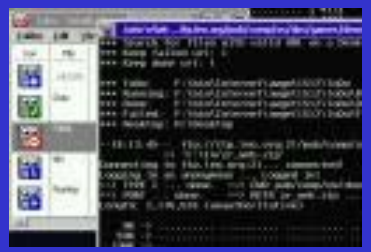
Programmation et projet encadré

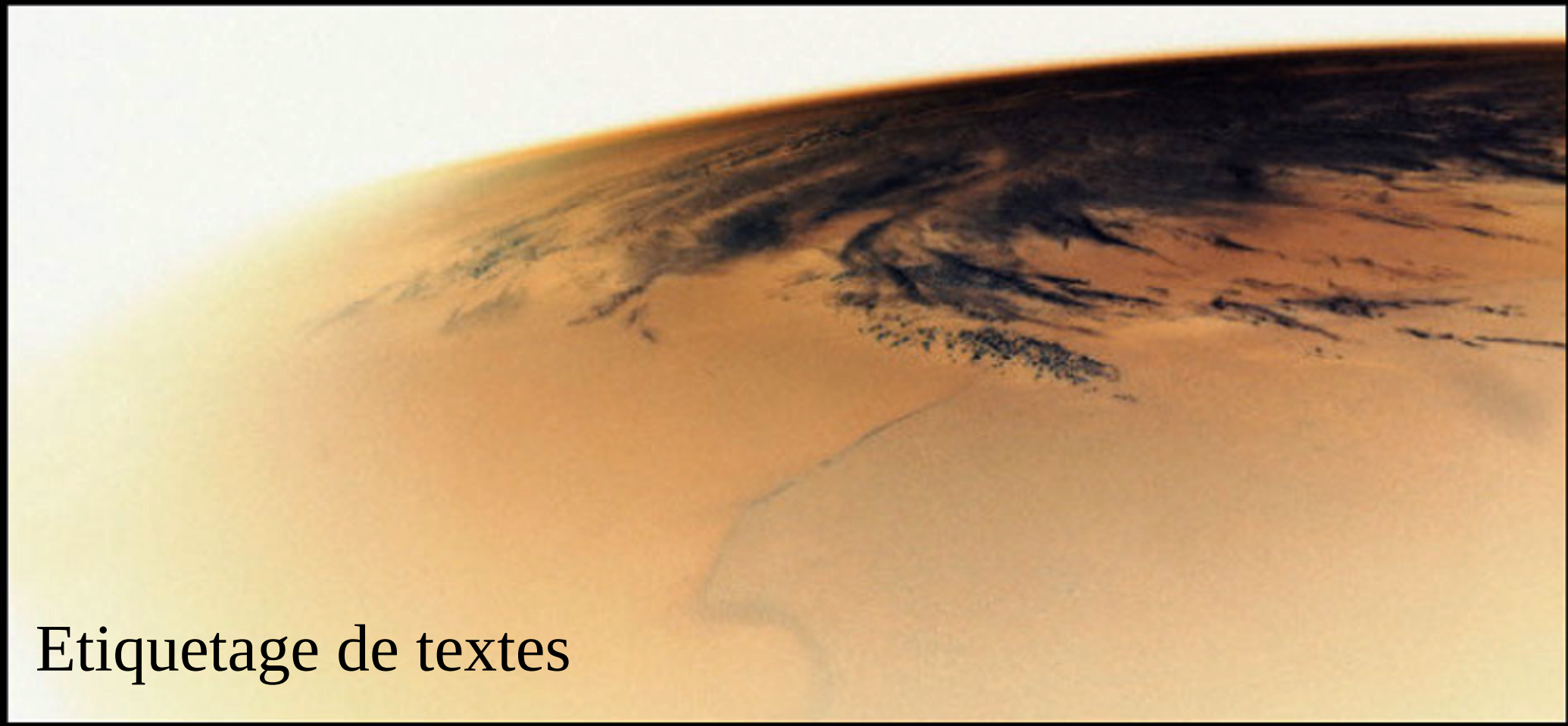
Boîte à outils

Série 2 : étiquetage



Bibliographie

- « Annotation automatique de corpus : panorama et état de la technique », Jean Véronis, *Ingénierie des langues* (ch. 4), J.M. Pierrel éditeur, Lavoisier Hermès, 2000
  - (mail à SF pour récupérer ce texte au format PDF)



