

INALCO

Cours sur le Traitement linguistique des Corpus

(TNML3)

Cours 8

Annotations morpho-syntaxique

Mercredi 16 Novembre 2016

Patrick Paroubek (LIMSI-CNRS)

Analyse

En TALN, analyser c'est de manière générale :

- Segmenter (*par ex. frontières de mots*)
- Identifier (*par ex. lister les étiquettes morpho-syntaxiques possibles*)
- Désambiguïser, éventuellement (*par ex. choisir la bonne étiquette morpho-syntaxique*)

Le problème est circulaire! (pour le texte et la parole)

A la base: Qu'est-ce qu'un mot ?

Importance en TALN du choix des unités élémentaires,
les (*tokens*).

Analyse

Un des mécanismes fondamentaux de l'analyse du langage est la **classification** des mots d'après leur comportement dans la langue. Cette classification peut être de nature très variée, combinant des éléments **morphologiques** (groupes de conjugaison des verbes), **morpho-syntaxiques** (distinction entre noms et verbes), **syntaxiques** (schémas de sous-catégorisation verbale : transitif, intransitif etc.), voire **sémantiques** (distinction entre toponymes et patronymes) et quelques fois mêmes **pragmatiques** (George Sand est un auteur féminin).

En TALN, une des analyses les plus simple que l'on peut effectuer et qui se trouve à la base de nombreuses applications plus complexes, en particulier l'analyse syntaxique, est **l'étiquetage morpho-syntaxique** (*Part-Of-Speech tagging, POS tagging*) .

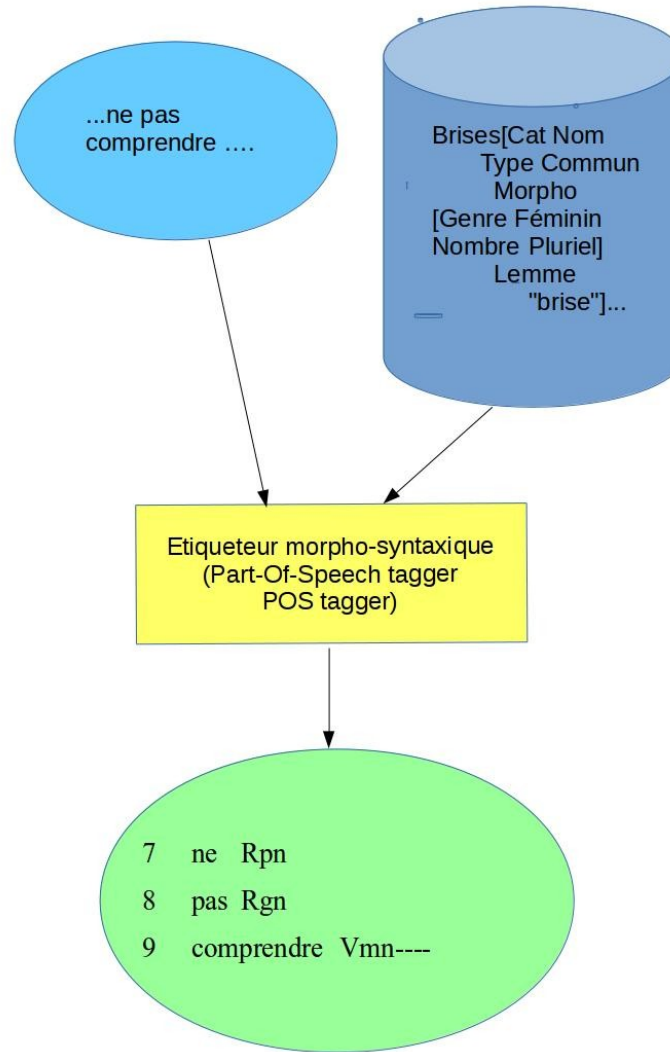
L'étiquetage morpho-syntaxique consiste à identifier la classe morpho-syntaxique qui est associée aux mots dans leur contexte d'énonciation.

Par exemple une phrase comme :

Le programme affiche des résultats

pourrait être annotée de la façon suivante :

Le [article défini masculin singulier] programme [nom commun masculin singulier] affiche [verbe conjugué 3ème personne singulier indicatif présent] des [article indéfini masculin pluriel] résultats [nom commun masculin pluriel].



