SL0720X
Étiquetage morphosyntaxique : fonctionnement de TreeTagger

Franck Sajous/CLLE-ERSS

http://w3.erss.univ-tlse2.fr/membre/fsajous/

Étiqueteurs morphosyntaxiques

Autres étiqueteurs (composants d'analyseurs syntaxiques)

- Cordial analyseur (Synapse) : analyseur performant, proriétaire et payant
- Talismane (Urieli, 2013) : analyseur *open-source*, gratuit, développé à CLLE-ERSS
- Analyseurs d'ALPAGE comme MaltParser, adapté de l'anglais

TreeTagger

- Étiqueteur ancien (Schmid, 1994), probablement le plus utilisé pour le français
- Pas le plus précis, propriétaire, mais :
  - gratuit
  - rapide
  - disponible pour plus de 15 langues
  - une interface graphique disponible (pour Windows uniquement)
- on ne sait rien des données utilisées pour la version française de TreeTagger !

Construction d'un étiqueteur

Cas général

- règles écrites à la main
- ou apprentissage automatique :
  - un corpus étiqueté « manuellement »
    (en réalité, automatisation le plus possible puis correction manuelle)
  - un système d'apprentissage
    → n'apprend pas par cœur les cas rencontrés mais généralise

Le	chat	mange	la	souris
DET	NOM	VER	DET	NOM
Le	pompier	éteint	le	feu
DET	NOM	VER	DET	NOM
Le	programmeur	conçoit	l'	étiqueteur
DET	NOM	VER	DET	?

TreeTagger : un article (Schmid, 1994)

- décrit les grandes lignes du processus d'étiquetage
- ne précise :
  - ni le corpus d'entraînement
  - ni le lexique utilisé
  - ni comment les deux types d'informations sont utilisés conjointement

Fonctionnement

Ressources

- lexique : liste de termes, avec pour chaque terme :
  - un/des couples <lemme, POS>
  - la probabilité pour chaque couple
  - eg. *souris* :

POS	lemme	proba
VER:pres	sourire	0.45
NOM	souris	0.28
VER:simp	sourire	0.15
VER:impe	sourire	0.12
...		

Fonctionnement

Ressources

- lexique : liste de termes, avec pour chaque terme :
  - un/des couples <lemme, POS>
  - la probabilité pour chaque couple
  - eg. *souris* :

POS	lemme	proba
VER:pres	sourire	0.45
NOM	souris	0.28
VER:simp	sourire	0.15
VER:impe	sourire	0.12
...		

- Suffixes (pour les mots inconnus)
  eg. en anglais : *-ness* VS *-less*

ness		less	
NN	0.97	JJ (Adj)	0.82
NP	0.03	RB (Adv)	0.12
		NP	0.06

Fonctionnement

Ressources

- lexique : liste de termes, avec pour chaque terme :
  - un/des couples <lemme, POS>
  - la probabilité pour chaque couple
  - eg. *souris* :

```
POS          lemme      proba
VER:pres     sourire    0.45
NOM          souris     0.28
VER:simp     sourire    0.15
VER:impe     sourire    0.12
...
```

- Suffixes (pour les mots inconnus)
  eg. en anglais : *-ness* VS *-less*

ness		less	
NN	0.97	JJ (Adj)	0.82
NP	0.03	RB (Adv)	0.12
		NP	0.06

- arbre de décision : contexte syntaxique (mots connus et inconnus)

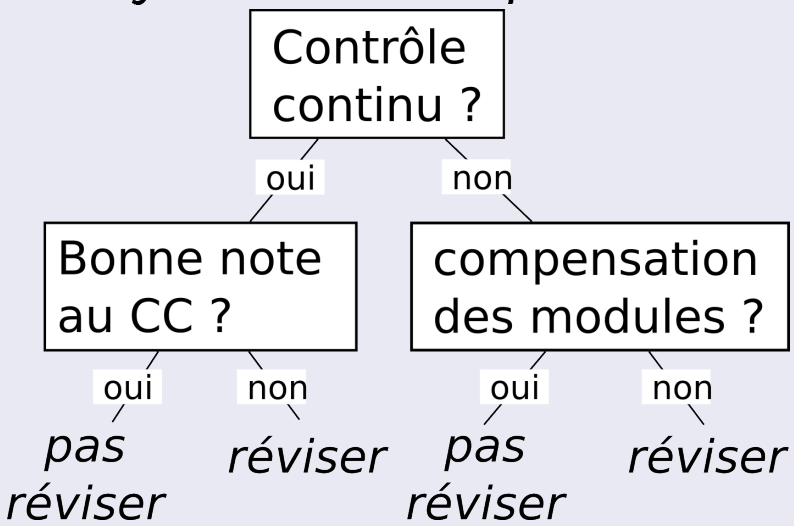
Arbres de décision

Définition

- Outil d'aide à la décision et à la classification
- Noeuds : tests sur les valeurs de paramètres
- Branches : résultat des tests

Exemple 1 : aide à la décision

Dois-je réviser mon partiel ?