Etiquetage de textes

Boîte à outils : série 2

Objectif

- Etiqueter un texte
- Automatisation

sommaire

Etiquetage morpho-syntaxique

- Étant donné un ensemble de couples (graphie, CMS) et un texte, choisir pour chacun des mots (graphies) du texte parmi ses CMS associées celle(s) qui correspond(ent) au contexte.
  - suppose que « celle(s) qui correspond(ent) au contexte » ait un sens, par exemple « confirmée(s) par un expert humain »
- Plusieurs approches possibles :
  - À bases de règles : Le « tagger de Brill »
  - Probabiliste : Chaîne de Markov cachées (HMM)

Principe général

- Soit la phrase : *Jean a mangé des pommes*.
- Etape 1 : *segmentation*
  - **U1** = *Jean*, **U2** = *a mangé*, **U3** = *des*, **U4** = *pommes*, **U5** = . (point)
- Etape 2 : *étiquetage morpho-syntaxique*
  - on associe des ***Informations morpho-syntaxiques*** aux **Ui (i = 1, 2, 3, ...)**, comme par exemple :
    - **U1** = *Jean* : ***Informations morpho-syntaxiques*** : nom propre, masculin, singulier.
    - **U2** = *a mangé* : ***Forme lemmatisée*** : *manger*. ***Informations morpho-syntaxiques*** : verbe, passé composé, indicatif, 3ème personne, singulier, constructions : transitif, ...

Les étiqueteurs (utilisés dans ce cours)

- Cordial
  - Version Université 6.00
  - www.synapse-fr.com
  - Brève présentation avec exercices de prise en main de Cordial 6 Université)
- TreeTagger
  - http://www.ims.uni-stuttgart.de/projekte/corplex/DecisionTreeTagger-de.html
  - *The TreeTagger is a tool for annotating text with part-of-speech and lemma information which has been developed within the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger has been successfully used to tag German, English, French, Italian, Greek and old French texts and is easily adaptable to other languages if a lexicon and a manually tagged training corpus are available.*
  - A NOTER :
    - une version de TreeTagger *online* est disponible à cette adresse :
    - http://www.cele.nottingham.ac.uk/~ccztk/treetagger.php
      - application en Flash permettant d'étiqueter des textes de moins de 1000 mots

Travail personnel [*série 2*]:
étiquetage des contenus des fils

- Objectif :
  - Vous devez construire un programme qui parcourt une arborescence de fichiers et applique un traitement d'étiquetage sur chacun des fichiers rencontrés au moment du parcours
  - En sortie, le programme doit construire un fichier structuré (XML) contenant une trace du traitement réalisé sur les fichiers
  - Application :
    - **Ressources fournies :**
      - Une arborescence de fils RSS
        - Les 2 transparents suivants montrent l 'allure de l'arborescence et le contenu des fils
      - Un squelette minimal du programme de parcours
  - **Traitement** :
    - étiqueter les contenus textuels des balises DESCRIPTION et TITLE (*i.e.* votre programme de filtrage construit précédemment)
    - IMPORTANT : on « conservera » aussi le titre de la « rubrique » du fil (balise *title* sous *channel* cf présentation du corpus)

