SL0720X
Étiquetage morphosyntaxique/TreeTagger

Franck Sajous/CLLE-ERSS

http://fsajous.free.fr/

Préambule

RAPPEL

- Noms de fichiers : seulement des lettres (non accentuées), des chiffres et des tirets
- Pas d'espace, pas d'accent, pas de signe de ponctuation autre que tiret (-) et tiret-bas (_)
- Concerne TOUS vos fichiers/répertoires (données traitées et produites par TreeTagger, mais aussi archives, compte-rendus, pièces-jointes, etc.)

(dernier coup de semonce avant hostilités)

Format des compte-rendus

- Convertissez vos documents .doc, .docx, .odt, etc. en PDF (sous peine de perte de mise en page)
- Autant d'environnements que d'individus, mais des standards

Entrée/sortie

Format d'entrée :

Le gloubi-boulga est la nourriture préférée de Casimir.

Format de sortie « par défaut »:

```
Le              DET:ART     le
gloubi-boulga   NOM         <unknown>
est             VER:pres    être
la              DET:ART     le
nourriture      NOM         nourriture
préférée        VER:pper    préférer
de              PRP         de
Casimir         NAM         Casimir
.               SENT        .
```

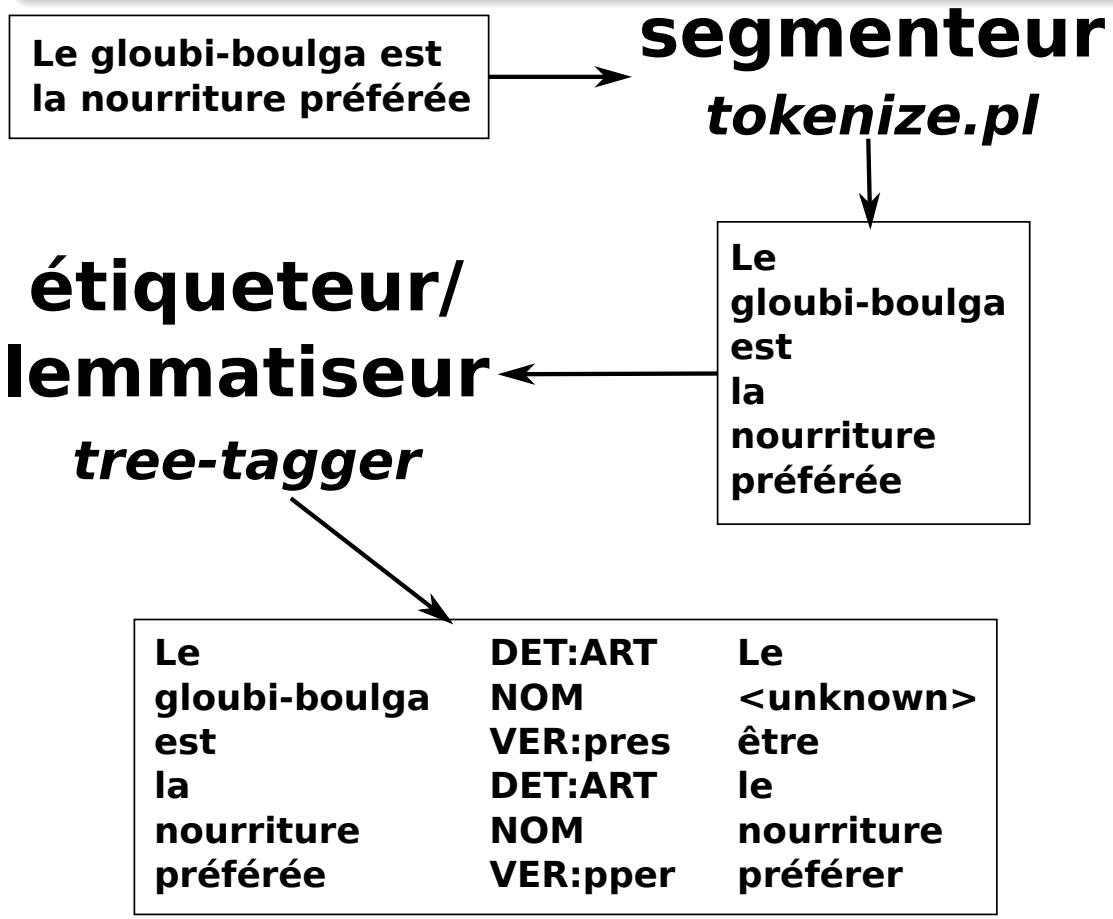
Étiquetage

- Information fournies : parties du discours, lemmes, frontières de phrases
- Étapes préalables : segmentation en phrases, puis en tokens (des traitements non triviaux !)