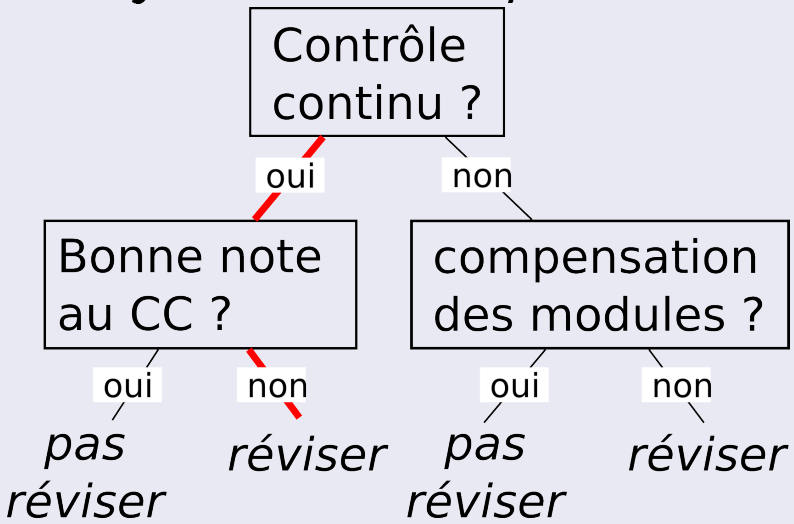
Arbres de décision

Définition

- Outil d'aide à la décision et à la classification
- Noeuds : tests sur les valeurs de paramètres
- Branches : résultat des tests

Exemple 1 : aide à la décision

Dois-je réviser mon partiel ?



Un chemin de l'arbre = une règle :

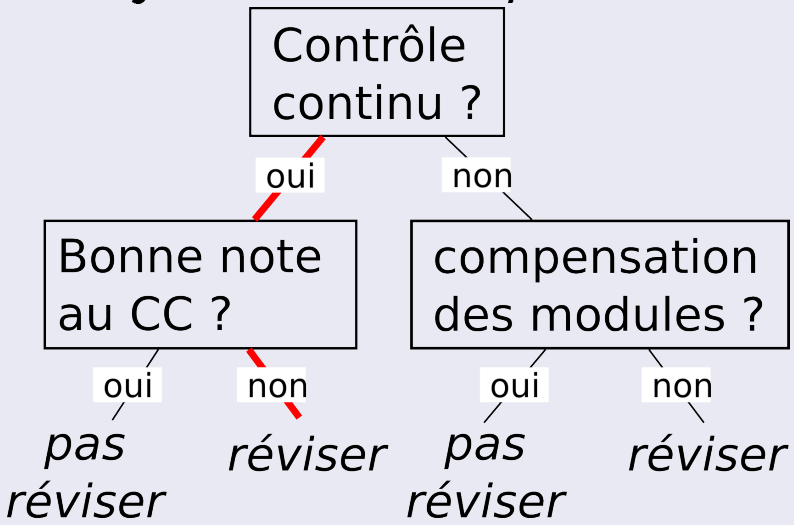
Arbres de décision

Définition

- Outil d'aide à la décision et à la classification
- Noeuds : tests sur les valeurs de paramètres
- Branches : résultat des tests

Exemple 1 : aide à la décision

Dois-je réviser mon partiel ?

