

Workpackage 2 - Guidelines for the Annotation of Keywords

Lothar Lemnitzer
Universität Tübingen

March 2006

1 Introduction

2 What is a keyword?

I will start with some definitions of *keyword*. I will discuss them afterwards.

Keyword definition 1: The word or words that relate to a particular topic.

Keyword definition 2: A significant word in the abstract, title, subject headings (descriptors), or text of an entry in a bibliographic database which can be used as a search term in a free-text or natural language search.

Keyword definition 3: A word or phrase that a user believes is relevant to the information he or she is seeking.

Keyword definition 4: The user enters keywords into an online search form. The search engine then examines each record in its database to find those documents that match the keywords. A keyword search (as opposed to a concept search) is a search for documents containing one or more words specified by the user.

The first definition focusses on the semantic qualities of a keyword. According to this, it has to represent the topic of theme of the document with which it is linked. Let us assume that we have a collection of documents about eLearning. Is *eLearning* a good keyword for any document in this collection? From the point of view of definition 1, yes. It represents the topic of probably any text in the collection. From another point of view which is expressed in the latter definitions, it is not.

The second definition stresses the point that a keyword belongs to the descriptive data for a text and appears quite often in some salient passages of the text, e.g. the abstract or the title. We will see later that we do not follow completely this definition. We assume that keywords appear prominently at salient places, but they may appear in other parts of the document as well. It will not suffice to scan these salient parts of a document in search for keywords. The definition introduces the word *significant*. This narrows down the set of "good" keywords. To come back to our example, the word *eLearning* will not be significant for a collection of texts about eLearning. However, once these texts become part of a larger collection of texts which include other topics, *eLearning* will be a good keyword. You see that the wider context of the application or usage scenario matters. Our advice for now is to mark these keywords because they represent the topic of the text(s) somehow regardless in which context they will be used.

The third definition focusses on the user's perspective. In order for the user to use "our" keywords, (s)he has to know this word and to consider it relevant. Instead of using "our" keyword *eLearning*, he might use *web-based learning* (a rough synonym), *learning* (a hyperonym) or *web-based training* (a hyponym). A good application should take this behaviour into account by linking all these words by lexical-semantic relations.

What you can do for now is to mark all related terms in text as keywords (i.e. do not miss a keyword because it is "just" a synonym of another word you have already marked).

The fourth definition describes the use of keywords in a typical information retrieval scenario: keywords, which are provided by the user, are somehow matched with the documents which contain one or more of these terms.

3 Why should you annotate keywords?

In the LT4EL project, we want to build software which detects keyword automatically. We want the computer to find, for any text, the words which are salient and related to the topic(s) of this text.

This software should be, in the ideal case, as good as a human who scans a text and finds and marks the keywords manually. In order to achieve this, and to check how good the software is already, we need manually annotated texts. In other words: your performance will be the reference against which the performance of our software will be evaluated.

4 The input to your task

You will get from your colleagues a set of texts in which you will mark the keywords. Be aware that this will not be just plain text. The documents which you will get have undergone linguistic annotation. This means that each word, sentence etc. is already marked with tags. These tags provide useful linguistic information, e.g. about the part of speech of a word.

Here is an example of a linguistically annotated text:

```
example 1: <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 the 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.

The information might also be helpful for your task. A noun will in most cases be a better keyword than an adjective or adverb, and you can glean information about the category of each word from these tags.

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 also been done automatically, and all automatic processes are prone to errors. Do not waste your time correcting these errors. Our automatic keyword finding 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* mark as a keyword even if it received the wrong linguistic description (for example, it might have been classified as an adjective instead of a noun).

5 How you have to annotate

Your major tasks will be to:

- find keywords
- mark them

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. It might be even useful to use pen and paper to graphically depict the topics of it.

For the first reading it will be extremely helpful to get the text without all the 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 the best solution.

When you mark the terms, you have to go through the text with the tag view "switched on". The reason is that you have to add, for each keyword, a pair of tags that marks the enclosed word as a keyword.

You should work with a printout of the html version of the text, mark the keywords with a text marker, and after this insert the information into the annotated text.

In the html-version of the text you might encounter words which are emphasized (i.e. written in italics or bold font, or underlined). You can also easily see where the headings (titles and subtitles) are. You should pay special attention to emphasized words because the highlighting might signal that these words have high importance for the text. These highlighted parts and the headings are the places where you might find the best keywords.

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

The last thing you have to do, once you have identified the keywords, is: enclose them in a pair of tags. We will detail this in next section.

5.1 The *markedTerm* element

All tokens (or sequences of tokens) which you have identified as keywords should be enclosed in a pair of tags. To demonstrate this, we copy the above example and make the term *eLearning* a keyword.

```
example 2: <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="m12" kw="y" status="?" comment=
"I am not quite sure whether this is a good keyword, it is probably too general"><tok id="t2">
<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 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. Keywords which span more than one token are treated the same as single word token. Let us look at another example. We will now mark the term *eLearning system* as a keyword.

```
example 3: <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="m13" 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 *markedTerm*-tag now encloses two tokens: *eLearning* and *system*.

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

5.2 The attributes of the *markedTerm* element

The tag you will use to enclose the keywords is called *markedTerm*. Why did we use this name and not simply *keyword*? The reason is that in the course of this project we will not only annotate keywords but also words which are defined. To be more precise, we will annotate definitions which consist of a defined term and its definition. We give one example for this.

example 4: Now let us define **part of speech**. This term *signifies one of the traditional grammatical categories, like noun, verb, adjective, etc.*

The bold text in the example is the defined term and the text in italics is the definition.

Thus we have two types of terms to be marked in our documents: keywords and defined terms. This is why we have chosen to call the tag *markedTerm*.

As you have seen in the examples above, the opening tag allows you to add information in the form of several attributes. Look again at the opening tag in example 2 above, repeated here:

example 5: `<markedTerm id="m45" kw="y" status="?" comment="I am not quite sure whether this is a good keyword, it is probably too general" >`

You should specify a unique identifier (*id*) for each *markedTerm*. The value of the identifier should start with an *m* and be followed by a number. It is easiest if you use sequential numbers. You can start with 'm1' in each new document.

When you want to point out that the marked term is a keyword, you write *kw="y"*. With the attribute *status* you can mark how sure you are that this is a keyword. Currently you have the following three choices:

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

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. In example 3 above, we used the *markedTerm* with neither the *status* nor the *comment* attribute because we were quite sure that the marked term is a keyword and we did not have any further comments.

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 invent keywords. All the keywords should be marked directly in the text, which implies that they are there. If you miss a keyword which would best describe the text but is not there, you can still add a comment which points us to this additional keyword.
- Do not try to annotate keywords 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 keyword, please annotate them as separate keywords, i.e. enclose each token in a separate *markedTerm*-tag (and add a comment, please).
- Please do not alter the orthographic form of any token. It might occur that a keyword 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.