CQP(web) installation @ corptedig-glif

This is a rough guideline to index, compress, and install corpora on the corptedig-glif CQPweb interface, written by Thomas Brochhagen.

Local setup

Software

You will need some tools from the the IMS Open Corpus Workbench. Get the latest **CWB main package** as well as the **Perl API & Support packages**. The latter is optional but recommended.

Data preparation

For POS-tagged data, the cwb tools expect: one or multiple tab-separated vertical files (`.vrt`-files), where each row is a single word, together with any additional information for it (POS-tag, lemma, e.g.).

Additionally, each `.vrt`-file should be enclosed by a `<text>`-tag. The tools do a good job at recognizing XML information, if you tell them to expect it. A minimal example of a ready to be processed file looks as follows:

```
<text>
She             PRP     she
sells           VRZ     sell
seashells       NNS     seashell
</text>
```

You may want to at least add an `id` attribute to the tags. In this way you can later restrict your queries to particular fragments and see where your matches come from. The file will then look like this:

```
<text id='tales_of_sales_by_the_sea' year='2019'>
She                                     PRP     she
sells                                   VRZ     sell
seashells                               NNS     seashell
</text>
```

If you have no information to add, a single `<text>`-tag enclosing all the text will do.

Indexing

We'll call our corpus `SEA`. Assume that its vertical files are in `/corpus/vrt/`. Create a folder to store the indexed corpus:

```
mdir /corpus/binaries
```

Assuming the second format, including an `id` and a `year` tag, generate the corpus:

```
cwb-encode -c utf8 -d corpus/binaries -F corpus/vrt -R corpus/sea -xsB -P pos -P lemma -S text:0+id+yea
```

In the order in which the flags appear, this tells CWB to encode the corpus as utf8; to store the indexed corpus at `corpus/binaries`; to take all data from `corpus/vrt`; to store the corpus' registry as `corpus/sea`;

to parse the XML; to expect, after the first column, reserved for words, a POS tag and a lemma; and to parse `<text>`-tags with their information appropriately. Further XML-markup needs to be added with more `-S` flags.

Now, index and compress the corpus:

```
cwb-make -r corpus/ -V SEA
```

Skip the `-V` flag if your corpus is big and pay attention to upper- and lower-case conventions.

Testing

See whether if everything works locally:

```
cqp -r corpus/
show corpora;
CDE;
"seashells";
```