

Workpackage 2 - Guidelines for the Annotation of Definitions

Lothar Lemnitzer
Universität Tübingen

March 2006

1 Introduction

These Guidelines are one of two parts. The accompanying second part will show you how to construct local grammars for definitions. In this part we will explain how definitions which appear in your texts should be annotated.

You should also consult the *Specification of a target format for the linguistic annotation* and in particular the DTDs therein to fully understand the format of the linguistic specification.

2 What is a definition?

definition 1: a concise explanation of the meaning of a word or phrase or symbol

definition 2: A definition gives the essential semantic features of a concept as well as those that distinguish the concept from all others.

definition 3: A definition delimits or describes the meaning of a concept or term by stating the essential properties of the entities or objects denoted by that concept or term. The word or phrase to be defined is the *definiendum* and the phrase defining it is the *definiens*. For example, in the definition "a bachelor is an unmarried man", the *definiendum* is "bachelor", and the *definiens* is "unmarried man". Often, as in this example, the definition is a statement that expresses a logical equivalence between the *definiendum* and the *definiens*.

definition 4: A definition by genus and difference is one in which a word or concept that indicates a species – a specific type of item, not necessarily a biological category – is described first by a broader category, the genus, then distinguished from other items in that category by *differentia*.

There are two important points here with definitions. The first is that a definition somehow relates the term to be defined (called *definiendum* above) with a text which paraphrases the meaning of this term (in a certain context). The second is that the the defining text (called *definiens* above) specifies what kind the defined term is (by referring to a higher level concept, the *genus proximum*) and which features distinguishes it from the other objects of the same kind (by referring to its specific features, the *differentia*). To use an oversimplified example:

example 1: A dog is [an animal] (the genus proximum) which [barks] (the differentia).

Now let us go over some real life examples to round up the picture:

example 2: The TARDIS is a fictional time machine and spacecraft in the British science fiction television programme Doctor Who.

example 3: An applet is a small program or application that runs on a browser and enables additional features like animation.

example 4: Aquilaria is a genus of eight species of trees in the Thymelaeaceae native to southeast Asia.

example 5: In Tudor and Early Stuart English architecture a banqueting house is a separate building reached through pleasure gardens from the main residence, whose use is purely for entertaining.

You see from these examples, most of them taken from the Wikipedia, that the most frequent type of pattern for definitions in English is to link the defined term and the defining text with the copula verb. The copula verb can of course be used in the plural and in the past tense.

The last example is of particular interest. First, it starts with an adverbial phrase. Second, in the strict sense of a definition with *genus proximum* and *differentia specifica*, the defining text will end after the word *residence*. We recommend that you mark the adverbial phrase (*In Tudor ...*) as part of the defining text, but leave out the part of the sentence after the word *residence*. Otherwise, the local grammars which you will build from the examples will become too complicated.

We recommend that you consult some articles in the Wikipedia for your language to get a first impression of how definitions in your language look like. The "random article" button on the Wikipedia main page is particularly helpful for this task.

Definitions in learning objects might however differ from encyclopedic texts. In learning objects, authors often use definitions:

- to introduce terms and concepts which are supposed to be new to the reader;
- to give a known term a new meaning, which is specific to the theory of the author who wrote the learning object.

Furthermore, there are no such strict rules for defining as there are in the Wikipedia.

3 Why should you annotate definitions?

In the LT4EL project, we want to build software which detects definitory contexts automatically. The author who submits a text to our software will get a list of phrases which, with high probability, will be definitions or parts of it. The software will thus help authors to compile a glossary for their learning materials.

These glossaries are extremely helpful for the users of the learning materials. They can, while reading parts of the text, refer to it if they are not sure whether they understood a term correctly.

To work properly, the software needs patterns of which definitions typically consist. In the last section we explained that a definition has two parts: the defined term and the defining text. With the examples which you will provide we will learn how the *defining text* looks like. Your examples will be the basis for the abstraction which eventually will lead to a local grammar of definitory contexts for your language. We will explain this latter step in the second part of this manual. For now, it should suffice to know what we need the annotation for.

4 The input to your task

You will get a set of texts in which you will mark both the defined terms and the defining text. Be aware that this will not be just plain text. The documents which you will get have already undergone linguistic annotation. This means that each word, sentence etc. is already marked with tags. These tags provide useful linguistic information, e.g. about sentences, chunks and individual words. We will use this information later on to construct the local grammars mentioned above.

