Achille Falaise

Introduction à l'annotation automatique de corpus textuels

Lausanne, 30 mai 2016

http://pro.aiakide.net/cours/L2016

Plan

1) Une annotation, pour quoi faire ?
2) Fonctionnement d'un tagger
3) Exemple pratique

Quel type d'annotation ?

Chaque être humain naît libre.

Forme	*POS*	*Lemme*
Chaque	PRO:IND	chaque
être	NOM	être
humain	ADJ	humain
naît	VER:pres	naître
libre	ADJ	libre
.	SENT	.

Pour quoi faire ?

Utilisation

- Dans un tableur
- Avec des outils graphiques d'exploitation de corpus
  - TXM, Hyperbase...
- Des scripts Python, Perl, R...
- En ligne de commande

Pour quoi faire ?

Lister les occurrences d'une POS ou d'un lemme

- Tous les adjectifs

Forme	*POS*	*Lemme*
dernier	ADJ	dernier
puissante	ADJ	puissant
vaste	ADJ	vaste
terrible	ADJ	terrible
immenses	ADJ	immense
large	ADJ	large
parallèle	ADJ	parallèle
nord	ADJ	nord
sud	ADJ	sud
renversées	ADJ	renversé

L'Île mystérieuse (Jules Verne)
Wikisource → Version imprimable

Pour quoi faire ?

Lister les occurrences d'une POS ou d'un lemme

- Tous les adjectifs

Forme	*POS*	*Lemme*
dernier	ADJ	dernier
puissante	ADJ	puissant
vaste	ADJ	vaste
terrible	ADJ	terrible
immenses	ADJ	immense
large	ADJ	large
parallèle	ADJ	parallèle
nord	ADJ	nord
sud	ADJ	sud
renversées	ADJ	renversé

L'Île mystérieuse (Jules Verne)
Wikisource → Version imprimable

Pour quoi faire ?

Lister les occurrences d'une POS ou d'un lemme

- Fréquence des adjectifs

	Lemme
332	autre
272	grand
185	bon
184	nouveau
182	même
170	petit
144	seul
128	dernier
106	large
102	jeune
99	haut
96	beau

Pour quoi faire ?

Lister les occurrences d'une POS ou d'un lemme

Verbe support : faire(VER) + DET + NOM

Lemme	*Lemme*	*Lemme*
faire	le	description
faire	le	effet
faire	le	ciel
faire	un	signe
faire	notre	affaire
faire	le	rivière
faire	notre	provision
faire	le	récit
faire	le	fumeur
faire	le	joie
faire	un	litière

Pour quoi faire ?

Lister les occurrences d'une POS ou d'un lemme

- Verbe support : faire(VER) + DET + NOM

	Lemme	*Lemme*	*Lemme*
7	faire	le	tour
5	faire	un	signe
4	faire	le	ingénieur
3	faire	un	visite
3	faire	un	mouvement
3	faire	le	récit
3	faire	le	coup
3	faire	le	ciel
2	faire	un	voyage
2	faire	un	sorte

De nombreux outils

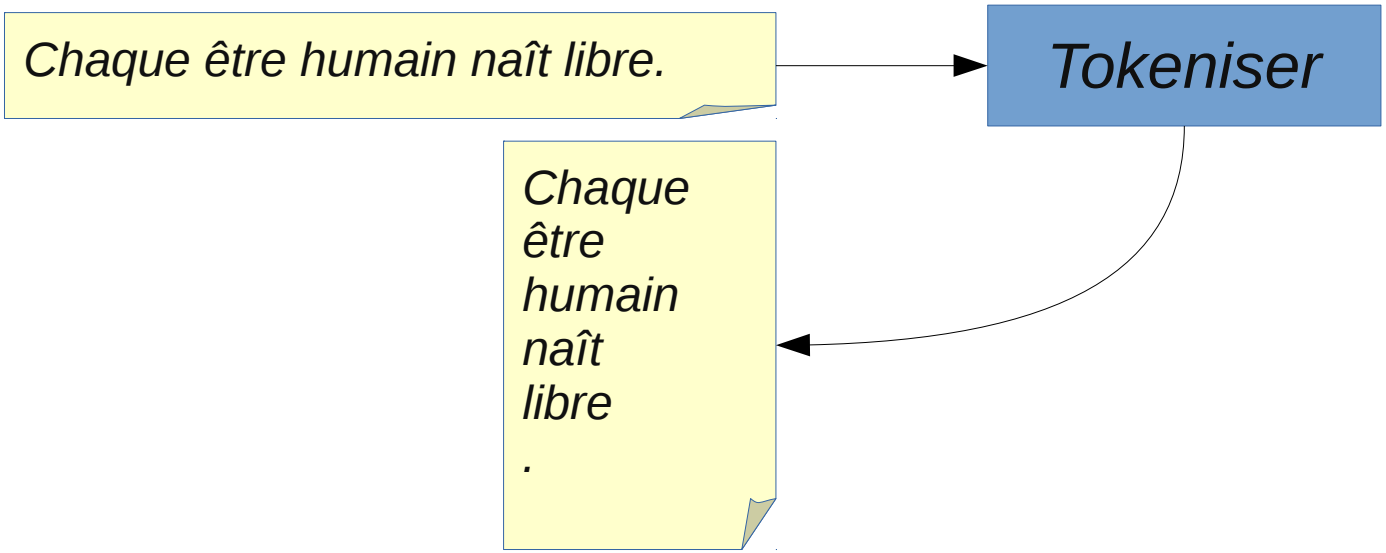
Parties du discours + lemmes (*tagger*)

- TreeTagger
  - http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
- MElt Tagger
- ...

