TEI et analyse linguistique

Novembre 2011

Quelques éléments TEI

- `<s>` (phrase)
- `<cl>` (proposition)
- `<ph>` (syntagme)
- `<w>` (mot, token)
- `<m>` (morphème)
- `<c>` (caractère)
- `<pc>` (symbole de ponsctuation)

Le cas particulier de l'oral

La notion de phrase n'est pas forcément très pertinente...
La notion de symbole de ponctuation pas forcément non plus
On risque donc d'en reveir souvent à des `<seg>` ou des `<span>` auquels on donnera une interprétation et dans certains cas à des `<div>`

Un point de repère

Je n'entrerai pas dans des choses très compliquées !
Un (bon) exemple de corpus annoté linguistiquement en TEI : NKJP (corpus national du polonais)
Références :

- Stand-off TEI Annotation: the case of the National Corpus of Polish (proc. ling. annot. workshop, ACL-IJCNLP 2009)
- http://nlp.ipipan.waw.pl/TEI4NKJP/

Segmentation en mots

- on utilise `<w>`
- éventuellement, `<w>` dans `<w>` (pomme de terre)

Attributs de `<w>` :

- ‘lemma’
- ‘lemmaRef’ (pointe vers une définition)

L'annotation en parties du discours

La plus courament pratiquée (peu coûteuse) et bénéfices immédiats !
Ce qui semblerait simple :

- un attribut 'pos' sur `<w>`

n'est pas autorisé !

L'attribut 'ana'

Ce qui est autorisé, c'est l'attribut 'ana' (analyse), du module 'analysis'

- Attention : cet attribut peut-être porté par un (très) grand nombre d'éléments et peut donc potentiellement avoir des interprétations (très) diverses !
- remarque : le type de 'ana' est un indice ! 'ana' est de type data.pointer.

La (ma ?) philosophie TEI

On essaie de garder le maximum d'information à l'intérieur du fichier TEI

- Si on a un jeu d'étiquettes morphologiques, il est probablement défini quelque part !
- pourquoi pas dans le fichier TEI ?
- evidemment, *dans* veut seulement dire, on se donne l'accès vers...

<interp> et <interpGrp>

<interp> permet de donner le sens d'une interprétation.
<interpGrp> permet de regrouper de telles interprétations.
Exemple :

```
<interpGrp>
  <desc>Etiquettes morphosyntaxiques</desc>
  <interp  xml:id="ART">Articles</interp>
  <interp  xml:id="N">Noms communs</interp>
  <interp  xml:id="ADJ">Adjectifs</interp>
</interpGrp>
```

On peut maintenant faire le lien !


```
<w  ana="#ART">le</w>
<w  ana="#ADJ">petit</w>
<w  ana="#N">chat</w>
<interpGrp>
  <desc>Etiquettes morphosyntaxiques</desc>
  <interp  xml:id="ART">Articles</interp>
  <interp  xml:id="N">Noms communs</interp>
  <interp  xml:id="ADJ">Adjectifs</interp>
</interpGrp>
```

Et si on veut être encore plus précis

On peut faire le lien avec une base terminologique comme ISOCAT
On utilise alors l'attribut 'sameAs'

```
<interp  xml:id="ART"
   sameAs="http://www.isocat.org/datcat/DC-1892">Articles</i
```

Annotation linguitique ?

Beaucoup d'élements TEI peuvent correspondre à une analyse linguistique !
par exemple :

- `<name>`
- `<rs>`
- `<date>`
- etc.

Pas besoin donc de se demander comment annoter des entités nommées !

Exemple : coréférence


```
<p>
  <name  key="P">Pierre</name>
  <rs  key="P">se</rs> demanda si <rs  key="P">il</rs>...
</p>
```

Remarque, on aurait pu également utiliser ‘xml:id’ et ‘ref’

Un mot sur les structures de traits

Il y a en TEI un module ‘iso-fs’, commun entre la TEI et l’ISO pour définir des structures de traits. Un des intérêts est de pouvoir définir de façon modulaire des jeux d’étiquettes. Exemple, on voudrait annoter en morpho-syntaxe en descendant jusqu’aux genres, nombres, temps, personnes, etc.

```
<f  xml:id="pluriel">
  <binary  value="true"/>
</f>
<f  xml:id="nomCommun"  name="cat">
  <string>Nom commun</string>
</f>
<fs  xml:id="N:p"  feats="#nomCommun #pluriel"/>
```

```
<w  ana="#N:p">chaises</w>
```

Concrètement

- Les structures de traits se définissent dans un `<fDecl>`, lui-même dans un `<fsDecl>`, le tout dans `<fsdDecl>`
- et enfin, tout ça va dans le `<teiHeader>/<encodingDesc>`

L'éternel problème des structures concurentes

Quand on commence à multiplier les annotations, il y a toute chance qu'on finisse par tomber sur les hiérarchies multiples
Le sujet a fait couler des *litres* d'encre (électronique ou non).
D'un point de vue pratique, `<span>` permet souvent de s'en sortir

```
<span xml:id="s1" to="#a1" interp="#qqchose"/>du texte
<anchor xml:id="a1"/>
```

Inconvénient du

 est aussi peu informatif que <seg> ! On peut (il faut) préciser son sens via :

- ‘type’
- <interp>
- <join> (apparement déprécié et servirait plutôt à joindre des <seg>)

Autres niveaux d'analyse linguistique

A ma connaissance, peu de corpus annotés en syntaxe, désambiguïsation de sens ou sémantique encodés en TEI
Une exception notable : NKJP

- Annotation "stand-off", astucieuse (via xi:include)
- A ce niveau de complexité, le stand-off est probablement la seule solution viable
- Séparation de la structure habituelle (`<div>`, `<p>`, etc..) pour annoter commodément le contenu textuel

Utilisation extensive des `<seg>`, et des `<link>`
exemple :

```
<link
    xml:id="link17"
    type="subject"
    targets="ann_morphosyntax.xml#seg78 #group43"/>
```

Remarques NKJP

- Les niveaux d'analyse s'incluent les uns les autres
- gestion des disjonctions (ex morphosyntaxe) via des `<choice>`
- éventuellement, le niveau d'analyse suivant, ne reprend que l'un des termes du choix

En pratique

- On n'annote pas (entièrement) à la main ce genre de choses !
- Il faut des outils (ex : treetagger), et des scripts
- cas particulier de treetagger, mode 'sgml' permet d'ignorer le balisage (et donc en fait de le conserver).
- Autre solution : TXM

Exercice

- Encoder à ma main en parties du discours un petit bout d'un de vos fichiers TEI
- Si possible, expliciter les parties du discours par :
  - soit : un `<interpGrp>`
  - soit des structures de traits