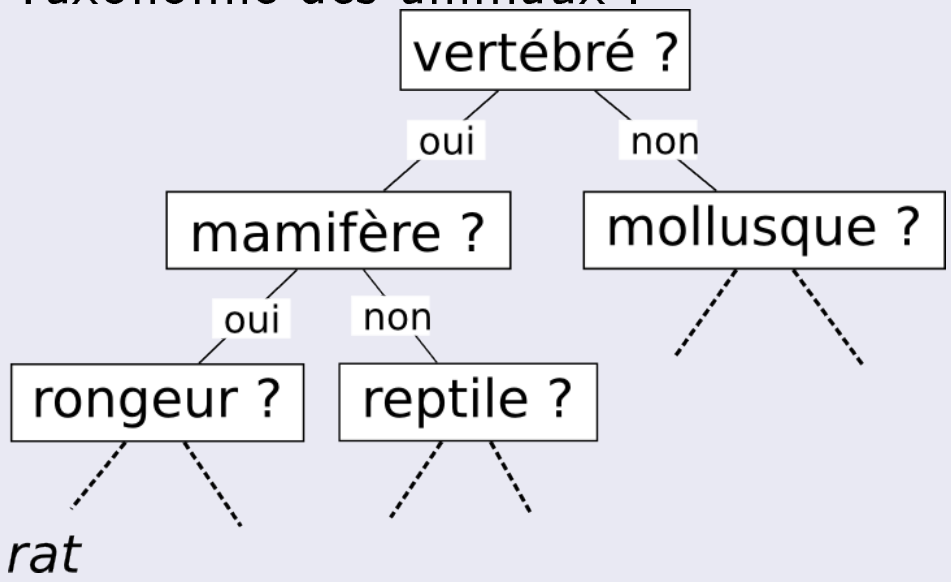


Un chemin de l'arbre = une règle :
s'il y a un CC
et que je n'ai pas eu une bonne note,
alors je révise le partiel

Arbres de décision (2)

Exemple 2 : classification

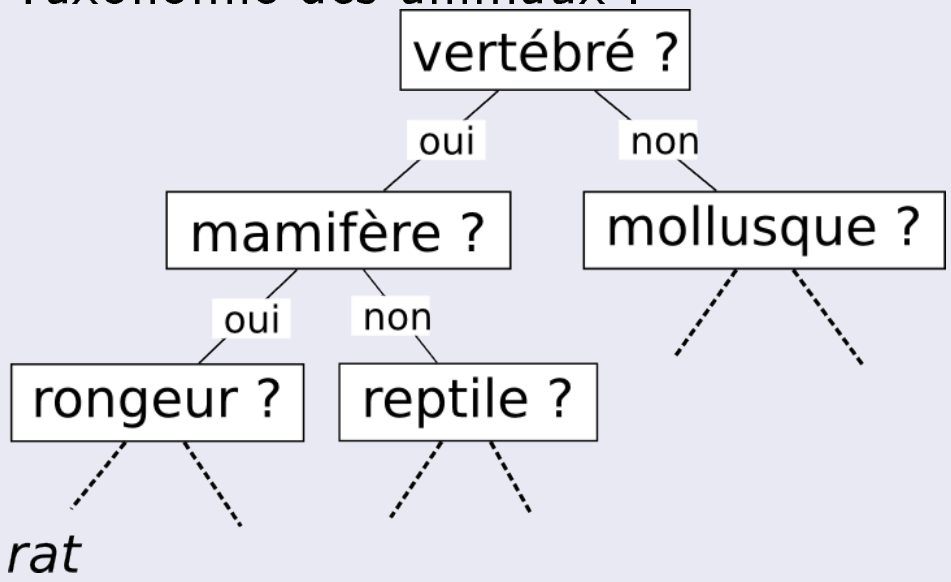
Taxonomie des animaux :



Arbres de décision (2)

Exemple 2 : classification

Taxonomie des animaux :



- Quel animal suis-je ?
- Es-tu un vertébré ?
- oui.
- Es-tu un mammifère ?
- etc.

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

- A des noeuds dans les cheveux ?

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

- A des noeuds dans les cheveux ?
  Non

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

- A des noeuds dans les cheveux ? Non
- A un chapeau à fleurs ?

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

- A des noeuds dans les cheveux ? Non
- A un chapeau à fleurs ? Non

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

- A des noeuds dans les cheveux ? Non
- A un chapeau à fleurs ? Non
- $\rightarrow$ tests déséquilibrés, décision potentiellement longue !

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

- A des noeuds dans les cheveux ? Non
- A un chapeau à fleurs ? Non
- $\rightarrow$ tests déséquilibrés, décision potentiellement longue !
- Femme ?