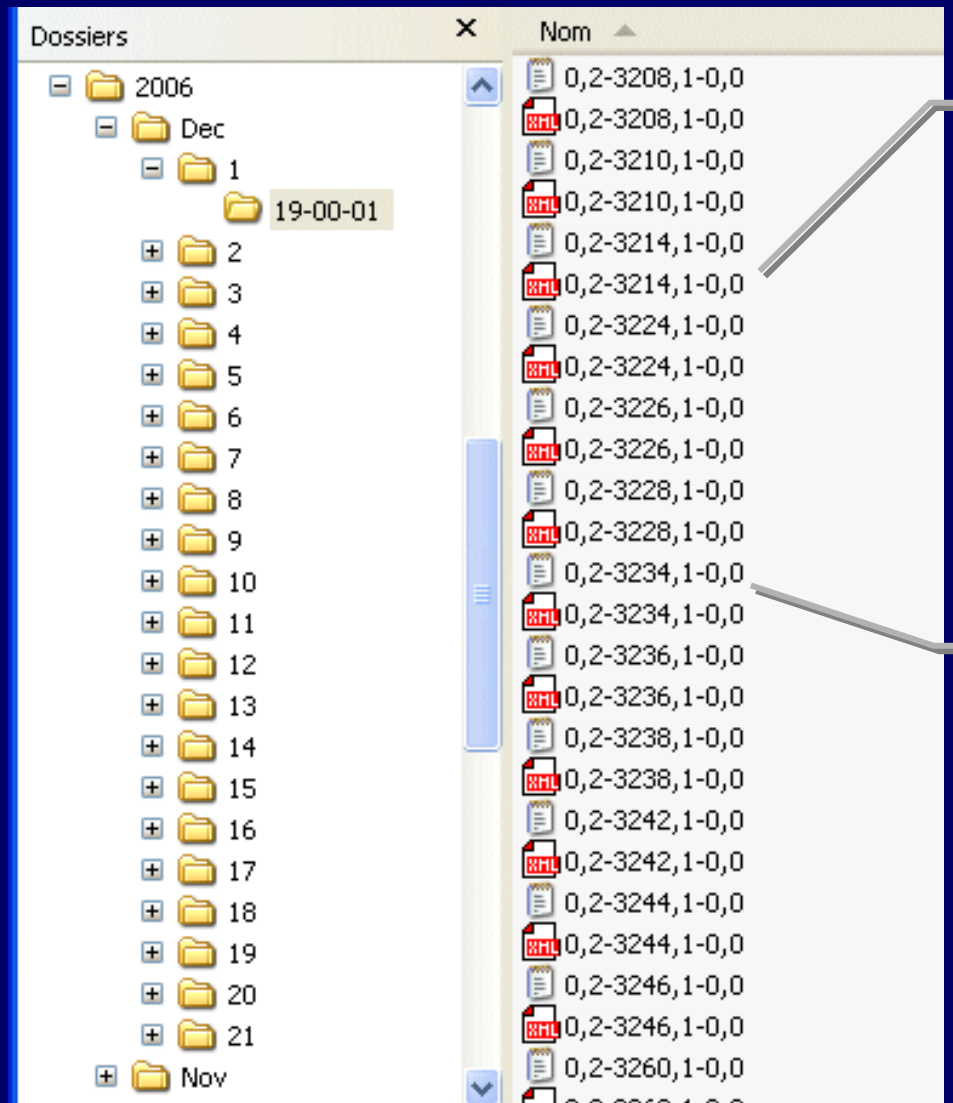
Les 17 fils RSS du journal *Le Monde* sur la période du 20/11/2006 au 21/12/2006.
Ces 17 fils ont été archivés tous les jours à 19h sur cette période.

Rubrique	Fils RSS	Fils format texte (Lexico3)
A la Une	0,2-3208,1-0,0.xml	0,2-3208,1-0,0.txt
International	0,2-3210,1-0,0.xml	0,2-3210,1-0,0.txt
Europe	0,2-3214,1-0,0.xml	0,2-3214,1-0,0.txt
France	0,2-3224,1-0,0.xml	0,2-3224,1-0,0.txt
Société	0,2-3226,1-0,0.xml	0,2-3226,1-0,0.txt
Environnement	0,2-3228,1-0,0.xml	0,2-3228,1-0,0.txt
Entreprises	0,2-3234,1-0,0.xml	0,2-3234,1-0,0.txt
Médias	0,2-3236,1-0,0.xml	0,2-3236,1-0,0.txt
Rendez-vous	0,2-3238,1-0,0.xml	0,2-3238,1-0,0.txt
Sports	0,2-3242,1-0,0.xml	0,2-3242,1-0,0.txt
Sciences	0,2-3244,1-0,0.xml	0,2-3244,1-0,0.txt
Culture	0,2-3246,1-0,0.xml	0,2-3246,1-0,0.txt
Technologies	0,2-651865,1-0,0.xml	0,2-651865,1-0,0.txt
Cinéma	0,2-3476,1-0,0.xml	0,2-3476,1-0,0.txt
Voyages	0,2-3546,1-0,0.xml	0,2-3546,1-0,0.txt
Livres	0,2-3260,1-0,0.xml	0,2-3260,1-0,0.txt
Présidentielle 2007	0,57-0,64-823353,0.xml	0,57-0,64-823353,0.txt

SOURCE : http://www.lemonde.fr/web/rss/0,48-0,1-0,0.html

L'arbre des fils

(Le Monde : 1 mois dans les fils)



Les fils au format XML sont stockés dans un dossier horodaté du type : 2006/Mois/Jour/Heure

ATTENTION : seuls les fils au format XML seront à traiter !!!!

