SL0720X
Étiquetage morpho-syntaxique

Franck Sajous (CLLE-ERSS) - sajous@univ-tlse2.fr
http://w3.erss.univ-tlse2.fr/membre/fsajous/

Notes :
- toutes les urls données dans ce document, ainsi que les textes à étiqueter sont présents sous forme de liens cliquables à l'adresse : http://w3.erss.univ-tlse2.fr/membre/fsajous/SDL/SL0720X/
- au cours des séances, vous utiliserez systématiquement l'éditeur de texte **TextPad** et le navigateur **Firefox**.

1 À la découverte de Treetagger

1. À la racine du disque `C:` doit se trouver un répertoire nommé `SL0720`. Si ce n'est pas le cas, créez-le : ouvrez le poste de travail, clic-droit sur `C:`, puis *ouvrir*. Clic-droit dans la fenêtre, puis *Nouveau/dossier*.
2. Ouvrez l'éditeur de texte Textpad (icône sur le bureau ou menu *Démarrer/Tous les programmes*).
3. Saisissez quelques phrases et enregistrez le fichier dans le répertoire `C:\SL0720`.
4. Ouvrez Firefox et saisissez l'url ci-dessous dans la barre d'adresse : http://cental.fltr.ucl.ac.be/treetagger/
(ou cherchez *treetagger cental louvain* dans un moteur de recherche) ;
5. Cliquez sur le bouton *Parcourir* en face de *texte à étiqueter*, puis sélectionnez le fichier texte que vous avez créé dans la boîte de dialogue qui s'affiche et validez.
6. Cliquez sur envoyer (ou *submit query*).
7. Cliquez sur le lien *Télécharger le fichier étiqueté* (clic-droit pour enregistrer le fichier sur votre machine, clic-gauche pour l'afficher dans le navigateur).
8. Lorsque vous affichez le résultat dans votre navigateur, il est possible que les caractères comportant des diacritiques s'affichent mal. Dans ce cas, allez dans le menu *affichage* de votre navigateur, puis *Encodage des caractères/Occidental ISO-8859-1*.
9. Observez le résultat : de quelles informations dispose-t-on en plus du texte initial ? Quelles sont, selon vous, les étapes préalables à l'étiquetage morpho-syntaxique ?

2 Installation de l'interface

1. En salle machine, TreeTagger doit être installé dans le répertoire `C:\TreeTagger`. Si ce n'est pas le cas, installez-le en vous reportant à l'annexe D.

2. Affichez dans un navigateur la page suivante :
`http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm`
3. Dans l'item *Graphical interface* de la section *Links*, cliquez sur *Download the Windows interface to the tagger program.*
4. Enregistrez le fichier `.exe` dans le répertoire `C:\TreeTagger\bin`.
5. Créez un raccourci vers `wintreetagger.exe` depuis le bureau (cliquer-glisser avec le bouton droit de la souris depuis le fichier `wintreetagger.exe` vers le bureau, puis *créer un raccourci...*)

3 Étiquetage

3.1 Prise en main

1. Dans TextPad, créez un document texte et saisissez quelques phrases. Enregistrez le document.
2. Dans l'interface WinTreeTagger, sélectionnez la bonne langue, puis le fichier d'entrée (cliquez dans la zone de texte sous *Input File*).
3. Sous *lexical information*, cochez *none*.
4. Pour indiquer le fichier de sortie (résultat de l'étiquetage), cliquez dans la zone de texte sous *Output File*, puis :
   - sélectionnez un fichier existant (dans ce cas, le fichier sera écrasé) ;
   - ou entrez manuellement le nom d'un nouveau fichier (en le suffixant par `.txt`).
5. Cliquez sur *Run*.
6. Ouvrez le fichier résultat.

3.2 Casimir

Téléchargez le fichier `gloubi-boulga.txt` [1] et étiquetez-le.

Quels sont les formats particuliers pour les lemmes ?