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

- A des noeuds dans les cheveux ? Non
- A un chapeau à fleurs ? Non
- $\rightarrow$ tests déséquilibrés, décision potentiellement longue !
- Femme ? Oui

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

- A des noeuds dans les cheveux ? Non
- A un chapeau à fleurs ? Non
- $\rightarrow$ tests déséquilibrés, décision potentiellement longue !
- Femme ? Oui
- Cheveux longs ?

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*

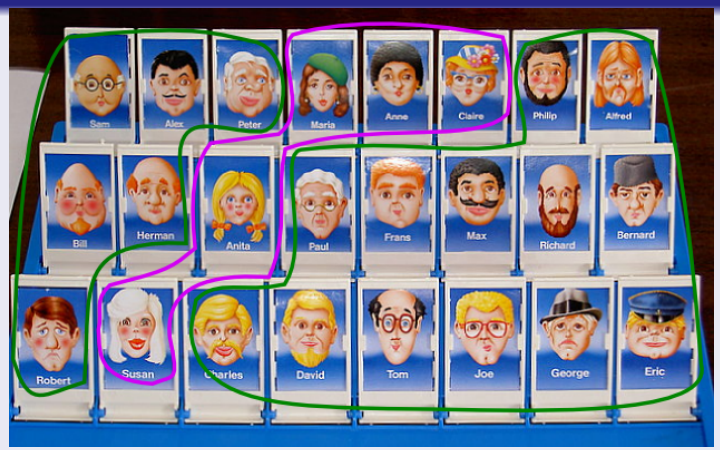
- A des noeuds dans les cheveux ? Non
- A un chapeau à fleurs ? Non
- $\rightarrow$ tests déséquilibrés, décision potentiellement longue !
- Femme ? Oui
- Cheveux longs ? Non

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*



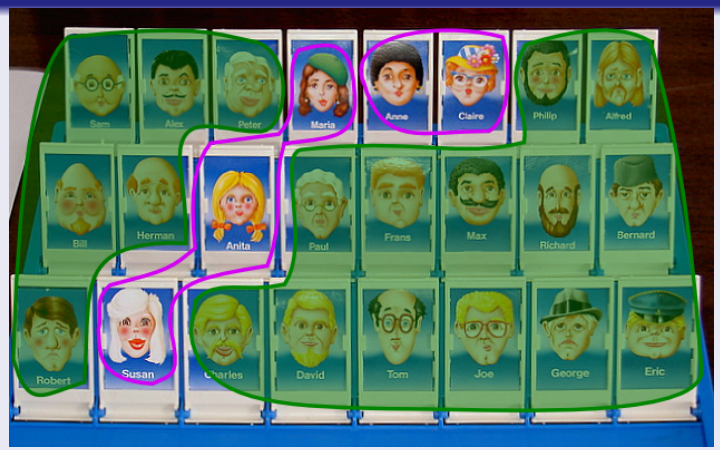
- A des noeuds dans les cheveux ? Non
- A un chapeau à fleurs ? Non
- $\rightarrow$ tests déséquilibrés, décision potentiellement longue !
- Femme ? Oui
- Cheveux longs ? Non
- tests moins déséquilibrés

Arbres de décision (3)

Structure de l'arbre

Pour une décision rapide : arbre peu profond, tests qui partitionnent équitablement (idéalement : 50/50)

Exemple : *Qui est-ce ?*



- A des noeuds dans les cheveux ? Non
- A un chapeau à fleurs ? Non
- → tests déséquilibrés, décision potentiellement longue !
- Femme ? Oui
- Cheveux longs ? Non
- tests moins déséquilibrés
- → décision + rapide

TreeTagger : mots « ambigus »

Séquence à étiqueter

La	petite	souris
Det:art	Adj	?