POS

Ressource : Descriptions morpho-syntaxiques et jeux d'étiquettes

Brises	[Cat	Nom	
	Type	Commun	
	Morpho	[Genre	Féminin
		Nombre	Pluriel]
	Lemme	"brise"]	
	[Cat	Verbe	
	Type	Principal	
	SousCat	Transitif	
	Morpho	[Temps	Présent
		Mode	Indicatif
		Voix	Active
		Personne	2ème
		Genre	Singulier]
	Lemme	"briser"]	

POS

Exemple d'annotation morpho-syntaxique automatique

0	I	Nkms	11	n	Rpn/1.2
1	où	Pr-mp--	12	'	Rpn/2.2
1	l	Pp3msn-/1.3	13	est	Vmip3s-
3	'	Pp3msn-/2.3	14	pas	Rgn
4	on	Pp3msn-/3.3	15	sans	Sp
5	commence	Vmip3s-	16	une	Da-fs-i
6	à	Sp	17	certaine	Ai-fs
7	ne	Rpn	18	émotion	Ncfs
8	pas	Rgn	19	que	Pr-fs--
9	comprendre	Vmn----	20	je	Pp1msn-
10	ce	Pd-ms--			

Les différentes approches pour l'étiquetage morpho-syntaxique :

1. Segmentation + étiquetage a priori
2. Méthodes à base de règles sans apprentissage
3. Méthodes à base de règles avec apprentissage (Brill)
4. Méthodes probabilistes (modèles de Markov)
5. Méthodes à base de réseaux de neurones artificielles
6. Autres méthodes (e.g. algorithmes génétiques etc.)

Plusieurs architectures $\xrightarrow{\hspace{1cm}}$ possible pour l'articulation: EMS / AS :

1. en pipe-line, EMS → AS (la plus simple et la plus fréquente)
2. en boucle, EMS → AS (la plus réaliste, mais la plus complexe)
3. descendante, AS → EMS (rare)
4. fusionnelle (pas de distinction explicite entre les informations propres à l'EMS et celle de l'AS, le meilleur modèle ?)

POS

Les performances sont mesurées en taux d'étiquetage correct.
Elles sont presque toujours supérieures à 90% et atteignent parfois 99%.

[DEROSE 1988] env. 96% sur l'anglais (corpus Brown) avec le système VOLSUNGA. De même, les trois meilleurs résultats en précision obtenus dans GRACE [Paroubek98] étaient de 97,8%, 96,7% et 94,8%.

!

POS

Difficulté de la tâche: un simple accès lexical à un dictionnaire sans désambiguïsation a obtenu 88% en précision dans GRACE, mais ce score est tombé à 59% lorsque quelques règles non contextuelles de désambiguïsation ont été appliquées.

Pour une phrase de 15 mots et un taux d'étiquetage correct de 96% on obtient en fait un taux approché d'étiquetage correct au niveau des phrases de seulement 54.2% et, inversement, pour garantir un taux de 95% au niveau des phrases, il faudrait un système permettant un taux d'étiquetage correct (au niveau des mots) d'au moins 99.67%

En TALN, analyser c'est de manière générale :

1. Segmenter (*par ex. frontières de mots*)
2. Identifier (*par ex. lister les étiquettes morpho-syntaxiques possibles*)
3. Désambiguïser, éventuellement (*par ex. choisir la bonne étiquette morpho-syntaxique*)

Le problème est circulaire!