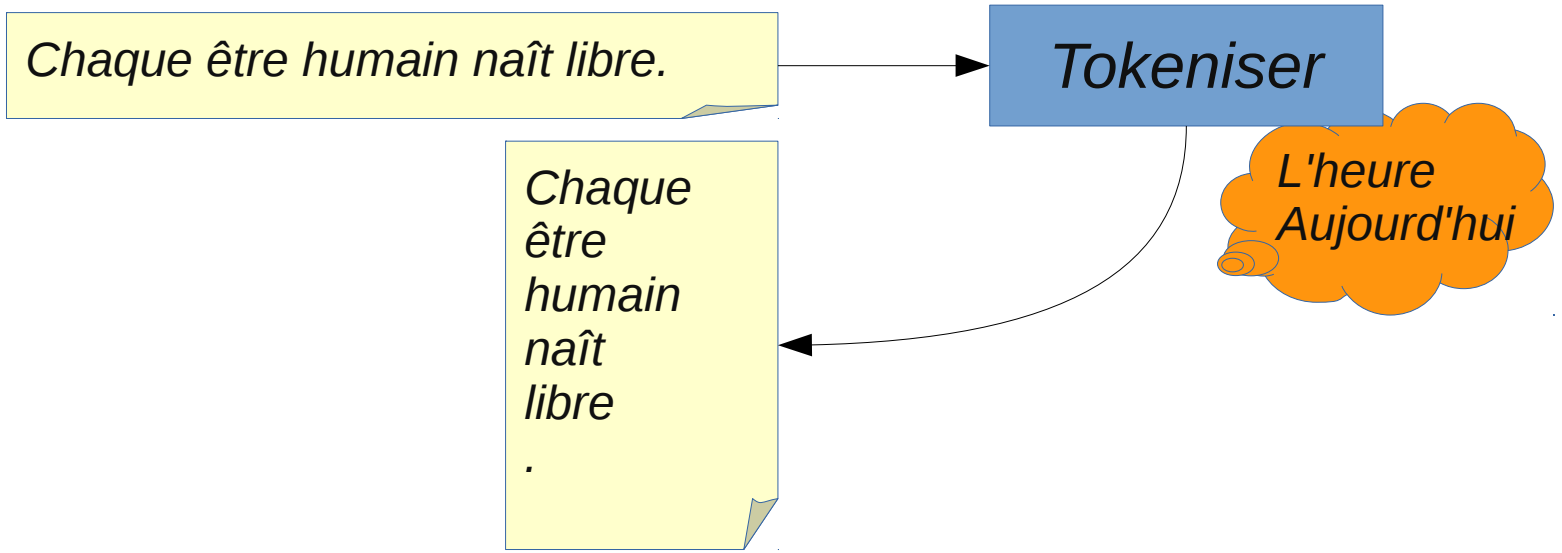
Dépendances (*parser*)

- Talismane
- Plateforme Bonsai
  - http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html
- ...