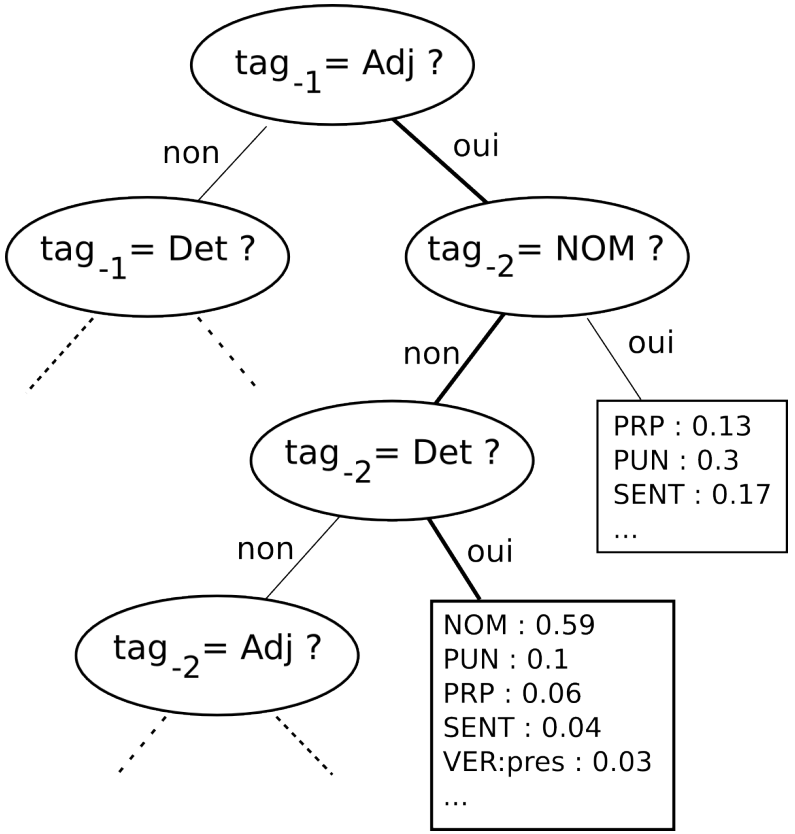
TreeTagger : mots « ambigus »

Séquence à étiqueter

La	petite	souris
Det:art	Adj	?

étiquetage : souris (token isolé)

VER:pres	sourire	0.517
NOM	souris	0.307
VER:simp	sourire	0.152
VER:impe	sourire	0.023



TreeTagger : mots « ambigus »

Séquence à étiqueter

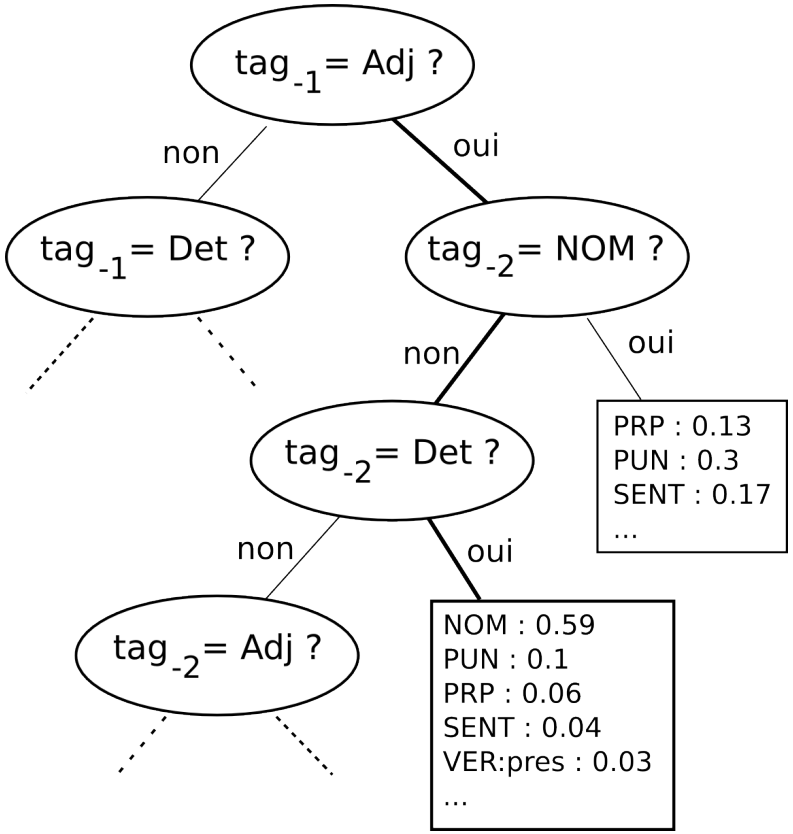
La	petite	souris
Det:art	Adj	?

étiquetage : souris (token isolé)

VER:pres	sourire	0.517
NOM	souris	0.307
VER:simp	sourire	0.152
VER:impe	sourire	0.023

Étiquetage final

NOM	souris	0.77
VER:pres	sourire	0.19
VER:simp	sourire	0.03
...		