Pré-traitements : segmentation

Tokenisation ou « *segmentation en mots* »

- ne fait pas partie à proprement parler de l'étiquetage.
- script distribué avec TreeTagger/script intégré à l'interface



La tokenisation

Tokenisation ou *« segmentation en mots »*

- étape non triviale : règles générales + lexiques d'exceptions
- différente selon les langues, y compris d'une même famille. Par exemple, comment traiter les clitiques ?
  - *« Pouvez-vous venir lundi ? »*
  - *« Pour vous inscrire, rendez-vous au secrétariat et suivez la procédure. »*
  - *« Prenez un rendez-vous avec le secrétariat. »*
- traitement nécessaire des unités polylexicales (*pomme de terre, grille-pain*), des locutions (*c'est-à-dire*)
- cas non ambigus (*aujourd'hui*) ou ambigus (*collectionneur d'œuvres d'art/hors d'œuvre*)
- cas d'ambiguïté : en fait, bien que
- si l'on fournit sa propre tokenisation, elle doit être en cohérence avec le lexique de TreeTagger. Sinon, on doit étiqueter les unités lexicales que l'on segmente différemment.

Segmentation en phrases

Un token particulier : la frontière de phrases

- marques de poncuations fortes : . ! ? ...
- un bon indice, mais pas suffisant
  - sigles : « *La S.N.C.F. augmente ses tarifs* »
  - abréviation : « *M. Machin a annoncé que. . .* »
  - nombres décimaux (notation anglo-saxonne) et autres : « *Une valeur approximée à 3.14 sera utilisée* », « *recours au 49.3 pour passer en force* »
- $\rightarrow$ règles + lexiques, comme pour la tokenisation
- Exemple : M. + une majuscule + une suite de minuscules = abréviation de Monsieur + nom propre $\rightarrow$ ne pas segmenter. Oui, mais :
  - « *Il a commandé un tee-shirt en taille M. Il est trop petit.* »
  - « *L'acide folique s'appelait vitamine M. Depuis, c'est vitamine B9.* »
  - etc.

Tokenisation : retour aux observations

à la sauce Casimirus. . . avec de la moutarde

sauce	NOM	sauce
Casimirus..	ABR	<unknown>
.	SENT	.
avec	PRP	avec

Tokenisation : retour aux observations

à la sauce Casimirus. . . avec de la moutarde

sauce	NOM	sauce
Casimirus..	ABR	<unknown>
.	SENT	.
avec	PRP	avec

mauvaise segmentation dûe uniquement à l'utilisation de l'interface.

Si l'on effectue la tokenisation et l'étiquetage en ligne de commande :

```
perl c:\Cours\Outils\TreeTagger\cmd\tokenize.pl -f gloubi.txt > gloubi.tok
```

```
Casimirus
...
avec
```

Tokenisation : retour aux observations

à la sauce Casimirus. . . avec de la moutarde

sauce	NOM	sauce
Casimirus..	ABR	<unknown>
.	SENT	.
avec	PRP	avec

mauvaise segmentation dûe uniquement à l'utilisation de l'interface.

