="auquel" idref="0 1"/>
```

2.4 Simple words and Compounds

In order to annotate ambiguities between compounds and simple words, we use an DAG. We encode these DAGs with a list of states decorated with an attribute to mark the final ones and the initial one, and a list of transitions between these states.

The corresponding elements are respectively **state** and **transition**.

For example, the French string *pomme de terre cuite* may be ambiguous between several word segmentations: (cooked potato (*pomme_ de_ terre cuite*) or ceramic apple (*pomme de terre_ cuite*)). Finally, we could want to describe every simple word ("cooked earth apple", *pomme de terre cuite*). One representation of all these possibilities could be:

```
<s>
<token form="pomme" id="0">Pomme</token>
<token form="de" id="1">de</token>
<token form="terre" id="2">terre</token>
<token form="cuite" id="3">cuite</token>
        <state id="S0" type="initial"/>
        <state id="S2"/>
        <state id="S3"/>
        <state id="S4"/>
        <state id="S5" type="final"/>
        <transition source="S0" target="S4">
                <w entry="pomme de terre" idrefs="0 1 2"/>
        </transition>
        <transition source="S0" target="S1">
                <w entry="pomme" idrefs="0"/>
        </transition>
        <transition source="S1" target="S2">
                <w entry="de" idrefs="1"/>
        </transition>
        <transition source="S3" target="S5">
                <w entry="terre cuite" idrefs="1"/>
        </transition>
        <transition source="S3" target="S4">
                <w entry="terre" idrefs="1"/>
        </transition>
        <transition source="S4" target="S5">
                <w entry="cuite" idrefs="1"/>
        </transition>
</s>
```

2.5 S-units

At a morpho-syntactic level of description, we do not segment text in clauses, phrases or sentences. But we need a means to annotate a chunk of text which corresponds to one Finite State Machine. This element is called **s** (as in the CES) without any linguistic definition. It can be for example a graphical string matching a regular expression, texts units defined by a syntactic description, etc.

A **s** element contains a list of tokens, and a Finite State Automaton as Directed Acyclic Graph.

2.6 Linguistic categories

Each word is annotated with a set of tags and feature structures which denote linguistic categories for this unit. It is not appropriate here to give an exhaustive list of such tags. We just propose to classify these tags in two categories: A head category called "Part Of Speech", and sub-categories which are attributes of the head category. All these tags are organized in a features structure complemented with an optional human-readable summary.

2.6.1 Part of speech

We generally use distributional criteria to define the POS, but in case the classical notion of *partes orationis* can be used, logical criteria can also be retained. The user of this norm may define a set of POS consistent with the theory he uses. So we do not define a finite set of categories, but we allow the user to define or to refer to his own set.

Possible split criteria for POS we can retain are:

- distribution properties
- meaning
- flexional morphology
- derivational morphology
- grammatical function or dependencies

The POS taxonomy is defined by the user and depends strongly on the language and on the linguistic theoretical choices. What is a POS for someone could be a grammatical property for another. For example, being a *numeral* could be seen as belonging to a specific distributional category or as having a semantic property while being a noun, an adjective or a pronoun for example. For this reason, we propose to annotate the POS in a way which allows subcategorisation of POS.

Thus, we propose to encode POS and grammatical properties with feature structures associated with compact tags that sum up in a human-readable way the feature structures.

```
<tag type="POS" name="NumNounMascSing">
<fs>
pos = Noun
subcat = Num


</fs>
```

Example of a list of tags

We draw up an example of a list of Part Of Speech.

- Nouns

  Substantive

  Subcategories:

  - Proper noun
  - Common Nouns

- Adjective

  In some descriptions, a distributional category **predeterminer** can have the functional property of an adjective as well as other categories. In French *Seuls les enfants restent* “Only the children stay” contains the predeterminer *Seuls* which can also be seen as an adjective or a part of a determiner.

- Adverb

  The French negative particle *ne* can be seen as an adverb or as a clitic in distributional studies.

  The Adverb category is not well defined, and boundaries with other categories are often vague.

- Verb

  Possible subcategories:

  - Auxiliary
  - Main verb
  - Raising verb
  - semi-auxiliary, modal verbs

- Pronoun

  The distributional category **Clitic** is not a subcategory of pronoun because a clitic is not only a conjunct pronoun but can also be an affix, a negative adverb, etc.

  The **numerals** can be subcategories of pronouns as well as adjectives or nouns. Alternatively, the category **numeral** can also be added to the list.

- Clitic

  Possible subcategories:

  - enclitic
  - proclitic

- Conjunction

  The category **Connector** could be used instead. But in this case, adverbs have to be defined differently.

  Possible subcategories:

  - Co-ordinating Conjunction

    - Subordinating Conjunction

- Determiner

  Article is possibly a subcategory of determiner, but could also be added to the POS list. In this case, possessives, demonstratives and other words have to be defined separately.

  The distributional category **specifier** has to be defined here or separately, if used.

- Ad-position

  Preposition and **postposition** are subcategories, but could also be added to the POS list.

- Word sentence

  Interjection

- Punctuation

  Possible subcategories:

    - Weak punctuation
    - Strong punctuation

- Complementizer

- Predeterminer

- Residual

- Specifier

2.6.2 subcategories

Subcategories define classes in a POS or grammatical properties like semantic properties, morphological properties, etc.

Nevertheless, some grammatical relations do not belong to the morphosyntactical annotation and shall not be used:

- sub-categorization

- selectional restriction

- co-occurrence restriction

- semantic marker

- etc.

Sample list of morphosyntactical grammatical categories (also called grammatical markers, syntactic markers):

- gender
- number
- person
- mood
- tense
- voice
- case
- aspect
- nominal classes
- etc.