Repérez les erreurs et classez-les en différents types et niveaux d'analyse. Repérez des incohérences dans l'étiquetage (des données similaires qui conduisent à des étiquetages différents).

Quels sont les différentes étiquettes pour les marques de ponctuations ? À quoi correspondent-elles ? Quels sont les cas particuliers dans la colonne des lemmes ?

Dans quels cas l'étiquetage des mots inconnus paraît correct/incorrect ? Cet étiquetage paraît-il cohérent (systématique) ?

3.3 Intervenir dans l'étiquetage

3.3.1 Pré-étiquetage

Il est possible de fournir à TreeTagger un texte pour lequel certains éléments ont été pré-étiquetés : on impose ainsi que tel ou tel token ait, dans un contexte

1. Ce texte provient de l'article Wikipédia :
`http://fr.wikipedia.org/wiki/Gloubi-boulga`

donné, tel ou tel lemme ou étiquette. Il faut pour cela produire, en entrée de TreeTagger, un fichier déjà segmenté (un token par ligne). Les lignes contenant les tokens que l'on veut pré-étiqueter doivent avoir le format suivant :
`token TAB étiquette`
ou
`token TAB étiquette TAB lemme`
(TAB représente une tabulation).

Reprendre la phrase *« Des soirées destinées aux adulescents combinant cosplay, rediffusion de dessins animés des années 1980 et fête étaient appelées Gloubi-boulga »* de l'exemple précédent, tokenisez-la à la main (*i.e.* mettez-la dans un fichier au format un token par ligne) [2]. Attention, dans le format "un token par ligne", les tokens ne doivent être ni précédés, ni suivis d'espace. Étiquetez-là avec l'option suivante : *Tokenization Options/none*. Répétez l'opération en faisant en sorte que *dessins animés* constitue un seul token. Observez le résultat. Ajoutez, dans le fichier d'entrée, l'étiquette et le lemme de *dessins animés* (chaque information étant séparée par une tabulation). Cochez *Input options/Manual tags have lemma*. Observez le résultat.

3.3.2 Lexique auxiliaire

La documentation de TreeTagger (cf. fichier `README.txt`, à la racine du répertoire `TreeTagger`) précise qu'il est possible d'utiliser un lexique auxiliaire dont elle précise le format. Un extrait de la documentation est reproduit ci-dessous :

```
Further optional command line arguments:
* -lex <f>: The file <f> contains additional lexicon entries to be used
  by the tagger. The file format is identical to the format of the lexicon
  argument of the training program (see below).

  <f>: name of a file which contains the fullform lexicon. Each line
  of the lexicon corresponds to one word form and contains the word form
  and a sequence of tag-lemma pairs. Each tag is preceded by a tab character
  and each lemma is preceded by a blank or tab character.
  Example:

aback RB aback
abacuses NNS abacus
abandon VB abandon VBP abandon
abandoned JJ abandoned VBD abandon VBN abandon
abandoning VBG abandon
```

Considérez l'exemple précédent (version non tokenizée) et créez un fichier lexique qui contienne les informations relatives aux néologismes présents dans la phrase. Pour étiqueter la version non tokenizée, sélectionnez à nouveau *built-in* dans *Tokenization Options*. Pour utiliser le lexique que vous venez de créer, cochez, dans *Tagging Options*, la case *Auxiliary lexicon*, cliquez dans le champ

2. Il est possible également d'utiliser le tokenizer fourni avec TreeTagger : reportez-vous pour cela l'annexe E.

texte à droite et sélectionnez votre fichier lexique. Lancez l'étiquetage et observez.

3.3.3 Pré-étiquetage et lexique auxiliaire

La documentation concernant l'utilisation d'un lexique auxiliaire précise que l'on ne peut pas inclure dans ce lexique d'unité polylexicale telle que *dessin animé*. Pourquoi, selon vous ?
Reprenez le fichier tokenizé de la section 3.3.1 et étiquetez-là en utilisant conjointement le lexique créé à la section 3.3.2. Veillez à sélectionner dans l'interface les options adéquates de format d'entrée, de tokenisation et d'étiquetage.