Si l'on effectue la tokenisation et l'étiquetage en ligne de commande :

```
perl c:\Cours\Outils\TreeTagger\cmd\tokenize.pl -f gloubi.txt > gloubi.tok
```

```
Casimirus
...
avec
```

```
c:\Cours\Outils\TreeTagger\bin\tree-tagger
    c:\Cours\Outils\TreeTagger\lib\french.par -token -lemma > gloubi.tag
```

Casimirus	NAM	<unknown>
...	PUN	...
avec	PRP	avec

Ponctuation et lemmes

Colonne des lemmes

- le lemme de la forme fléchie si elle fait partie du lexique de TreeTagger
- `@card@` pour les nombres
- `<unknown>` si le token ne fait pas partie du lexique

Ponctuation

`SENT`	ponctuations fortes (fins de phrase)
`PUN:cit`	marques de citation (guillemets)
`PUN`	autres

Jeu d'étiquettes

Quelques commentaires

- on ne sait pas trop d'où il vient (quel corpus d'apprentissage ?)
- pas d'information sur la flexion des noms/adjs (m/f, p/s)
- mais temps des verbes

Observations sur l'étiquetage VER:pper/ADJ

la nourriture préférée	DET:ART NOM VER:pper	le nourriture préférer
la banane écrasée	DET:ART NOM ADJ	le banane écrasé
du chocolat rapé	PRP NOM VER:pper	le chocolat <unknown>

Observations sur l'étiquetage NOM/NAM

Christophe Izard, le créateur de Casimir

Christophe	NAM	Christophe
Izard	NAM	<unknown>
Casimir	NAM	Casimir

Observations sur l'étiquetage NOM/NAM

Christophe Izard, le créateur de Casimir

Christophe	NAM	Christophe
Izard	NAM	<unknown>
Casimir	NAM	Casimir

gâteau [...] que seul le Casimirus adore

le	DET:ART	le
Casimirus	NOM	<unknown>

Observations sur l'étiquetage NOM/NAM

Christophe Izard, le créateur de Casimir

Christophe	NAM	Christophe
Izard	NAM	<unknown>
Casimir	NAM	Casimir

gâteau [...] que seul le Casimirus adore

le	DET:ART	le
Casimirus	NOM	<unknown>

Gloubi-boulga Nights ou Soirées Gloubi-boulga

Soirées	NAM	<unknown>

Observations sur l'étiquetage NOM/NAM

Christophe Izard, le créateur de Casimir

Christophe	NAM	Christophe
Izard	NAM	<unknown>
Casimir	NAM	Casimir

gâteau [...] que seul le Casimirus adore

le	DET:ART	le
Casimirus	NOM	<unknown>

Gloubi-boulga Nights ou Soirées Gloubi-boulga

Soirées	NAM	<unknown>

saucisse de Toulouse

Toulouse	NAM	Toulouse

Observations sur l'étiquetage NOM/NAM

Christophe Izard, le créateur de Casimir

Christophe	NAM	Christophe
Izard	NAM	<unknown>
Casimir	NAM	Casimir

gâteau [...] que seul le Casimirus adore

le	DET:ART	le
Casimirus	NOM	<unknown>

Gloubi-boulga Nights ou Soirées Gloubi-boulga

Soirées	NAM	<unknown>

saucisse de Toulouse

Toulouse	NAM	Toulouse

À la Libération, en découvrant...

Libération	NOM	libération

Étiquetage de mots « ambigus »

Ici, ambigu = mot connu du lexique de TT comme pouvant avoir au moins deux étiquettes possibles.

des jaunes d'oeufs

des	PRP:det	du
jaunes	NOM	jaune
d'	PRP	de
oeufs	NOM	oeuf

Étiquetage de mots « ambigus »

Ici, ambigu = mot connu du lexique de TT comme pouvant avoir au moins deux étiquettes possibles.

des jaunes d'oeufs

des	PRP:det	du
jaunes	NOM	jaune
d'	PRP	de
oeufs	NOM	oeuf

crème chantilly

crème	VER:pres	crémer

Étiquetage en anglais

Langues différentes, jeux d'étiquettes différent

- peut refléter des différences inter-langues
- aussi souvent (surtout) dû au corpus d'apprentissage utilisé

Exemples

- `PRP:det` (Preposition plus article, e.g. *au, du, des*) spécifique au français
- Pas de `WP` (Wh-pronoun) en français. Mais pas non-plus de catégorie *pronom interrogatif*
- Pas de `RBR/RBS` (Adverb comparative/superlative), ni de `PDT` (Predeterminer) dans le tagset français. Pourtant, ça existe.
- `UH` (Interjection) et `FW` (Foreign Word) spécifiques à l'anglais
- Pas d'info flexionnelle pour les noms en français. Singulier/pluriel différencié pour l'anglais (`NN/NNS`)

Jeux d'étiquettes : des standards, et d'autres

TreeTagger

- d'une langue à l'autre, étiquettes différentes pour même catégorie (e.g. `PRO:PER=PP`)
- tagsets issus des corpus d'apprentissage : anglais→PTB, français→???

