The IMS Open Corpus Workbench (CWB)

WWW: http://cwb.sourceforge.net/

Manuels:

a. Corpus Encoding Tutorial
http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial.pdf

b. CQP Query Language Tutorial
http://cwb.sourceforge.net/files/CQP_Tutorial.pdf

Préparer un corpus pour l'exploiter

On prépare le corpus pour l'exploiter avec CWB en 2 étapes:

1. Encodage: la commande `cwb-encode`
2. Indexation et compression: la commande `cwb-make`

1. **`cwb-encode`**

Options

Paramètres obligatoires	
`-f/F`	fichier/dossier contenant les fichiers à encoder
`-d`	le dossier où on veut sauvegarder le corpus
`-R`	création automatique du fichier de registre dans le catalogue de registre, par exemple: `registry/PATROLOGIE`
Paramètres pour les fichiers XML	
`-x`	le mode compatible avec XML
`-s`	néglige les espaces
`-B`	supprime les espaces en token
Attributs de corpus	
`-S`	attribut structurel dont le nom correspond au nom d'élément XML; on enchaîne les attributs avec la syntaxe +ATTR
`-P`	attributs de tokens (e.g. *lemme*, *pos* etc.); la première colonne est automatiquement nommée *word*

Exemple d'usage

```
cwb-encode
   -xsB
   -F corpora/nonencode/patrologie/XML
   -d corpora/encode/patrologie
   -R corpora/registre/patrologie
   -P pos -P lemma
   -S s -S text:0+id+periode+type+titre+auteur+vol
```

2. **cwb-make**

Options

-V	nom du corpus
-r	dossier du registre

Exemple d'usage

```
cwb-make -V PATROLOGIA -r corpora/registre/
```

3. Pour tester le corpus

```
cwb-describe-corpus -r corpora/registre PATROLOGIE
```

Exploiter un corpus avec CWB

1. Lancement de CWB: `cqp`

Options

-r	dossier du registre
-e	mode interactif

Exemple d'usage

```
cqp -r corpora/registre/ -e
```

2. L'affichage des corpora installés dans le système

show corpora;	montre les corpora installés dans le système
info NOM_DE_CORPUS;	informations sur le corpus donné
show cd;	info sur le contexte
NOM_DE_CORPUS;	entre dans le corpus donné

3. Recherche simple sur token

```
"mot"
[word = "mot"]
```

Options

%c	Ignorer la casse
%d	Ignorer les signes diacritiques

Exemple d'usage

```
[word = "temp.+"%cd]
[lemma = "tempus"]
[pos = "SUB"]
```

4. Opérateurs logiques

=	égal	**&**	AND
!=	non égal	|	OR

Exemples d'usage

```
[lemma != "t.+" & pos !="SUB.*"];
```

5. Chercher des séquences de mots

Exemples d'usage

```
[word="tempus"][word="fugit"];
[word="tempus"][][pos="VBE"];
[word="tempus"][]{0,1}[pos="VBE"];
```

6. Assigner les variables

Exemples d'usage

a. afficher les derniers résultats

```
Last;
```

b. attribuer les résultats à une variable

```
Tempus = [lemma="tempus"][]{0,3}[pos="VBE"];
```

c. évaluer la taille du sous-corpus

```
size Tempus;
```

d. afficher les variables

```
show named;
```

e. supprimer les variables

```
discard Aetas;
```

f. sauvegarder les résultats

```
set DataDirectory (.)
cat Tempus > „tempus.txt”;
```

7. Limiter la localisation et l’étendue à une phrase etc.

Exemples d'usage

a. en spécifiant l’élément XML

```
TempusXML = <s>[lemma="tempus"][]{0,3}[pos="VBE"];
```

b. avec le mot `within`

```
TempusWithin = [lemma="tempus"][]{0,3}[pos="VBE"] within s;
```

c. avec `lbound()`, `rbound()`

```
TempusFonc = [lemma="tempus" & lbound(s)][]{0,3}[pos="VBE"];
```

8. Trier, réduire, grouper, calculer, tabulariser

a. commande `sort`

Syntaxe

```
sort VARIABLE by ATTRIBUT_POSITIONNEL on ANCRE COMMENT_TRIER;
```

Exemples d'usage

```
Tempus = [lemma = "tempus"];
sort Tempus by word;
sort Tempus by word %cd on match[1];
sort Tempus by word %cd on match[-1] ..match[-5] ;
sort Tempus by word %cd on match[-2] descending reverse;
```

b. commande `reduce` - réduire le nombre d'occurrences

Syntaxe

```
reduce VARIABLE to POURCENTAGE|NOMBRE_DE_OCCURRENCES
```

Exemples d'usage

```
Tempus1 = Tempus;
reduce Tempus1 to 1%;
reduce Tempus1 to 1000;
```

c. commande `count` - calculer le nombre d'occurrences

Syntaxe

```
count VARIABLE by ATTRIBUT_POSITIONNEL [on ANCRE];
```

NB. (+) Permet d'utiliser `%cd`. (-) Ne permet pas de créer des tableaux croisés!

Exemples d'usage

```
count Tempus by lemma%cd on match[1];
```

d. commande `group` - grouper les occurrences selon des critères

Syntaxe

```
group VARIABLE ANCRE1 ATTR_POS_1 by ATTR_POS_2 on ANCRE2;
```

NB. (+) Permet de créer des tableaux croisée! (-) Ne permet pas d'utiliser `%cd`.

Exemples d'usage

```
group Tempus match[1] word by match[2] pos;
```

e. commande `tabulate` - créer des tableaux

Syntaxe

```
tabulate Tempus ANCRE1 ATTR_POS_1, ANCRE2 ATTR_POS_2 ...
```

Exemples d'usage

```
tabulate Tempus match[1] pos, match[2] pos, match pos;
tabulate Tempus match[5]..match[1] word, match word;
```

f. script extérieur `cwb-scan-corpus` - créer des tableaux au dehors de `cqp`

Fonction

```
Extract n-gram frequency information from CWB corpus
```

Syntaxe

```
cwb-scan-corpus -r DOSSIER_REGISTRE -C NOM_DU_CORPUS
   ATTR_POS_1[+POSITION_A_DROIT][=/"REGEX"/[cd]] ...
```

Exemples d'usage

```
cwb-scan-corpus -r -C PATROLOGIE word+0 word+1 word+2
   lemma+2=/"tempus"/ text_auteur=/"Beda"/
```