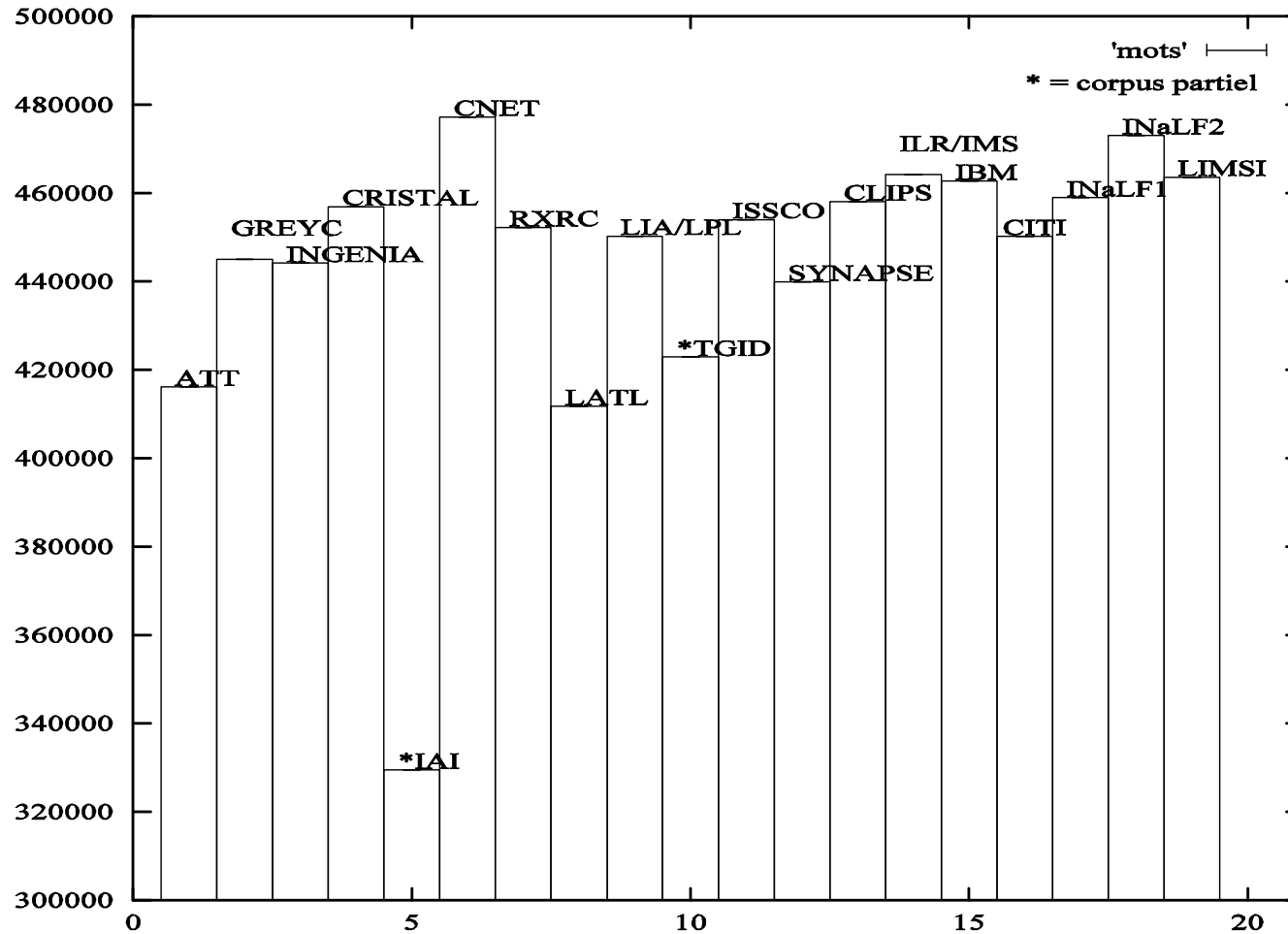
A la base: **Qu'est-ce qu'un mot ?**

Importance en TALN du choix des unités élémentaires, les (*tokens*).