GRACE (Grammaires et Ressources pour les Analyseurs de Corpus et leur Évaluation)

- campagne d'évaluation (comparaison) d'analyseurs syntaxiques → nécessité de mettre au point un tagset commun
- « *Format de description lexicale pour le français, partie 2 : description morpho-syntaxique* » (Rajman et al., 1997) document disponible sur IRIS
- exemples d'étiquettes : `N[cp][mf][sp]` (noms), `Af[pc][mf][sp]` (adj. qual.), `V[ma]i[pifs][123][sp]-` (verbes à l'indicatif)

Un autre étiqueteur : Cordial

Format de sortie de Cordial :

1. numéro de mot dans la phrase
2. forme
3. lemme
4. pos (catégorie grammaticale)
5. pos2 (catégorie + flexion, format GRACE modifié)
6. numéro de la tête du syntagme minimal [| du syntagme maximal]
7. fonction syntaxique
8. numéro de la proposition
9. pivot (verbe de la proposition)
10. type de proposition (eg. S→contient le sujet, V→verbe de base de la proposition, B→appartient à l'attribut du sujet, T→contient le sujet)
11. indications sémantiques (recours à des ontologies)

TreeTagger/Cordial

Comparaison sur *le dormeur du val*

(voir ref [Habert 2005] dans support de la séance 1)

- Différence de segmentation (l'exemple s'y prête !)
  6 phrases avec TT, 20 avec Cordial

TreeTagger/Cordial

Comparaison sur *le dormeur du val*

(voir ref [Habert 2005] dans support de la séance 1)

- Différence de segmentation (l'exemple s'y prête !)
  6 phrases avec TT, 20 avec Cordial
- *bouche ouverte, tête nue*
  - TT : nue → NOM/nue
  - Cordial : tête nue → ADV/tête nue

TreeTagger/Cordial

Comparaison sur *le dormeur du val*

(voir ref [Habert 2005] dans support de la séance 1)

- Différence de segmentation (l'exemple s'y prête !)
  6 phrases avec TT, 20 avec Cordial
- *bouche ouverte, tête nue*
  - TT : nue → NOM/nue
  - Cordial : tête nue → ADV/tête nue
- *étendu dans l'herbe sous la nue*
  - TT : nue → NOM/nue (OK)
  - Cordial : nue → NCFS/nue (OK)