Here is an example of a linguistically annotated text:

example 6: <par lang="en" id="p1"><s id="s1"> <chunk id="c1" category="NP"><tok id="t1">
<orth>Intensive</orth><lex><base>intensive</base><ctag>Adj</ctag></lex></tok>
<tok id="t2"><orth>eLearning</orth><lex><base>eLearning</base><ctag>Ncns</ctag>
</lex></tok></chunk><chunk id="c2" category="V"></tok><tok id="t3"><tok id="t3">
<orth>helps</orth><lex><base>help</base><ctag>Vfps</ctag></lex></chunk>
... </s> </par>

This passage is the beginning of a paragraph (par). The paragraph starts with a sentence (s) and the sentence starts with a chunk (chunk). A chunk is a part of a sentence, a kind of phrase. Refer to the annotation manual of your language for further details. Words are called "tokens" (tok) in these texts. For each word/token, you see the form in which it appeared in the text (orth), the base form, i.e. the

form in which it appears in a dictionary (base), its part of speech (ctag), and, for some languages, the morpho-syntactic description of the word (msd). The morphosyntactic description informs you about the case, number, tense, person etc. of an inflected word. The morpho-syntactic description is missing from our small example because it is not relevant for English. Please read and consult the annotation manual for your language to make sure that you understand all the information.

If you go through the text, you might discover that some of the linguistic tags are simply wrong. The reason is that the linguistic annotation has been done automatically, and all automatic processes are prone to errors. Do not waste your time correcting these errors. Our software has to cope with the same errors and it will also not try to correct them. You can of course mark a word, e.g. *eLearning* as a defined term even if it received the wrong linguistic description (for example, it might have been classified as an adjective instead of a noun).

Some texts will also have keywords marked. Here is an example of an annotated text with one keyword:

```
example 7: <par lang="en" id="p1"> <s id="s1"> <chunk id="c1" category="NP">
<tok id="t1"> <orth>An</orth> <lex> <base>a</base> <ctag>Di</ctag> </lex>
</tok> <markedTerm id="m12" kw="y"><tok id="t2"><orth>eLearning</orth>
<lex> <base>eLearning</base> <ctag>Ncns</ctag> </lex> </tok>
<tok id="t3"> <orth>system</orth> <lex> <base>system</base> <ctag>Ncns</ctag>
</lex> </tok></markedTerm> </chunk> <chunk id="c2" category="V"> </tok>
<tok id="t3"> <tok id="t3"> <orth>helps</orth> <lex> <base>help</base>
<ctag>Vfps</ctag> </lex> </chunk> ... </s> .... </par>
```

Note that the keyword is enclosed in a pair of tag with the name *markedTerm*. You will use the same tag to enclose the defined term. Indeed, a keyword can also be a defined term. We will explain this in more detail in a later section.

5 How you have to annotate

Your major tasks will:

- find a definition
- mark the defined term
- mark the defining text

You should read each text twice. The first reading will help you to identify the topics of the text and to make a mental map of the text. When re-reading it you will mark the definitions in it.

For the first reading it will be extremely helpful to get the text without all the linguistic annotation. Ask your project manager for an html version of the text. It does exist, and in this first phase your reading should not be impeded by all the tags. Probably you will work with a tool with which you can switch the tags on and off. Ask your project manager for such a solution.

We recommend you to work with a printout of the html version of the text, mark the definitions with a text marker, and afterwards insert the information into the annotated text.

When you mark the definitions in the file, you have to go through the text with the tag view "switched on". The reason is that you have to add, for each defined term and each defining text, a pair of tags that marks the enclosed text as either defined term or defining text.

In short: try everything which makes your work easier. Your project manager should support you in this effort.

5.1 The *markedTerm* element for defined terms

5.1.1 How to use the element

All tokens (or sequences of tokens) which you identified as defined terms should be enclosed in a pair of tags. To demonstrate this, we example 6 above example and mark the term *eLearning* as a defined term. Note that the term is already marked as a keyword, which is quite natural. Many defined terms qualify as keywords.

```
example 8: <par lang="en" id="p1"> <s id="s1"> <chunk id="c1" category="NP">
<tok id="t1"> <orth>Intensive</orth> <lex> <base>intensive</base> <ctag>Adj</ctag>
```

```
</lex> </tok> <markedTerm id="m45" kw="y" dt="y">
<orth>eLearning</orth> <lex> <base>eLearning</base> <ctag>Ncns</ctag>
</lex> </tok></markedTerm> </chunk> <chunk id="c2" category="V"> </tok>
<tok id="t3"> <tok id="t3"> <orth>helps</orth> <lex> <base>help</base>
<ctag>Vfps</ctag> </lex> </chunk> ... </s> .... </par>
```

Note that the tag that marks the word as a keyword and / or defined term starts **before** the *tok*-tag starts and ends **after** the *tok*-tag ends. Everything inside the *tok*-tag, i.e. the orthographic form, the base form and the part of speech information, are now "inside" the *markedTerm*-tag. This is as it should be.