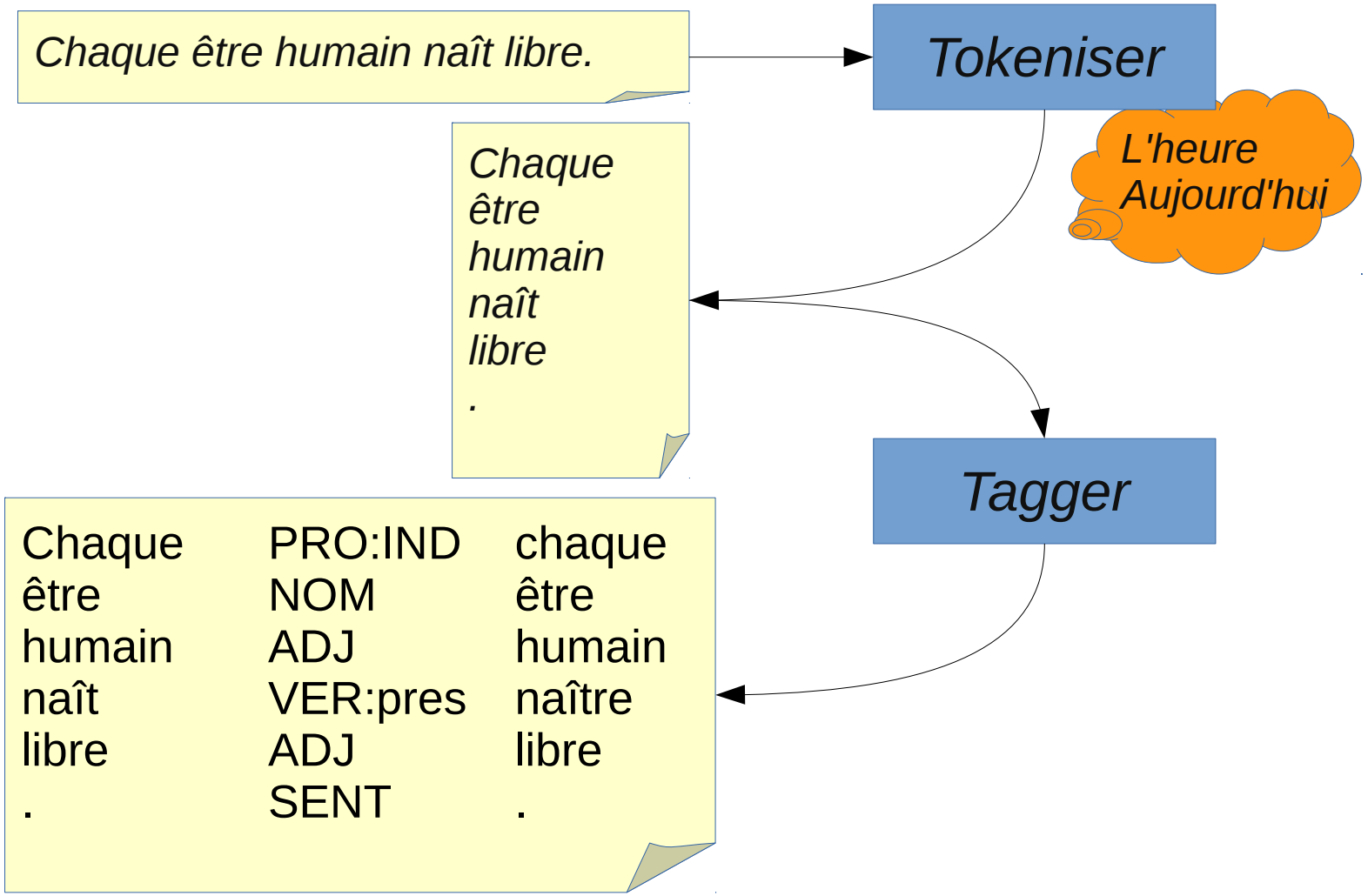
Anatomie d'un *tagger* stochastique



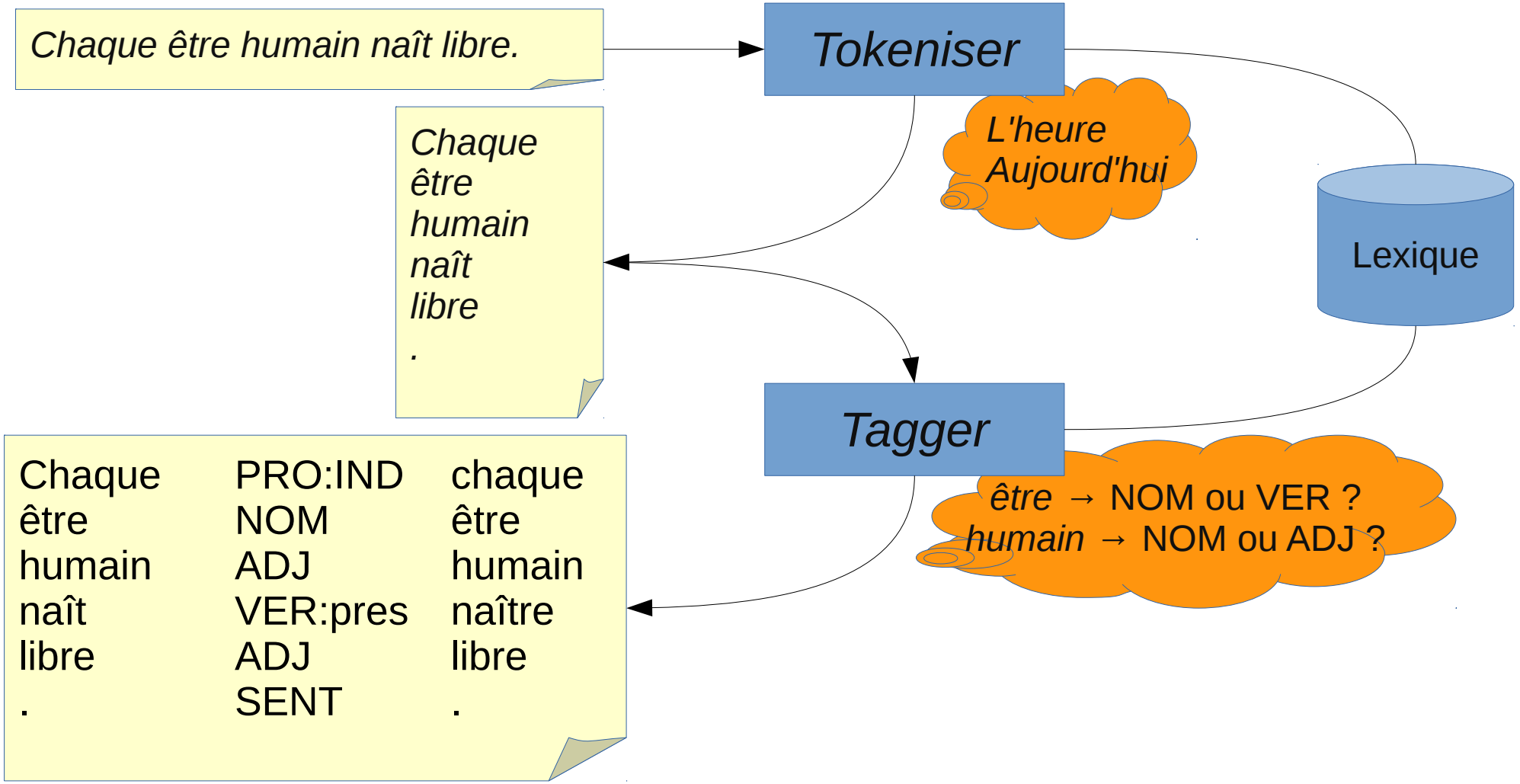
Anatomie d'un *tagger* stochastique



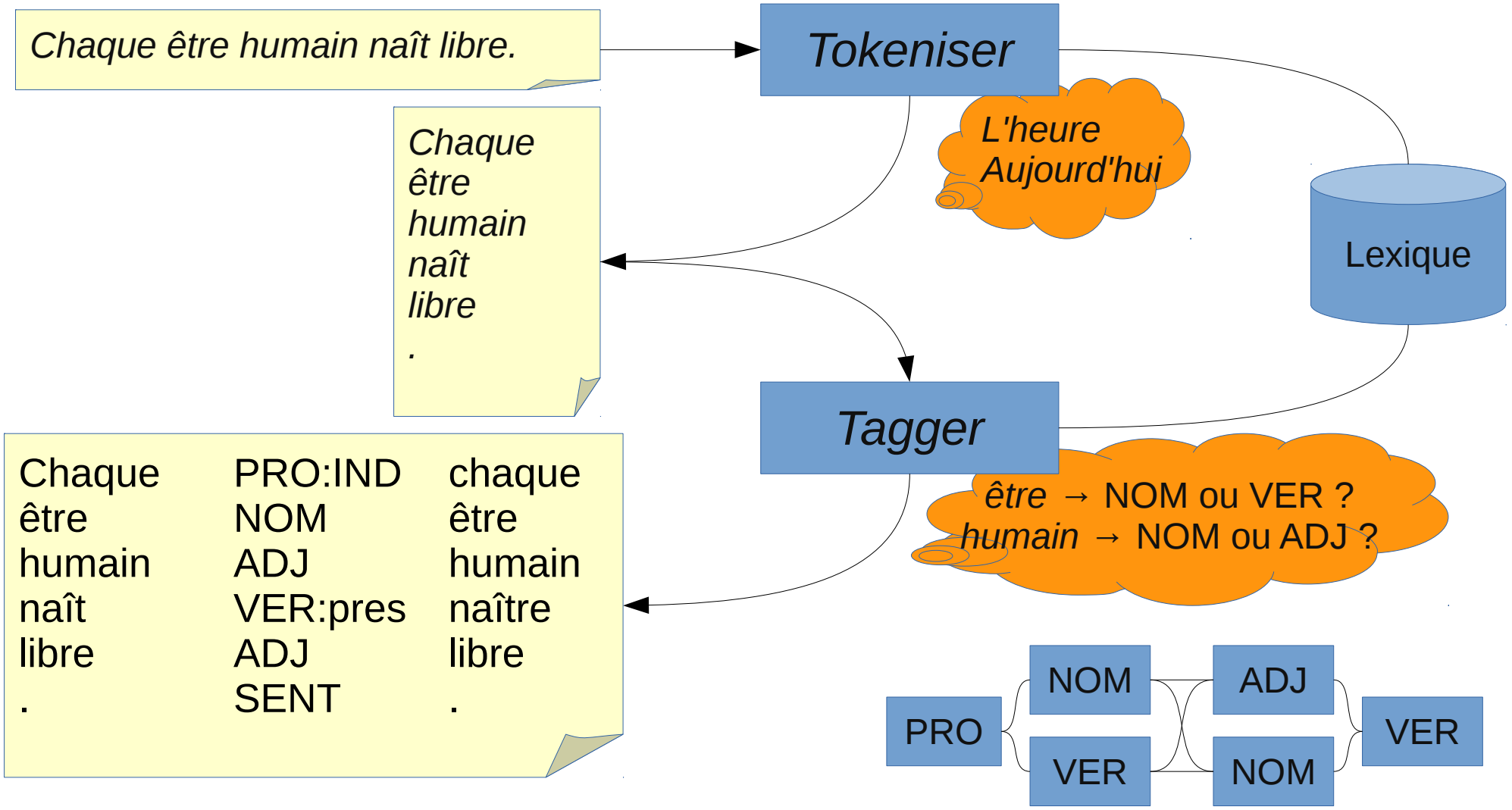
Anatomie d'un *tagger* stochastique



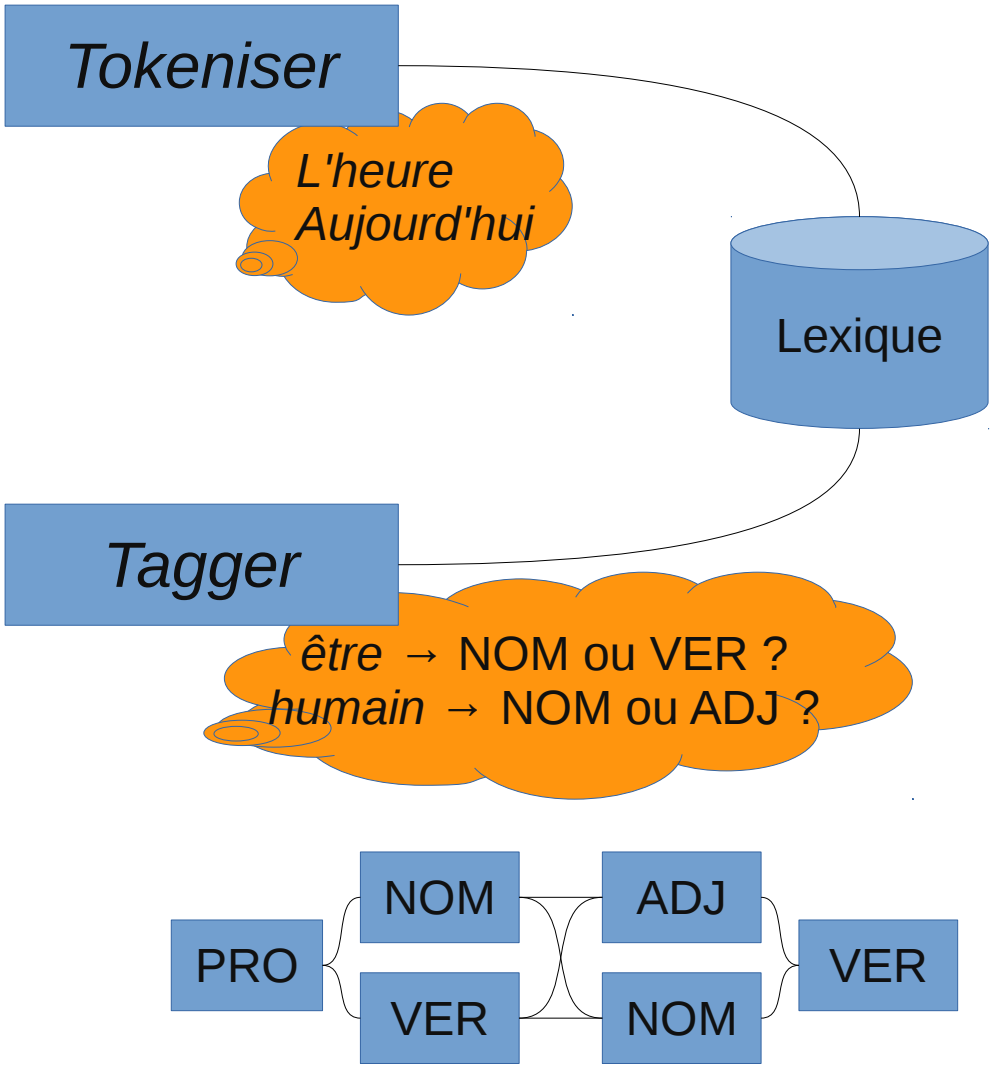
Anatomie d'un *tagger* stochastique



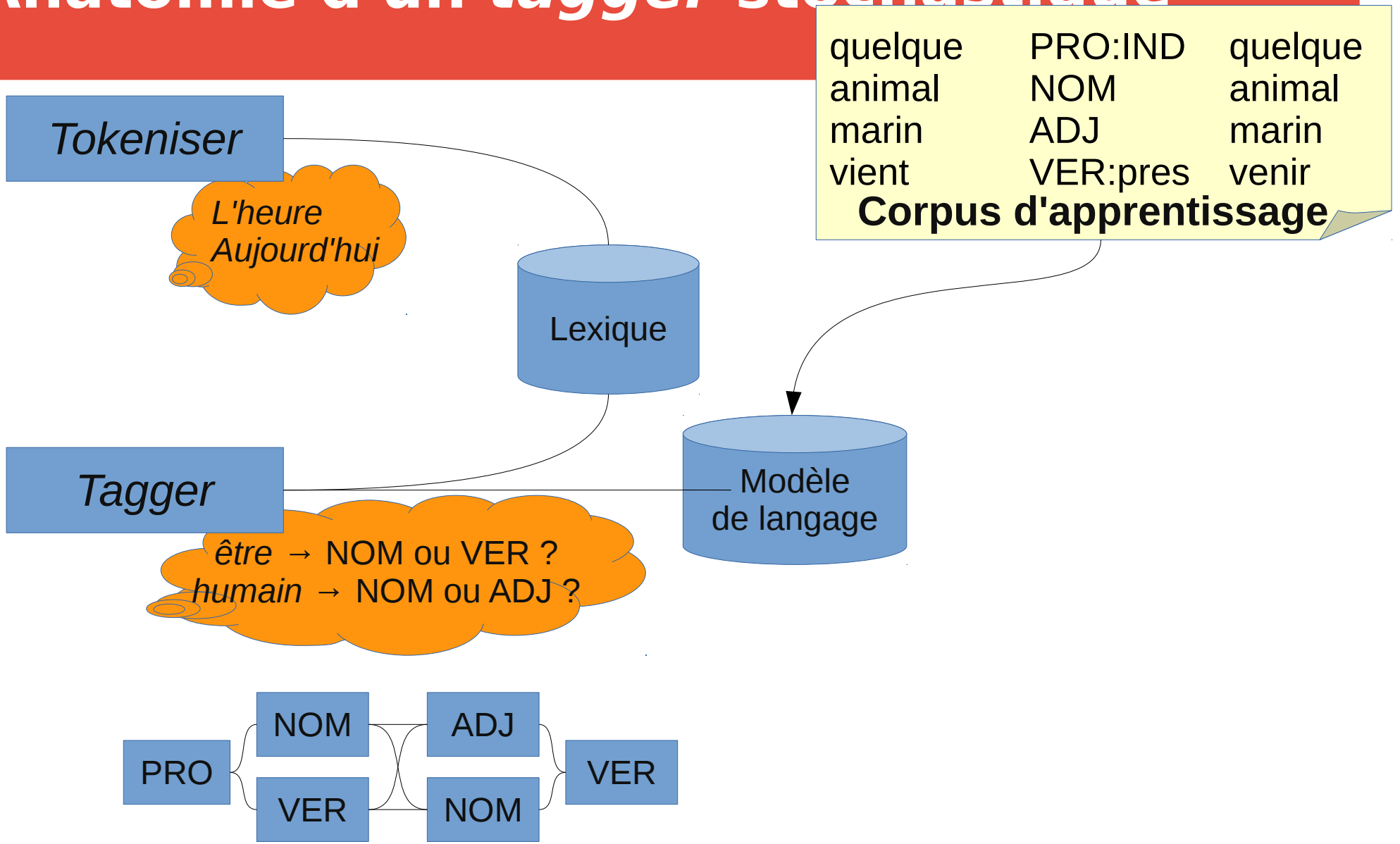
Anatomie d'un *tagger* stochastique



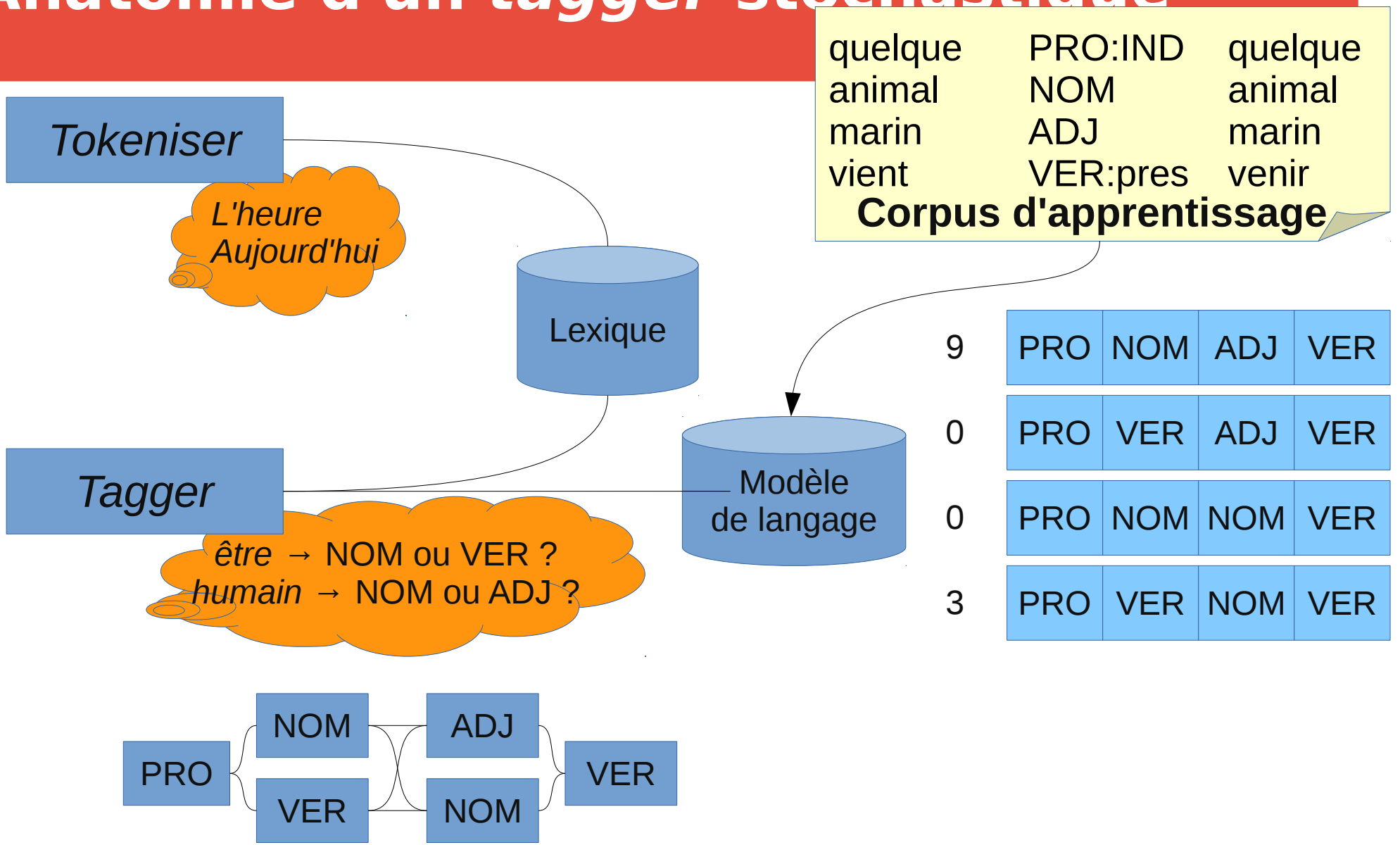
Anatomie d'un *tagger* stochastique



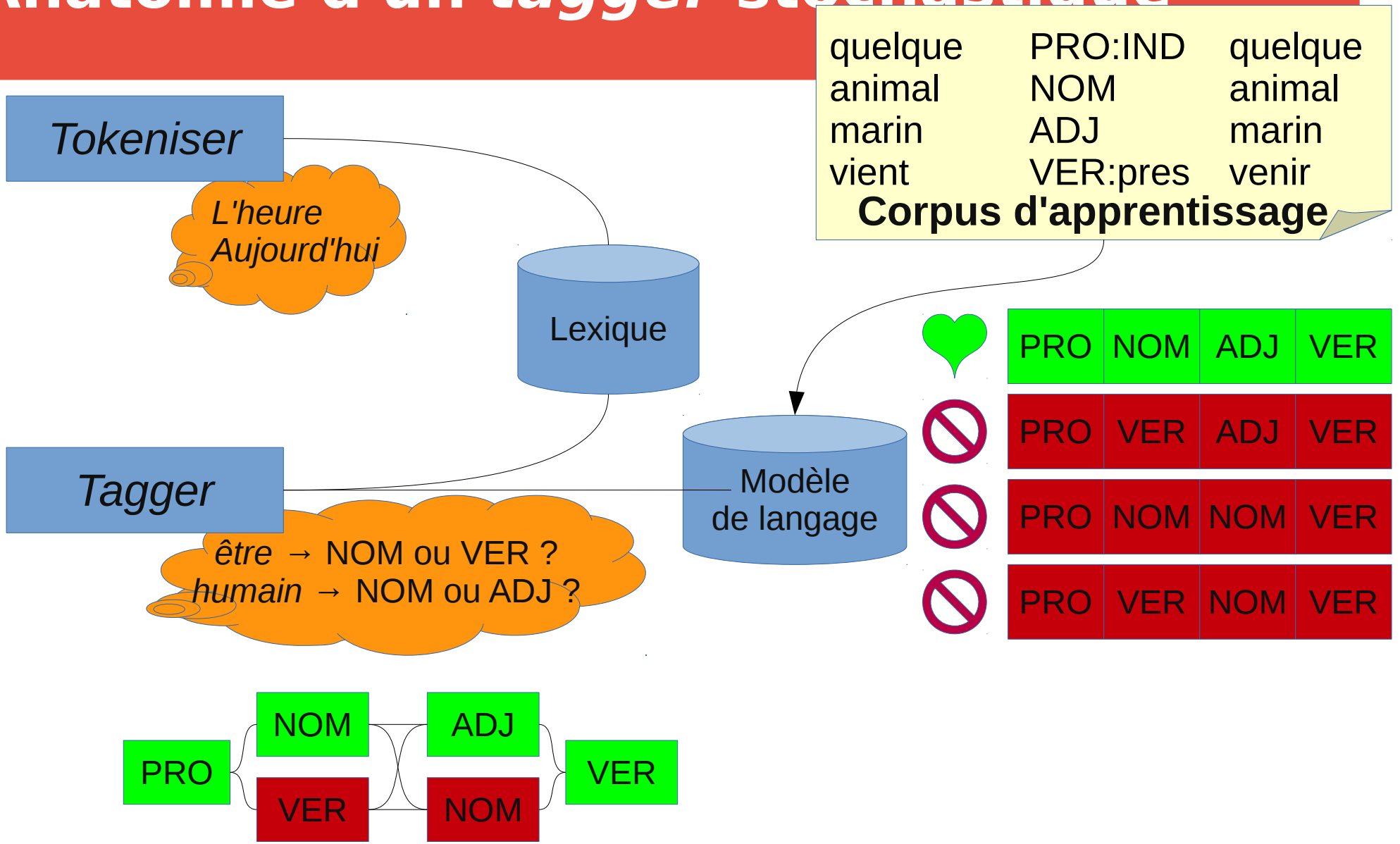
Anatomie d'un *tagger* stochastique



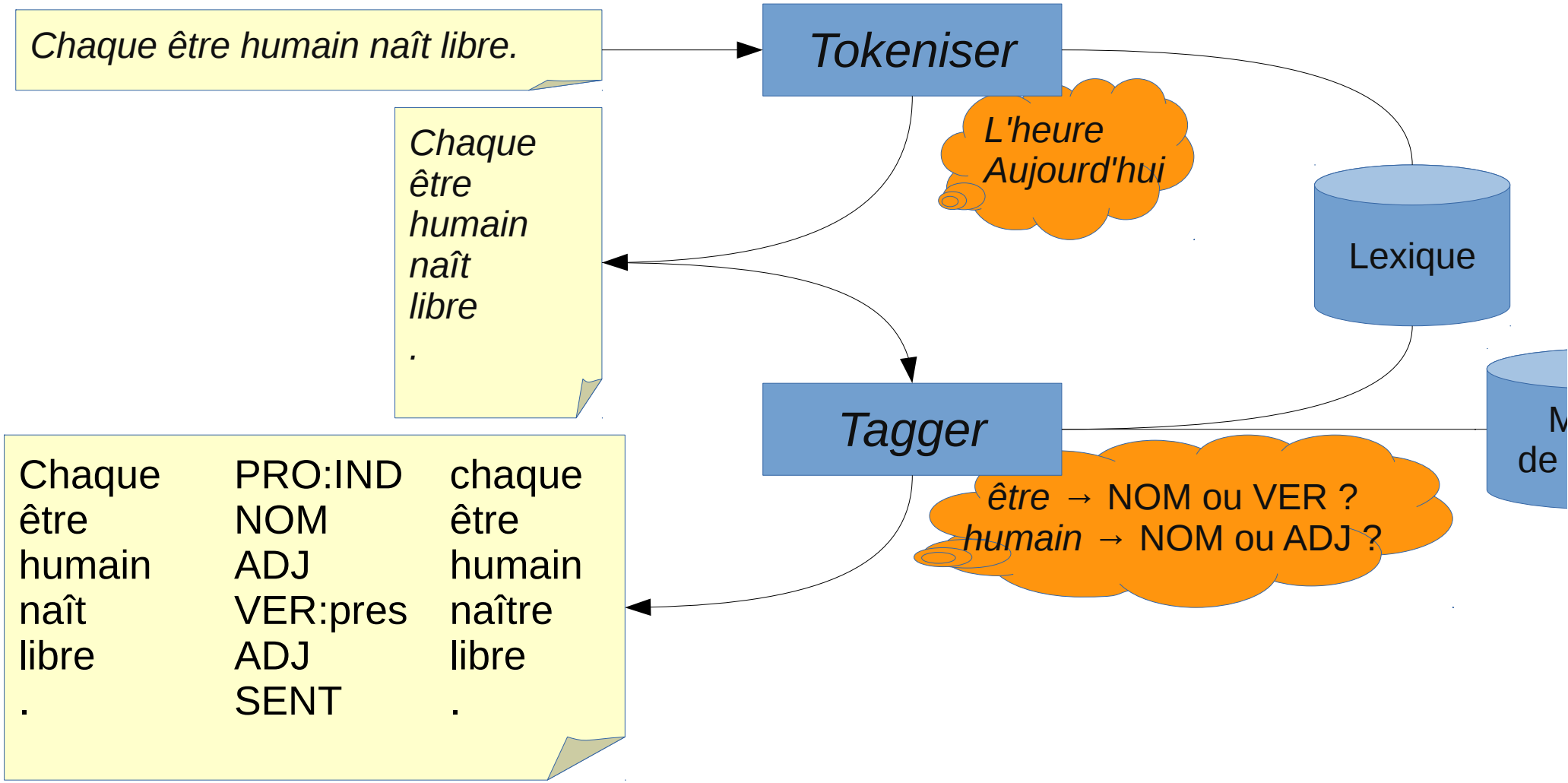
Anatomie d'un *tagger* stochastique



Anatomie d'un *tagger* stochastique