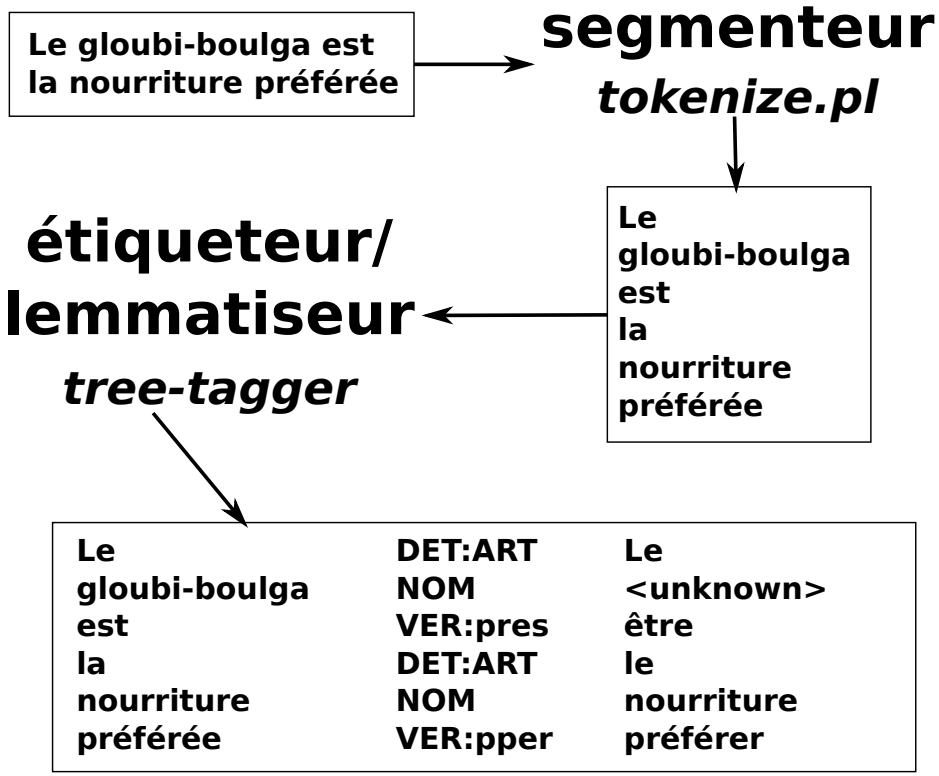
TreeTagger/Cordial

Comparaison sur *le dormeur du val*

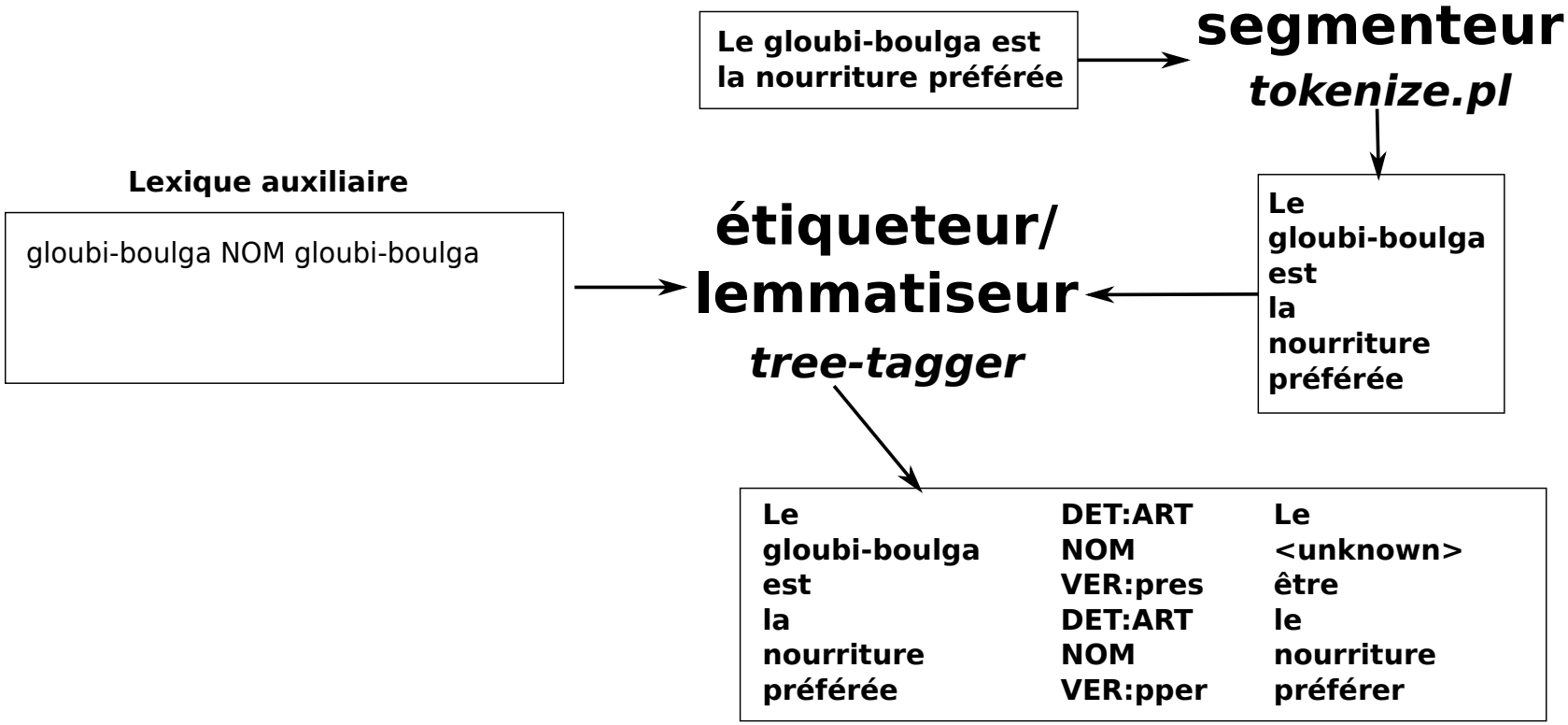
(voir ref [Habert 2005] dans support de la séance 1)