Le contenu des fils

« Rubrique du fil »

Balise TITLE

Balise DESCRIPTION

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <rss version="2.0" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  - <channel>
      <title>Le Monde.fr : A la Une</title>
      <link>http://www.lemonde.fr</link>
      <description>Toute l'actualité au moment de la connexion</description>
      <copyright>Copyright Le Monde.fr</copyright>
    - <image>
        <url>http://medias.lemonde.fr/mmpub/img/lgo/lemondefr_rss.gif</url>
        <title>Le Monde.fr</title>
        <link>http://www.lemonde.fr</link>
      </image>
      <pubDate>Fri, 02 Dec 2005 23:00:00 GMT</pubDate>
    - <item>
        <title>M. Chirac veut favoriser l'entrée en France des Africains
        hautement qualifiés</title>
        <link>http://www.lemonde.fr/web/article/0,1-0@2-3212,36-
        717310,0.html</link>
        <description>La France facilitera la délivrance de visas de longue
        durée à entrées multiples pour les entrepreneurs, cadres,
        chercheurs, professeurs et artistes africains, a annoncé samedi à
        Bamako le président français.</description>
        <pubDate>Sat, 03 Dec 2005 18:41:08 GMT</pubDate>
        <guid isPermaLink="false">http://www.lemonde.fr/web/article/0,1-
        0@2-3212,36-717310,0.html</guid>
      </item>
    - <item>
        <title>L'affaire des vols secrets de la CIA en Europe s'étend à
        l'Allemagne</title>
        <link>http://www.lemonde.fr/web/article/0,1-0@2-3214,36-
        717303,0.html</link>
        <description>Selon &#34;Der Spiegel&#34;, plus de 430 vols secrets
        transportant des prisonniers soupçonnés de terrorisme sont passés
        par l&#39;Allemagne, où Condoleezza Rice est attendue
        lundi.</description>
        <pubDate>Sat, 03 Dec 2005 15:03:41 GMT</pubDate>
        <guid isPermaLink="false">http://www.lemonde.fr/web/article/0,1-
        0@2-3214,36-717303,0.html</guid>
      </item>
    - <item>
        <title>Fiscalité : comment le gouvernement mène une réforme en
        catimini de l'épargne</title>
```

Le Monde en fil

Le programme d'étiquetage (1)