3.4 Étiquetage en anglais

Téléchargez le fichier `holy_grail.txt` [3] et étiquetez-le.

Comparez le jeu d'étiquettes pour l'anglais à celui du français. Comparez également les types d'erreurs comises entre les deux langues.

4 Comparaison d'étiqueteurs

Téléchargez le fichier `dormeur_duval.txt` et étiquetez-le avec TreeTagger [4].

1. Les deux premiers vers étiquetés par `Cordial` sont représentés figure 1. Quelles informations `Cordial` apporte-t-il en plus de `TreeTagger` ?
2. La version étiquetée (forme, lemme et partie du discours) par Cordial analyseur du *dormeur du val* est donnée dans l'annexe C. Quelle(s) différence(s) observez-vous ?
3. Relevez des erreurs d'étiquetage (ou résultats inattendus) pour chaque étiqueteur et comparez avec l'étiquetage de l'autre.

5 Étiquetage de textes balisés

Téléchargez le fichier « Extrait du corpus Air France » et décompressez-le. Visualisez-le dans TextPad, puis étiquez-le. Observez le résultat. Quel est le problème ?

Recommencez l'opération en cochant *Input Options/SGML tags present* et observez la différence.

3. Extrait issu du site *Unofficial Monty Python Home Page :*
http://www.mwscomp.com/python.html

4. Cet exemple est emprunté à Benoît Habert, *Instruments et ressources électroniques pour le français*. Ophrys (« L'Essentiel Français »), 2005

```
      ==== DEBUT DE PHRASE ====
1     C'     ce     PDS     Pd-..n  1|1   S     1     est          Principale
2     est    être   VINDP3S Vmip3s  2     V     1     est          Principale
3     un     un     DETIMS  Da-ms-i 4|4   B     1     est          Principale
4     trou   trou   NCMS    Ncms    4|4   B     1     est          Principale     ouverture/opening,unobstructed space
5     de     de     PREP    Sp      6|4   B     1     est          Principale
6     verdure verdure NCFS   Ncfs    6|4   B     1     est          Principale
7     où     où     PRI     Ptr-..- 7|7   -     2     chante       Relative
8     chante chanter VINDP3S Vmip3s  8     V     2     chante       Relative       pour les oiseaux/sing,warble
9     une    un     DETIFS  Da-fs-i 10|10 T     2     chante       Relative
10    rivière rivière NCFS   Ncfs    10|10 T     2     chante       Relative
      ===== FIN DE PHRASE =====
      ==== DEBUT DE PHRASE ====
1     Accrochant     accrocher      VPARPRES      Vmpp--  1     -     1          Indépendante    suspendre/hang,suspend
2     follement      follement      ADV     Rgp   -       -     1          Indépendante
3     aux    à le    DETDPIG Da-.p-d 4|4   -     1             Indépendante
4     herbes herbe   NCFP    Ncfp    4|4   -     1             Indépendante   plante/grass
5     des    de le   DETDPIG Da-.p-i 6|4   -     1             Indépendante
6     haillons       haillon NCMP    Ncmp  6|4   -     1          Indépendante
      ===== FIN DE PHRASE =====
```

FIGURE 1 – *Le dormeur du val* étiqueté par Cordial (deux premiers vers)

