

Tagging Advice Text

This text has been written in order to help users of the programme *Text Tool* when tagging texts. The first thing to bear in mind is that there are two modes for tagging with this programme:

- 1) *Automatic tagging* (Shift-Ctrl-F3)
- 2) *Interactive tagging* (F3)

Both tagging options are to be found in the *Tools* menu of the programme. If you like, you can use the shortcuts indicated in brackets above. Before starting tagging it is strongly advisable to consider what texts you wish to tag in what manner. The procedures available here are mechanical aids. They do not make any decisions on the contents of texts. Successful tagging is only possible if the user designs a sensible tagging system and the words in the texts to be tagged are unambiguous. Think first of all what your goal in tagging is. It is a time-consuming procedure, but can be greatly accelerated by functions such as the present two. Nonetheless, tagging is as good as you make it. Whether it yields the results you want depends on how you go about it.

1) Automatic tagging

This is the easiest form of tagging and will work best when the forms to be tagged in a text are unambiguous. For instance, if you assign a tag `_INTERROG_ADV` to a form like *which* then the results are liable to be incorrect if your text(s) contains a sentence like *The car which was stolen* where *which* is a relative pronoun. So be careful with automatic tagging if there are polyfunctional forms in your text(s).

Before automatic tagging you must write (or adapt) a list file in which the tags and the forms (words or parts of words) which are to be assigned these tags are specified. Please consult the supplied file `Tagging_Automatic_Test.lst` to see how this is done. When the *Automatic Tagging* window opens, you can select a list of tags (you can have several on disk if you wish). You must select the rows in table with tags and forms before tagging (or deselect "Only tag subset of forms"). You then select tags in the grid which appears in the usual way, by pressing either the Shift-key or the Ctrl-key along any of the arrow keys, just as in any *Windows* programme. You may also say if tags are only to be attached to whole words and you can choose to confirm each tag. The latter is useful if you wish to skip forms to which a tag does not apply. You can also specify whether tags are to be highlighted (by red marking). This is only retained if you stored the tagged file(s) to disk as an RTF file(s).

2) Interactive tagging

Tagging a text consists of attaching a grammatical label as a suffix to a word form. This is an important aspect of preparing text corpora for later linguistic retrieval tasks, either by the compiler(s) or others who have access to these corpora. However, tagging texts is time-consuming and its accuracy depends on the nature of the texts and the tagging scheme used.

With the option *Interactive tagging* the user decides what category of label is to be suffixed to what word forms. Once this operation has been carried out grammatical information can be retrieved from the texts of a corpus by referencing the tags suffixed onto words which have been tagged. In general you cannot reference semantic information in a corpus, i.e. a tagged corpus is primarily intended for retrieving morphological and possibly syntactic information.

To tag words/forms in the current text, you first select (get from disk) a set of words/forms to fill the list *Words to tag*. You then get a list for *Tags to use*. Now select some words/forms (in the *Words to tag* list) which are to be assigned to one of the tags. The words/forms are entered in the sub-list on the left by clicking on the button *Import checked forms*. Choose a tag to be attached to each of these forms. This appears in the top-left corner. Start the tagging process by clicking on *Start*.

The maximum number of tags and of input forms is 512 items in each case. The lists can be created with *Corpus Presenter Text Tool* itself and stored to disk for later use on this level of the programme. Take a look at the file `Tagging_Interactive_Test_(Tags).lst` which can be used for tags. Try the file `Tagging_Interactive_Test_(Word_Forms).lst` for a set of words/forms to be tagged. Use the file `Chaucer_Prologue.rtf` as a text with words/forms to be tagged. The files just mentioned can be downloaded by clicking on the following link. You can now experiment to see how the tagging actually works.

Tagging parameters

- 1) *Words or strings* Specifies if only words – or any string – can be tagged.
- 2) *Case-sensitive search* Determines whether small and capital letters are distinguished.
- 3) *Automatic or manual* Here you can decide whether *Corpus Presenter Text Tool* halts at each find and asks the user to confirm whether a form is to be tagged. Note that with manual tagging you can also edit the finds in the current text as you proceed.

Raymond Hickey
March 2006