- Application avec *treetagger*
  - Mode d'emploi (*cf* README) :

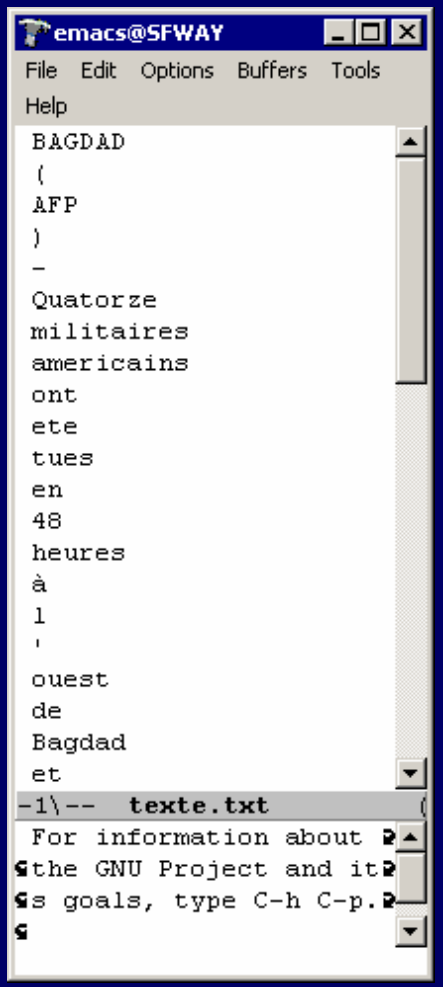
```
tree-tagger [options] <parametres> <textein> <texteout>
```

  - Le premier argument est le fichier paramètre (ici *french.par* dans le répertoire *lib*)…
  - Le deuxième argument est le texte à étiqueter (**avec un mot par ligne**)…
  - Le troisième argument est le nom du fichier de sortie…

Le programme d'étiquetage (2)

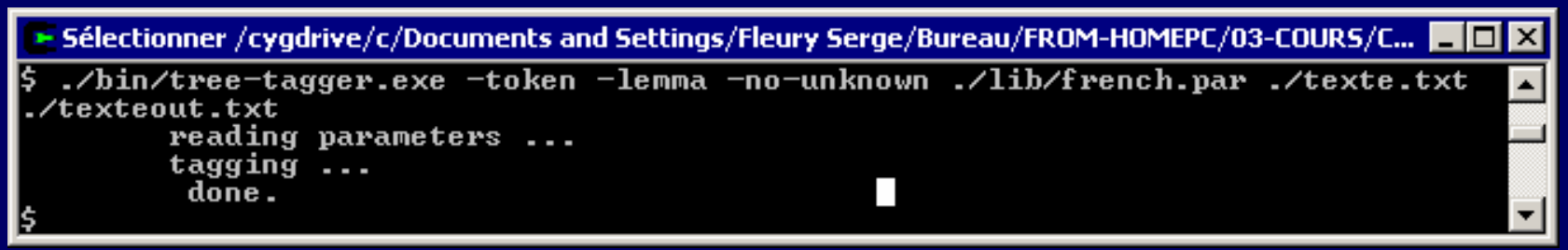
- Exemple de traitement
  - Fichier à étiqueter

texte.txt →



(*i.e* le contenu d'une balise DESCRIPTION d'un fil RSS/AFP (Nb mots < 200))

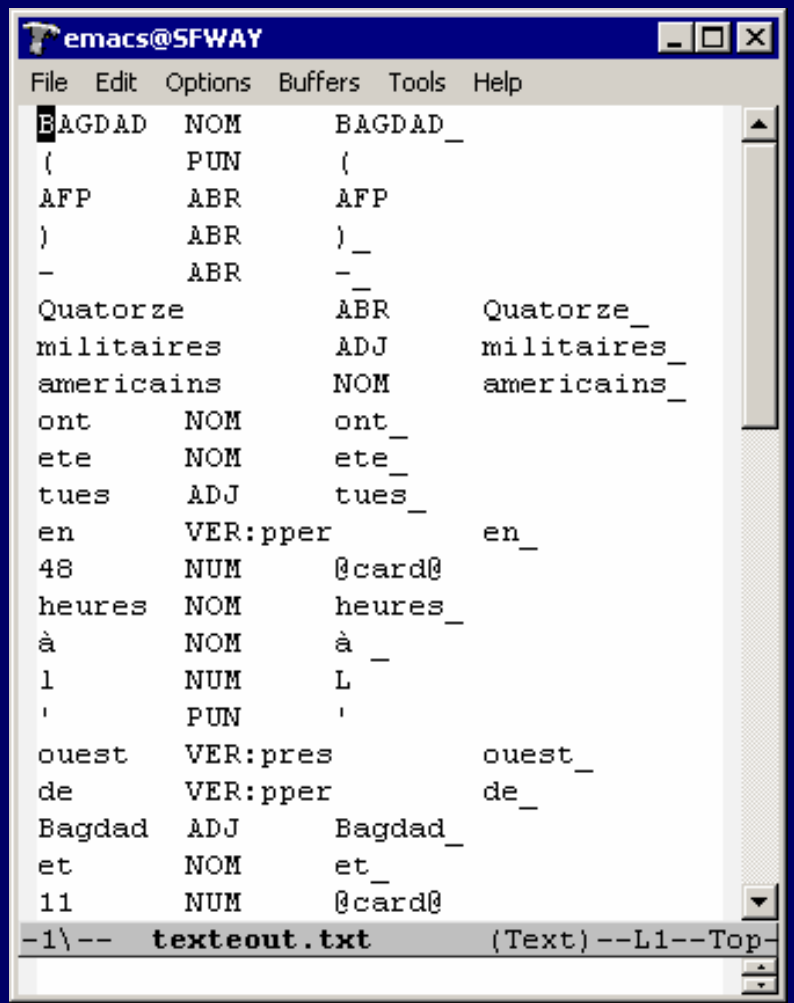
- Lancement du programme :
  - Paramètre : `french.par`
  - IN : `texte.txt`
  - OUT : `texteout.txt`
  - **Options** : `-token, -lemma, -no-unknown`



sommaire

Le programme d'étiquetage (3)

- Résultat du traitement



sommaire

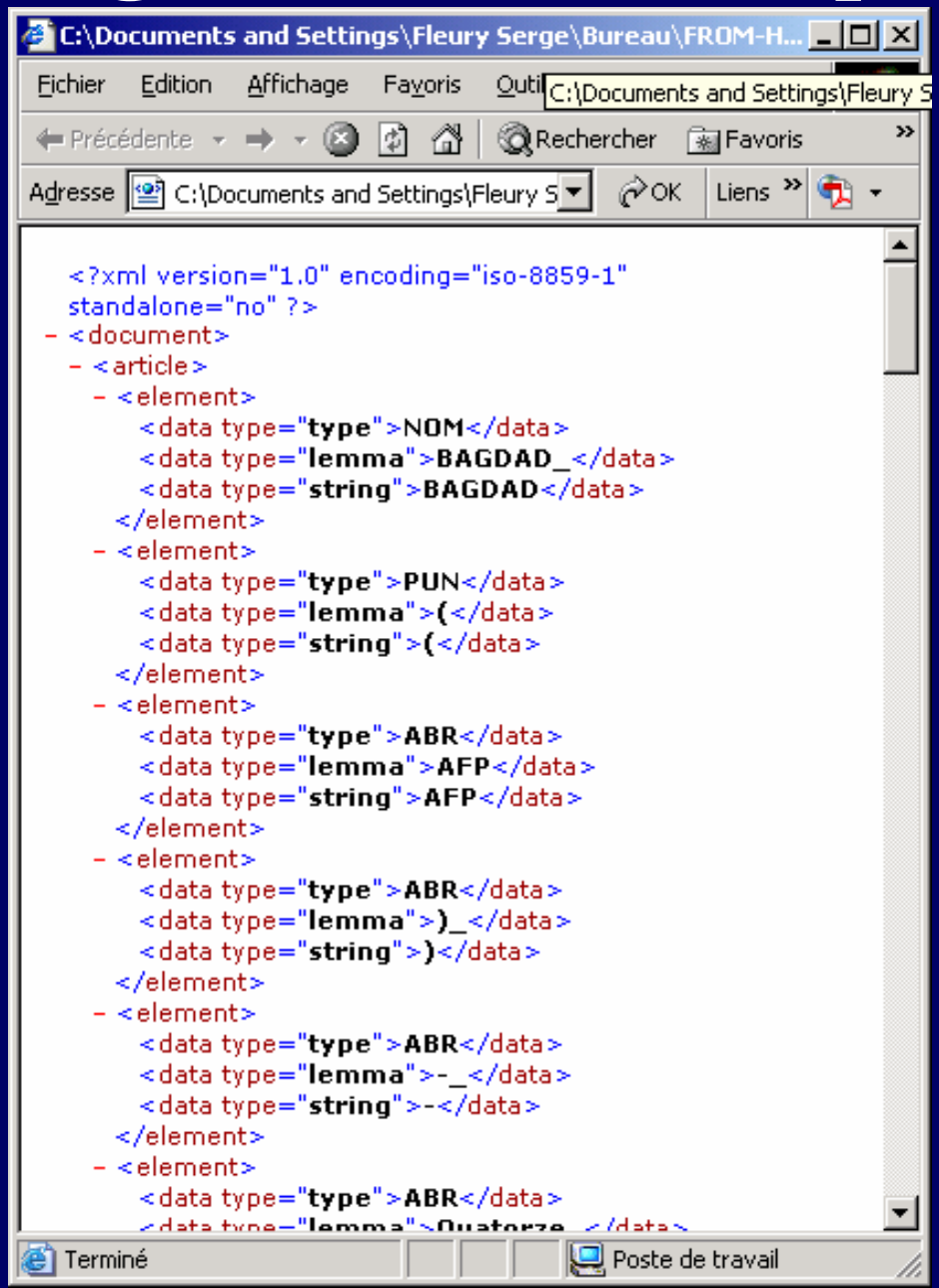
Le programme d'étiquetage (4)

- *Raffinement* : un script perl transforme la sortie du *treetagger* au format XML
  - Usage :