A Jeu d'étiquettes pour le Français

ABR	abreviation
ADJ	adjective
ADV	adverb
DET:ART	article
DET:POS	possessive pronoun (ma, ta, ...)
INT	interjection
KON	conjunction
NAM	proper name
NOM	noun
NUM	numeral
PRO	pronoun
PRO:DEM	demonstrative pronoun
PRO:IND	indefinite pronoun
PRO:PER	personal pronoun
PRO:POS	possessive pronoun (mien, tien, ...)
PRO:REL	relative pronoun
PRP	preposition
PRP:det	preposition plus article (au, du, aux, des)
PUN	punctuation
PUN:cit	punctuation citation
SENT	sentence tag
SYM	symbol
VER:cond	verb conditional
VER:futu	verb futur
VER:impe	verb imperative
VER:impf	verb imperfect
VER:infi	verb infinitive
VER:pper	verb past participle
VER:ppre	verb present participle
VER:pres	verb present
VER:simp	verb simple past
VER:subi	verb subjunctive imperfect
VER:subp	verb subjunctive present

B Jeu d'étiquettes pour l'Anglais : Penn Treebank Tagset

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

C *Le dormeur du val* étiqueté par Cordial analyseur

```
==== DEBUT DE PHRASE ====
C'      ce      PDS
est     être    VINDP3S
un      un      DETIMS
trou    trou    NCMS
de      de      PREP
verdure verdure NCFS
où      où      PRI
chante  chanter VINDP3S
une     un      DETIFS
rivière rivière NCFS
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Accrochant  accrocher  VPARPRES
follement  follement  ADV
aux     à le    DETDPIG
herbes  herbe   NCFP
des     de le   DETDPIG
haillons        haillon NCMP
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
D'      de      PREP
argent  argent  ADJINV
;       ;       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
où      où      PRI
le      le      DETDMS
soleil  soleil  NCMS
,       ,       PCTFAIB
de      de      PREP
la      le      DETDFS
montagne  montagne  NCFS
fière   fier    ADJFS
,       ,       PCTFAIB
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Luit    luire   VINDP3S
:       :       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
c'      ce      PDS
est     être    VINDP3S
un      un      DETIMS
petit   petit   ADJMS
val     val     NCMS
qui     qui     PRI
mousse  mousser VINDP3S
de      de      PREP
rayons  rayon   NCMP
.       .       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Un      un      DETIMS
soldat  soldat  NCMS
jeune   jeune   ADJSIG
,       ,       PCTFAIB
bouche  bouche  NCFS
ouverte ouvert  ADJFS
,       ,       PCTFAIB
tête nue        tête nue        ADV
,       ,       PCTFAIB
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Et      et      COO
la      le      DETDFS
nuque   nuque   NCFS
baignant        baigner VPARPRES
dans    dans    PREP
le      le      DETDMS
frais   frais   ADJMIN
cresson cresson NCMS
bleu    bleu    ADJINV
,       ,       PCTFAIB
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Dort    dormir  VINDP3S
;       ;       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
il      il      PPER3S
est     être    VINDP3S
étendu  étendre VPARPMS
dans    dans    PREP
l'      le      DETDFS
herbe   herbe   NCFS
,       ,       PCTFAIB
sous    sous    PREP
la      le      DETDFS
nue     nue     NCFS
,       ,       PCTFAIB
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Pâle    pâle    ADJSIG
dans    dans    PREP
son     son     DETPOSS
```

```
lit     lit     NCMS
vert    vert    ADJMS
où      où      PRI
la      le      DETDFS
lumière lumière NCFS
pleut   pleuvoir        VINDP3S
.       .       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Les     le      DETDPIG
pieds   pied    NCMP
dans    dans    PREP
les     le      DETDPIG
glaïeuls        glaïeul NCMP
,       ,       PCTFAIB
il      il      PPER3S
dort    dormir  VINDP3S
.       .       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Souriant  sourire  VPARPRES
comme   comme   SUB
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Sourirait  sourire  VCONP3S
un      un      DETIMS
enfant  enfant  NCSIG
malade  malade  ADJSIG
,       ,       PCTFAIB
il      il      PPER3S
fait un somme   sourire VINDP3S
:       :       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Nature  Nature  NPMS
,       ,       PCTFAIB
berce   bercer  VIMPP2S
-le     le      PPER3S
chaudement  chaudement  ADV
:       :       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
il      il      PPER3S
a froid avoir   VINDP3S
.       .       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Les     le      DETDPIG
parfums parfum  NCMP
ne      ne      ADV
font    faire   VINDP3P
pas     pas     ADV
frissonner      frissonner      VINF
sa      son     DETPOSS
narine  nar     NCFS
;       ;       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Il      il      PPER3S
dort    dormir  VINDP3S
dans    dans    PREP
le      le      DETDMS
soleil  soleil  NCMS
,       ,       PCTFAIB
la      le      DETDFS
main    main    NCFS
sur     sur     PREP
sa      son     DETPOSS
poitrine        poitrine        NCFS
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Tranquille      Tranquille      NPMS
.       .       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
Il      il      PPER3S
a       avoir   VINDP3S
deux    deux    ADJNUM
trous   trou    NCMP
rouges  rouge   ADJPIG
au      à le    DETDMS
côté    côté    NCMS
droit   droit   ADJMS
.       .       PCTFORTE
===== FIN DE PHRASE =====
==== DEBUT DE PHRASE ====
===== FIN DE PHRASE =====
```