Anatomie d'un *tagger* stochastique



Anatomie d'un *tagger* stochastique

Conséquences

- Le *tagger* va faire des erreurs
- Beaucoup de choses dépendent des données d'apprentissage
  - La tokénisation
  - Le jeu d'étiquettes
  - Les lemmes
  - Les performances du tagger
- Plusieurs modèles pour le français
  - Français standard : http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
  - Français parlé : http://cnrtl.fr/corpus/perceo/
  - Français XVIème-XVIIIème : http://presto.ens-lyon.fr/?page_id=197
  - Français IXème-XVème : http://srcmf.org/

Mise en pratique

Installer TreeTagger

Aller sur :http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

1. Télécharger et extraire les *Tagging scripts*
2. Télécharger et extraire le *Tagger package*
3. Inutile de ~~Télécharger le script install-tagger.sh~~
4. Remplacer le script ./cmd/utf8-tokenize.perl (qui est bogué) par http://pro.aiakide.net/cours/L2016/utf8-tokenize.perl

Télécharger et extraire le modèle de langage *French parameter file (UTF-8)*

TreeTagger est prêt !

Test :

```
echo 'Chaque être humain naît libre.'  | ./cmd/utf8-tokenize.perl -f -a
lib/french-abbreviations-utf8 | bin/tree-tagger -token -lemma french.par
```

Mise en pratique

Corpus d'exemple

cat exemple.txt

Chaque être humain naît libre.

Mise en pratique

Tokeniser le corpus d'exemple

cat exemple.txt | cmd/utf8-tokenize.perl -f -a lib/french-abbreviations-utf8