```
perl treetagger2xml sortietreetagger.txt
```

=> Création d'un fichier en sortie qui a pour nom : `sortietreetagger.txt.xml`

Le programme d'étiquetage (5)



sommaire

Votre travail (*partie 1*)…

- A partir du programme parcours-arborescence-fichiers.pl (*vu dans BàO série 1*)
  - Intégrez le traitement d'étiquetage avec *treetagger* sur les contenus des balises DESCRIPTION de tous les fichiers de votre « arborescence de fils ».
    - **Solution 1** : traitement d étiquetage « à la volée »
      - ***Extraire le contenu des balises DESCRIPTION*** . Dans votre script, utilisez la fonction « rechercher » sur un patron décrivant la balise visée, il est aisé ensuite de récupérer le contenu de la balise : `/<description>([^<]+)<\/description>/ ($1 = le contenu de la balise)`
      - ***Le reformater*** : un mot par ligne. Plusieurs solutions possibles : par exemple, *via* rechercher/remplacer (remplacer les « frontières » de mots par un retour à la ligne

        `s/([^tout-caractère-dans-un-mot])/\n$1\n/;` (à compléter…)
      - ***L 'étiqueter***. Dans votre script : Écrire le résultat du reformatage dans un fichier sur lequel vous lancez le *treetagger* puis vous récupérez le résultat de l'étiquetage en ouvrant le fichier associé
    - **Solution 2** : On peut aussi traiter globalement l'étiquetage en réalisant au préalable l'extraction des zones textuelles à étiqueter. Intégrez dans le script les processus d'extraction du contenu des balises DESCRIPTION et d'écriture en sortie de l'ensemble de ces contenus textuels.
  - Vous devrez construire en sortie un fichier structuré regroupant l'ensemble des traitements d'étiquetage
    - Exemple de sortie : projet-etiquetage-2\SORTIE-etiquetage.xml
    - On pourra reprendre et modifier le code de *treetagger2xml*

Votre travail (*partie 2*)…

- A partir du programme parcours-arborescence-fichiers.pl
  - Extraire dans un fichier les contenus des balises TITLE et DESCRIPTION de tous les fichiers de votre « arborescence de fils »
  - Etiqueter ce fichier avec Cordial (pour obtenir au minimum FORME/CATEGORIE/LEMME, *cf* mode d'emploi : par exemple celui-ci)
  - Essayez de construire en sortie (*via* un script Perl) un fichier structuré regroupant et reformatant l'ensemble des traitements d'étiquetage produit par Cordial (on pourra s'inspirer de celui construit dans la partie 1)

Votre travail (*raffinements*)…

- Construire une feuille de style XSLT pour afficher les résultats produits au format HTML
  - On pourra par exemple reformater l'arborescence XML disponible dans les fichiers de sortie pour reconstruire les structures phrastiques initiales dans lesquelles les étiquettes sont désormais intégrées
    - Exemple de sortie construite à partir du résultat produit dans la partie 1 : Sortie XML avec feuille XSLT : reformatage en phrase concaténant l'ensemble des mots de manière suivante `forme[lemme|cat]` (lien vers feuille de style XSL)
  - On pourra aussi mettre en avant certaines parties du discours
    - Exemple de sortie construite à partir du résultat produit dans la partie 1 : Sortie XML avec feuille de style XSLT : la catégorie NOM est mise en rouge (lien vers feuille XSLT)
- Vous nous envoyez par mail, une archive contenant une page web avec votre nom et le contenu de vos programmes (et leurs sorties)

Ressources pour démarrer…

- Le programme :
  - parcours-arborescence-fichiers.pl
- L'arborescence des fils :
  - Le corpus à utiliser sera disponible au LABO
- Un répertoire contenant le programme *treetagger* pour windows (à dézipper dans un dossier nommé *treetagger* par exemple)
  - projet-etiquetage-2\treetagger-win32.zip
  - (contient aussi le programme *projet-etiquetage-2\treetagger2xml.pl*)
- Vous envoyez (à SF) par mail, une archive contenant une page web avec votre nom et le contenu de votre programme (et ses sorties)