GLOSSA Administrator's Quickstart

Lars Nygaard, Anders Nøklestad

17th April 2008

Abstract

This is a brief introduction to making a corpus searchable with the GLOSSA interface. For further information, refer to the GLOSSA *Administrators Manual.*

1 CWB files

The corpus input files must be encoded in the CWB input format:

- one token per line, with token annotation in tab-separated columns
- structural annotation in XML-like format, always starting on a new line

For example:

```
<text id=”text1”>
<s id=”s1”>
I      I      pron     pers
am     be     verb     pres
a      a      art
fish   fish   noun     sg
</s>
```

It is recommended that you use the 'text' and 's' structural attributes, with ids. You can use any structural annotations, but that will require some more configuration.

This file must be converted into CWB binary format using cwb-encode and cwb-makeall. GLOSSA has a utility script that simplifies this process: bin/cqpify_monoling.pl (but it must be modified to each corpus).

1.1 Alignment

CWBs input format for alignment data consists of a file (for each language pair) containing tab-separated pairs of token start-stop positions of aligned regions regions. For example:

```
SAMNO_NORSK     s       SAMNO_SAMISK    s
2       11      4       11      2:1
12      16      12      16      1:1
17      28      17      31      1:2
```

This means that the region 2-11 in the corpus SAMNO_NORSK is aligned with the region 4-11 in the corpus SAMNO_SAMISK, and that this is an alignment of 2 s-regions to 1 s-region.

The program bin/create_cwb_alg.pl can create the CWB input file from an CES alignment file.

2 MySQL database

The MySQL database is optional.

You can create three bibliographic tables. Firstly, the main table: <projectname>text, and secondly the two subsidiary table <projectname>author and <projectname>class, for encoding one-to-many relationships.

In addition, you can precompile lexical statistics, in the table <projectname>_<corpusname>lexstat, and allow user annotation with the tables <projectname>annotations.

3 Configuration files

To make a new Glossa project (ie. a new interface), you need to go through several steps, some of them fairly complex. However, the easiest way to approach it is to simply find another existing project that is similar to yours, copy the configuration, and make a few changes.

3.1 PHP

- Add some lines to index_dev.php.
- Create the <projectname>.inc (and optinally <projectname>_cred.inc php files.

3.2 Config files

- Create the js/<projectname>.conf.js file
- Create the $GLOSSA/dat/<projectname>/cgi.conf file
- (Optionally:) Create the $GLOSSA/dat/<projectname>/meta.conf file

3.3 Menu file

- Create the menu definition javascript file: js/<projectname>.js file
- Edit js/dynamic_form_dev.js

The meny definition file can either be copied from another corpus with similar annotation, or created with the bin/create_menu_item.pl program.