- Différence de segmentation (l'exemple s'y prête !)
  6 phrases avec TT, 20 avec Cordial
- *bouche ouverte, tête nue*
  - TT : nue → NOM/nue
  - Cordial : tête nue → ADV/tête nue
- *étendu dans l'herbe sous la nue*
  - TT : nue → NOM/nue (OK)
  - Cordial : nue → NCFS/nue (OK)
- *sourirait un enfant malade, il faut un somme*
  - Cordial : fait un somme → sourire/VINDP3S
    ce n'est pas une erreur d'analyse, c'est un bug !

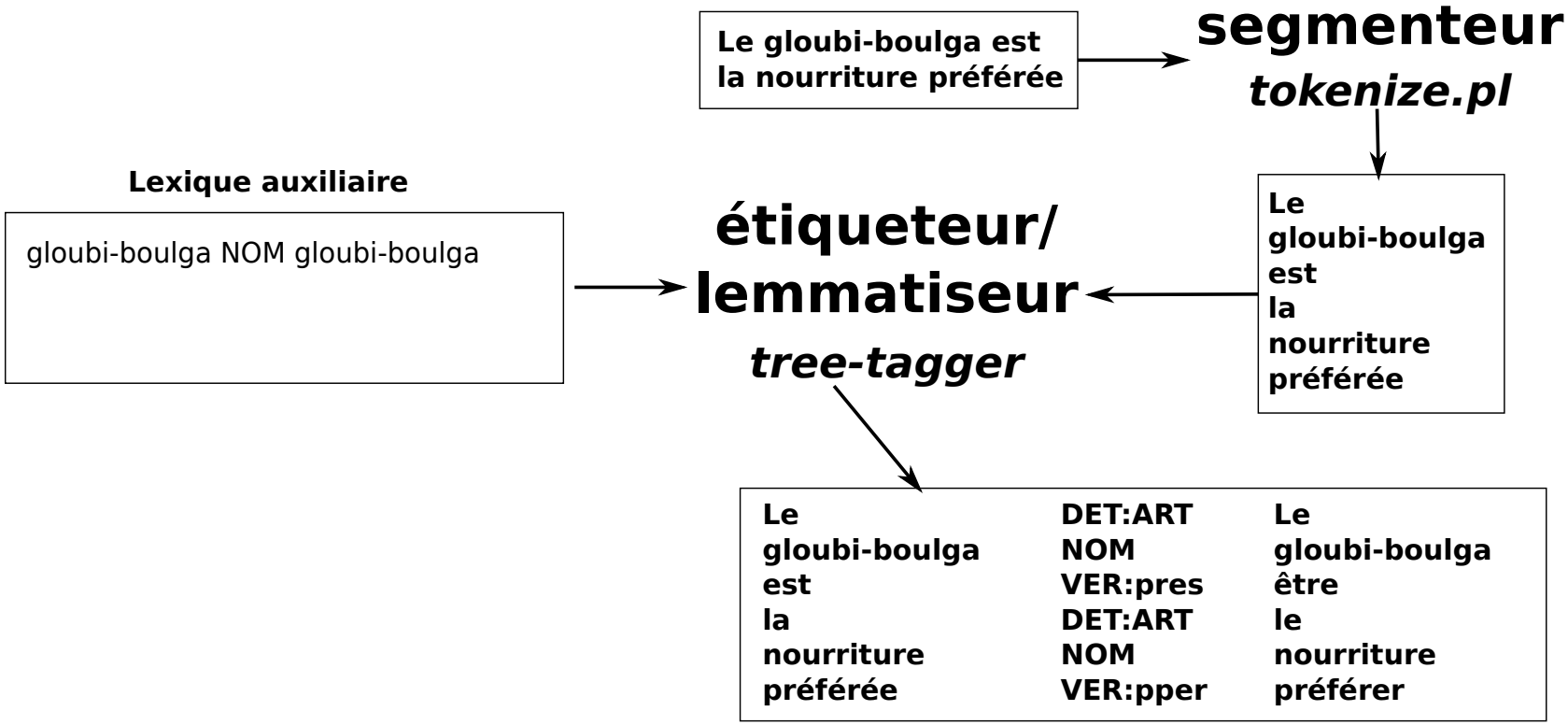
Lexique auxiliaire



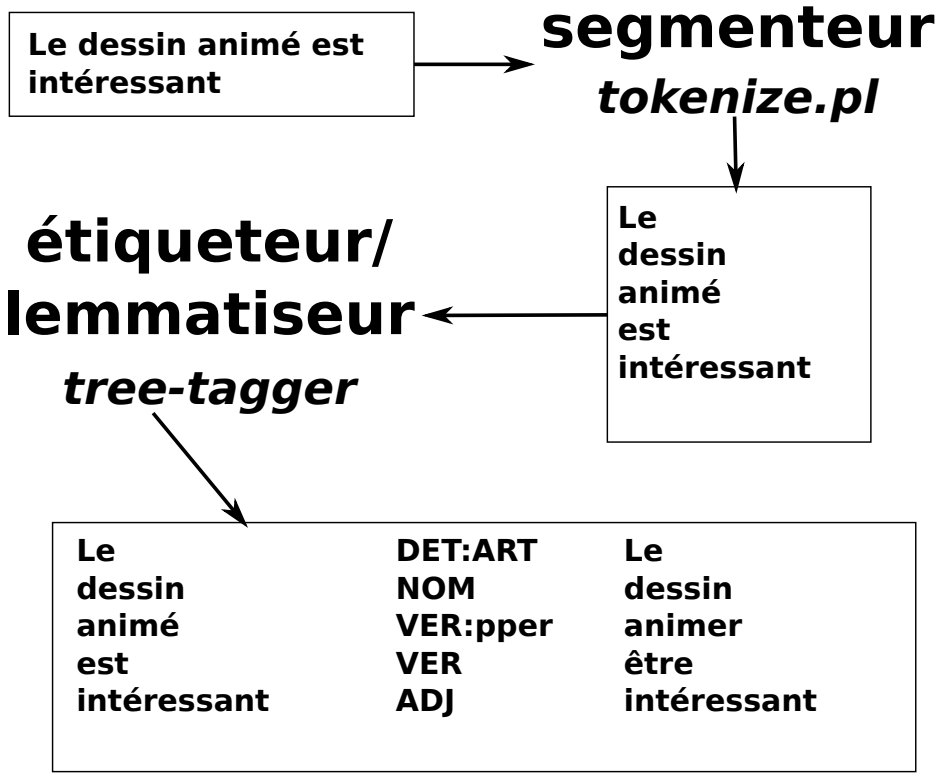
Lexique auxiliaire



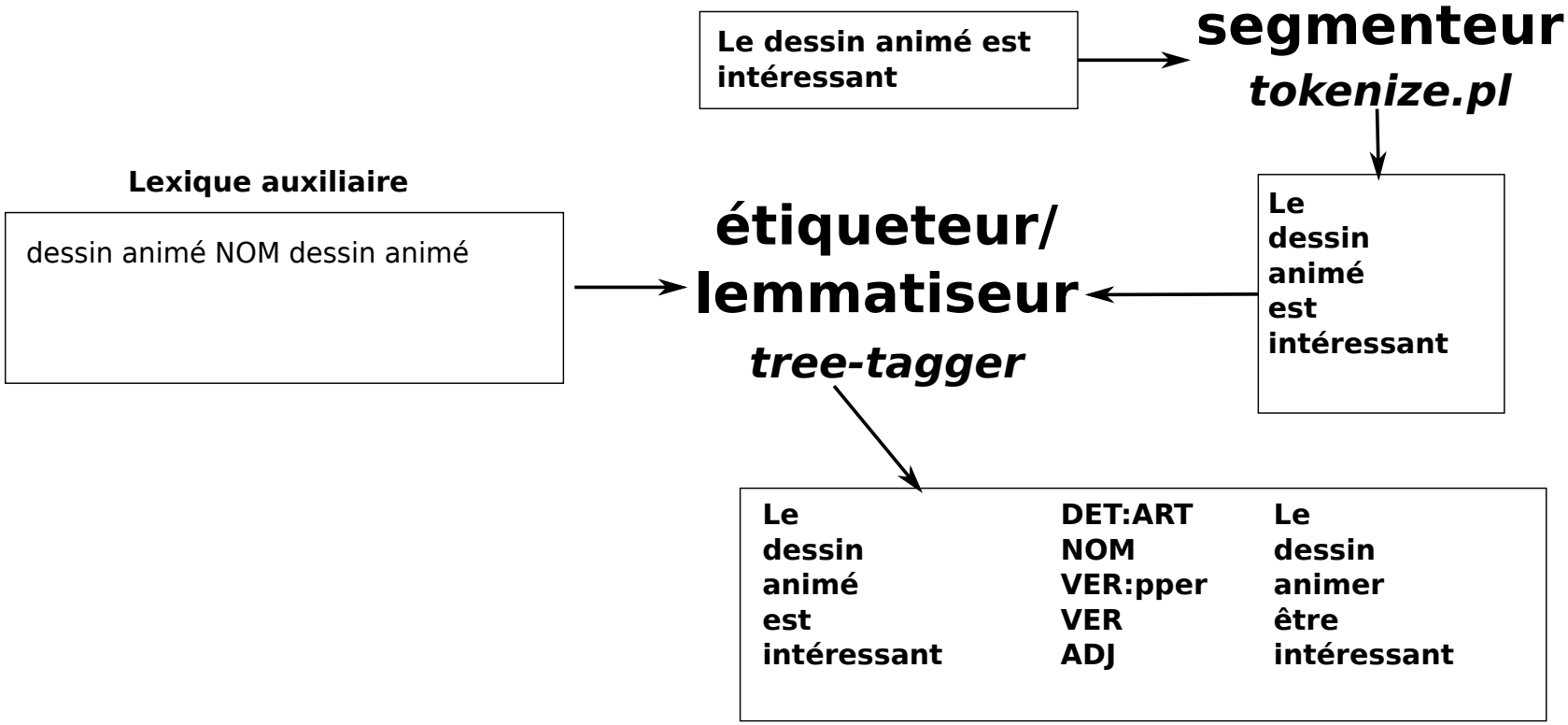
Lexique auxiliaire



Lexique auxiliaire



Lexique auxiliaire



Lexique auxiliaire vs préétiquetage

Lexique auxiliaire