TreeTagger : mots inconnus

Séquences à étiqueter

Des	très	grandes	ramiotes
PRP:det	ADV	ADJ	(1)
Tu	te	ramiotes	
PRO:PER	PRO:PER	(2)	
Des	voitures	très	ramiotes
PRP:det	NOM	ADV	(3)

TreeTagger : mots inconnus

Séquences à étiqueter

Des	très	grandes	ramiotes
PRP:det	ADV	ADJ	(1)
Tu	te	ramiotes	
PRO:PER	PRO:PER	(2)	
Des	voitures	très	ramiotes
PRP:det	NOM	ADV	(3)

Lexique de suffixes

```
iotes
        NOM         0.540404
        ADJ         0.287879
        VER:subp    0.085859
        VER:pres    0.085859
```

```
tag[-1] = ADJ
    tag[-2] = ADV
        absent !

tag[-1] = PRO:PER
    tag[-2] = PRO:PER
        VER:impf 0.222706
        VER:pres 0.599565
        ...

tag[-1] = ADV
    tag[-2] = NOM
        ADJ 0.239856
        ADV 0.094772
        VER:pres 0.122768
        ...
```

TreeTagger : mots inconnus

Séquences à étiqueter

Des	très	grandes	ramiotes
PRP:det	ADV	ADJ	(1)
Tu	te	ramiotes	
PRO:PER	PRO:PER	(2)	
Des	voitures	très	ramiotes
PRP:det	NOM	ADV	(3)

Lexique de suffixes

```
iotes
        NOM         0.540404
        ADJ         0.287879
        VER:subp    0.085859
        VER:pres    0.085859
```

```
tag[-1] = ADJ
    tag[-2] = ADV
        absent !

tag[-1] = PRO:PER
    tag[-2] = PRO:PER
        VER:impf 0.222706
        VER:pres 0.599565
        ...

tag[-1] = ADV
    tag[-2] = NOM
        ADJ 0.239856
        ADV 0.094772
        VER:pres 0.122768
        ...
```

Étiquetage final

1. NOM <unknown> 0.942 ADJ <unknown> 0.0489
2. VER:subp <unknown> 0.481 VER:pres <unknown> 0.366
3. ADJ <unknown> 0.846 NOM <unknown> 0.146

TreeTagger : mots inconnus (2)

Séquences à étiqueter

La	petite	zouris
Det:art	Adj	?
Tu	me	zouris
PRO:PER	PRO:PER	?

TreeTagger : mots inconnus (2)

Séquences à étiqueter

La	petite	zouris
Det:art	Adj	?
Tu	me	zouris
PRO:PER	PRO:PER	?

Lexique de suffixes

```
ouris
        NOM         0.892857
        VER:simp    0.035714
        VER:pres    0.035714
        VER:impe    0.035714
```

étiquetage : zouris (token isolé)

```
NOM    <unknkown>    1.000000
```

TreeTagger : mots inconnus (2)

Séquences à étiqueter

La	petite	zouris
Det:art	Adj	?
Tu	me	zouris
PRO:PER	PRO:PER	?

Lexique de suffixes

```
ouris
        NOM         0.892857
        VER:simp    0.035714
        VER:pres    0.035714
        VER:impe    0.035714
```

étiquetage : zouris (token isolé)

```
NOM   <unknkown>   1.000000
```

```
tag[-1] = ADJ
    tag[-2] = DET:ART
        NOM 0.633527
        VER:pres 0.031766

tag[-1] = PRO:PER
    tag[-2] = PRO:PER
        NOM 0.000358
        VER:pres 0.599565
        VER:impf 0.222706
```

TreeTagger : mots inconnus (2)

Séquences à étiqueter

La	petite	zouris
Det:art	Adj	?
Tu	me	zouris
PRO:PER	PRO:PER	?

Lexique de suffixes

```
ouris
        NOM         0.892857
        VER:simp    0.035714
        VER:pres    0.035714
        VER:impe    0.035714
```

étiquetage : zouris (token isolé)

```
NOM   <unknkown>   1.000000
```

```
tag[-1] = ADJ
    tag[-2] = DET:ART
        NOM 0.633527
        VER:pres 0.031766

tag[-1] = PRO:PER
    tag[-2] = PRO:PER
        NOM 0.000358
        VER:pres 0.599565
        VER:impf 0.222706
```

Étiquetage final

```
NOM <unknown> 1.00000
```

dans les deux cas !?