D Installation de TreeTagger

1. **Téléchargement de TreeTagger**
   Ouvrez un navigateur web et saisir dans la barre d'adresse l'url :
   `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
   ou cherchez *"treetagger"* dans un moteur de recherche.
   Trouvez le lien *Windows version* (entre les sections *Parameter files* et *tagsets*, ou Ctrl+F et *Windows version* dans la barre de recherche) et effectuez un clic-droit, puis sélectionnez *enregistrer la cible du lien sous* dans le menu contextuel. Enregistrez le fichier `tree-tagger-windows-3.1.zip` dans le répertoire `C:\`.
2. **Extraction de l'archive TreeTagger**
   Effectuez un clic-droit sur le fichier `.zip` et sélectionnez dans le menu contextuel *7-zip/extract here*.
   Le répertoire `C:\TreeTagger` contient à présent un sous-répertoire `TreeTagger` qui contient lui-même les sous-répertoires `bin`, `cmd` et `lib`.
3. **Téléchargement des fichiers de paramètres**
   Pour utiliser TreeTagger avec une langue donnée, il faut lui fournir un fichier de paramètres correspondant. Dans la section *Parameter files for PC* de la page web de TreeTagger, effectuez un clic-droit sur les liens *English parameter file* et *French parameter file (Latin1)* et enregistrez les fichiers dans `C:\TreeTagger\lib`.
4. **Extraction des fichiers de paramètres**
   Le répertoire `C:\TreeTagger\lib` doit contenir les fichiers suffixés `.bin.gz`. Décompressez chacun de ces fichiers par un clic-droit, puis *7-zip/extract here*. Une fois décompressés, les fichiers portent l'extension `.par`. Renommez-les (clic-droit, puis *Renommer* ou sélection, puis F2) respectivement en `french.par` et `english.par` s'ils portent des noms différents.

E Utiliser séparément le segmenteur de TreeTagger

TreeTagger fournit un programme qu'il utilise pour *tokeniser* les textes avant de les étiqueter. La procédure pour segmenter un texte est décrite ci-dessous. On suppose que TreeTagger est installé dans le répertoire `C:\TreeTagger`, que le texte à segmenter est `C:\SL0720\texte.txt` et que l'on veut produire le résultat de la segmentation dans le fichier `C:\SL0720\texte.tok`.

1. Ouvrir l'invite de commande : sous Windows, bouton *Démarrer/Exécuter*, puis saisir `cmd` et validez avec la touche `Entrée`.
2. Saisissez la commande suivante :
   `perl c:\TreeTagger\cmd\tokenize.pl -f c:\720\texte.txt > c:\720\texte.tok`
3. Pour l'anglais, remplacer l'option `-f` par `-e`