Chaque

être

humain

naît

libre

.

Mise en pratique

Analyser le corpus d'exemple

cat exemple.txt | cmd/utf8-tokenize.perl -f -a lib/french-abbreviations-utf8 | bin/tree-tagger -token -lemma -quiet -sgml french.par

Chaque	PRO:IND	chaque
être	NOM	être
humain	ADJ	humain
naît	VER:pres	naître
libre	ADJ	libre
.	SENT	.

Mise en pratique

Analyser le corpus d'exemple

cat exemple.txt | cmd/utf8-tokenize.perl -f -a lib/french-abbreviations-utf8 | bin/tree-tagger -token -lemma -quiet -threshold .01 -prob french.par

Chaque	PRO:IND	chaque	1.00			
être	NOM	être	0.67	VER:infi	être	0.32
humain	ADJ	humain	0.97	NOM	humain	0.02
naît	VER:pres	naître	1.00			
libre	ADJ	libre	1.00			
.	SENT	.	1.00			

Adjectifs d'un corpus

Liste des adjectifs (diapo 6)

```
cat exemple.txt | ./cmd/utf8-tokenize.perl -f -a lib/french-abbreviations-utf8 | bin/tree-tagger -token -lemma -quiet french.par | grep -P '\tADJ.*' | cut -f 3
```

Fréquence des adjectifs (diapo 7)

```
cat exemple.txt | ./cmd/utf8-tokenize.perl -f -a lib/french-abbreviations-utf8 | bin/tree-tagger -token -lemma -quiet french.par | grep -P '\tADJ.*' | cut -f 3 | sort | uniq -c | sort -nr
```

Séquences *faire(VER) + DET + NOM*

Liste des séquences *faire(VER) + DET + NOM* (diapo 8)

```
cat exemple.txt | ./cmd/utf8-tokenize.perl -f -a lib/french-abbreviations-utf8 | bin/tree-tagger -token -lemma -sgml -quiet french.par | grep -Pzo '.*\tVER.*\tfaire\n.*\tDET.*\n.*\tNOM.*' | paste -d "\t" - - - | cut -f3,6,9
```

Fréquence des séquences (diapo 9)

```
cat exemple.txt | ./cmd/utf8-tokenize.perl -f -a lib/french-abbreviations-utf8 | bin/tree-tagger -token -lemma -sgml -quiet french.par | grep -Pzo '.*\tVER.*\tfaire\n.*\tDET.*\n.*\tNOM.*' | paste -d "\t" - - - | cut -f3,6,9 | sort | uniq -c | sort -nr
```

Et pour aller plus loin…

Les commandes Unix utiles en Lettres

Unix for Poets, Kenneth Ward Church, AT&T Bell Laboratories

Sur Linux, MacOS X, Windows 10

(→ Windows 10, manip à faire)

Création du fichier txt

Exportation → Tableur

Détail ligne de commande