- informations mise à disposition de TreeTagger (si plusieurs catégories pour une forme, c'est lui qui choisit !)
- mots absents de son lexique

Préétiquetage

- on effectue la segmentation
- on étiquette et lemmatise cetaines entrées seulement !
- e.g. étiqueter spécifiquement *métier* (figurant comme nom dans le lexique de TT) comme adjectif lorsqu'il suit *application*
- quelle serait la conséquence de rajouter métier NOM et ADJ dans un lexique auxiliaire ?

On peut utiliser conjointement les deux !

Problèmes de manipulations et options d'étiquetage

TreeTagger perdu par les majuscules/le manque de pontuation forte

- Poésie et versification : c'est prévisible
- Fichier "gloubiboulga"
  - copier-coller depuis l'article de Wikipédia dans un fichier texte.
  - problème notamment des titres non terminés par des ponctuations fortes (souvent le cas pour les articles de journaux également)
  - mais il y a des lignes vides comme séparateurs !
  - $\rightarrow$ le problème se situe au niveau de la segmentation
- Dans *Tagging Options*, cocher *Use capital heuristics* résout nombre de problèmes

Problèmes d'encodage

TreeTagger « pur » et interface

- nouvelle version de TreeTagger : UTF-8 en entrée, UTF-8 en sortie (UTF-8 pour une éventuel lexique auxiliaire)
- interface WinTreeTagger : ???
  ça dépend ! De votre environnement, notamment.
- selon l'environnement, l'utilisation de l'interface conduit à un étiquetage improbable (pas seulement pour les mots comportant des diacritiques !)
- $\rightarrow$ désormais, utilisation de TreeTagger en ligne de commande