Defined terms can also consist of more than one word. These words must then follow in a sequence, as in the following example.

```
example 9: <par lang="en" id="p1"> <s id="s1"> <chunk id="c1" category="NP">
<tok id="t1"> <orth>An</orth> <lex> <base>a</base> <ctag>Di<ctag> </lex>
</tok> <markedTerm id="m46" dt="y"><tok id="t2"> <orth>eLearning</orth>
<lex> <base>eLearning</base> <ctag>Ncns</ctag> </lex> </tok>
<tok id="t3"> <orth>system</orth> <lex> <base>system</base> <ctag>Ncns</ctag>
</lex> </tok></markedTerm> </chunk> <chunk id="c2" category="V"> </tok>
<tok id="t3"> <tok id="t3"> <orth>helps</orth> <lex> <base>help</base>
<ctag>Vfps</ctag> </lex> </chunk> ... </s> .... </par>
```

Note that the *markedTerm*-tag now encloses two tokens: *eLearning* and *system*.

We will now explain the format of the *markedTerm* tag.

5.1.2 The attributes of the *markedTerm* element

The tag you will use to enclose the defined terms is called *markedTerm*. Why did we use this name and not simply *defined term*? The reason is that through the course of this project we will annotate both keywords and defined terms. That is why we chose *markedTerm* as a hypernym for both *keyword* and *defined term*.

You should specify a unique identifier (*id*) for each *markedTerm*. The value of the identifier should start with a *m* and be followed by a number. It is easiest if you use sequential numbers. You can start with 'm1' in each new document. However, if the term is already marked as a keyword, then it has already an identifier assigned to it. In this case you do not have to do this.

When you want to point out that the marked term is a defined term, you write *dt="y"*.

You will see two further attributes called *status* and *comment*. These attributes are for use with keywords exclusively, so you should not use them. You should also ignore them if the word is marked as a keyword and they are already specified.

5.2 The *definingText* element

The defining text should always be confined within the borders of a sentence. In other words: the defining text should always be a sentence or part of a sentence.

If the defining text spans more than one sentence, a case which will rarely occur, you have to split the text in two or more parts and enclose each part in separate tags. The following example is a simplified example of this case. Note that we apply only those tags which are necessary to explain the structure.

```
example 10: <s ...><definingText id="dt12a" def="m12" part="1">A
<markedTerm id="m12" dt="y">dog</markedTerm> is a kind of animal.</definingText>
</s><s><definingText id="dt12b" def="m12" part="2">
It has four legs and usually barks.</definingText></s>
```

You see that the *definingText* element can enclose a sequence of tokens or chunks (not longer than a sentence). In example 10 above, the first part of the defining text encloses also the defined term. This is correct in the case that the defined term is part of the sentence. If the defined term is **not** part of the sentence, as in the following example, then the defining text does not include the defined term:

```
example 11: <markedTerm id="m47" dt="yes">assessment item</markedTerm>:
<definingText id="dt13" def="m47">A questionnaire or measurable activity used to determine
if the learner has mastered a learning objective.</definingText>
```

With this example, you also see which attributes you should use with the *definingText* element.

First, you should specify a unique identifier (*id*) for each definingText element. The value of the identifier should start with *dt* and be followed by numbers and/or letters. It is easiest if you use sequential numbers. You can start with 'dt1' in each new document.

With the *def* attribute you specify to which definingTerm the definingText is related. You should use the **id** of the markedTerm here to indicate the relation. The relation might be clear if the defining text surrounds the defined term. But this is not always the case. Therefore we decided to cross-reference the defining text with the defined term.

You can use the following additional attributes: *status* and *comment*.

With the attribute *status* you can mark how sure you are that this is a definition. Currently you have the following three choices:

- if you are sure that the marked text is a proper definition, then you do not use the attribute status at all
- if you are in doubt, but tend to accept the text as a definition, then you use *?* as the value of the status attribute.
- if you are in doubt and tend to reject the text as a definition, then you use *??* as the value of the status attribute.

The marking of dubious and borderline cases helps us to build a grammar of definitions. If an example does not fit easily some of the rules and is dubious anyway, then we can safely ignore it.

The last attribute which you can use is *comment*. You can use this attribute to give any comment you want. We can use the comments afterwards as a kind of "diary" of your work. So do not hesitate to use the *comment*-attribute if you want to make any remarks or comments. If you do not want to comment, that is fine too – just leave the *comment* attribute out.

6 Difficult cases

In this section we will discuss difficult and borderline cases as the result of your feedback.

In the moment we have only three issues to raise:

- Do not try to annotate defined terms which consists of two words which do not follow one another. In the moment I do not have an example for this, but during the work there might occur such cases. If there are two discontinuous words which in your opinion form one single defined term, please annotate the whole sequence as one single defined term (and add a comment, please).
- The same holds for the defining text – it should be one continuous sequence of text. If there are insertions, e.g. texts in parenthesis, they have to be treated as part of the defining text.
- Please do not alter the orthographic form of any token. It might occur that a word is written in *UPPERCASE* or in *MiXed CaSe* or an accent is missing. Do not worry about it. On the contrary: for the validation of our software we need the keywords exactly as they are in the text.