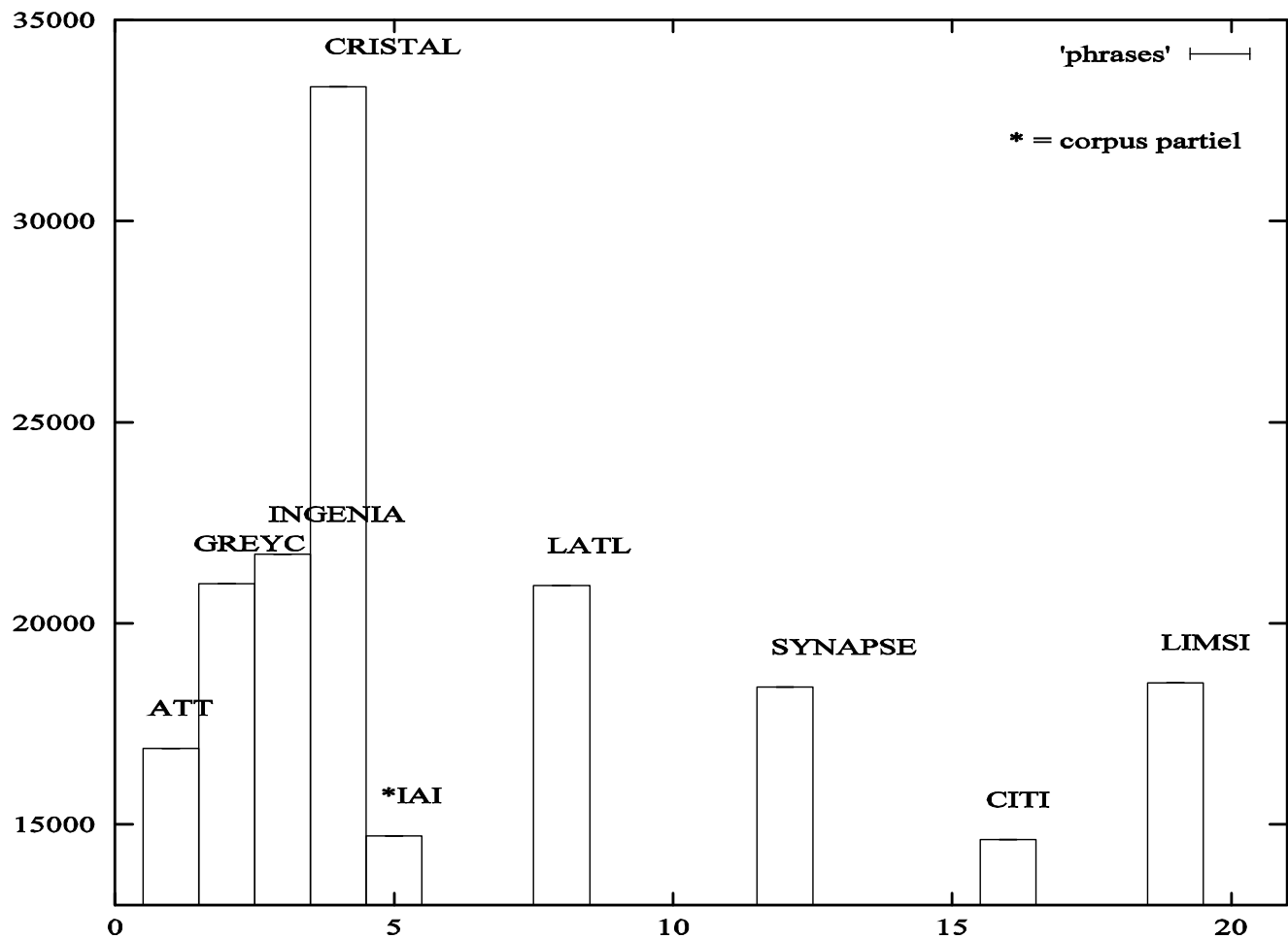
0 I Nkms
1 où Pr-mp--
1 I Pp3msn-/1.3
2 ' Pp3msn-/2.3
4 on Pp3msn-/3.3
5 commence
Vmip3s-
6 à Sp
7 ne Rpn
8 pasRgn
9 comprendre Vmn----
10 ce Pd-ms--

11 n Rpn/1.2
12 ' Rpn/2.2
13 est Vmip3s-
14 pas Rgn
15 sans Sp
16 une Da-fs-i
17 certaine Ai-fs
18 émotion Ncfs
19 que Pr-fs--
20 je Pp1msn-

Nombre de mots en fonction du participant (GRACE)



Nombre de phrases en fonction du participant (GRACE)



000000 Au DTC:sg
000001 cours SBC:sg
000002 de PREP

Alignement (15 systèmes
différents pour les tests)



000000 Au Sp+Da-ms-d
000001 cours Ncfs|Ncms
000002 de Da----i|Da-fp-i|Da-mp-i|Sp

Projection des étiquettes
dans le jeu GRACE



000000 Au Sp/1.3 6/14[0.428571]
000001 cours Ncms|Sp/2.3 6/15[0.4]
000002 de Sp 7/13[0.538462]

Combinaison
Vote &
mesure de
confiance

GRACE, évaluation d'étiquetage
morphosyntaxique pour le français, 21
participants, 5 pays:

3 phases: entraînements (10 millions de
mots), essais (450.000), test (836.500)

17 participants aux essais, 13 participants aux
tests finaux

mesure précision/décision, sur 20.000 mots,
puis 40.000 mots.étiquettes EAGLES et
MULTEXT

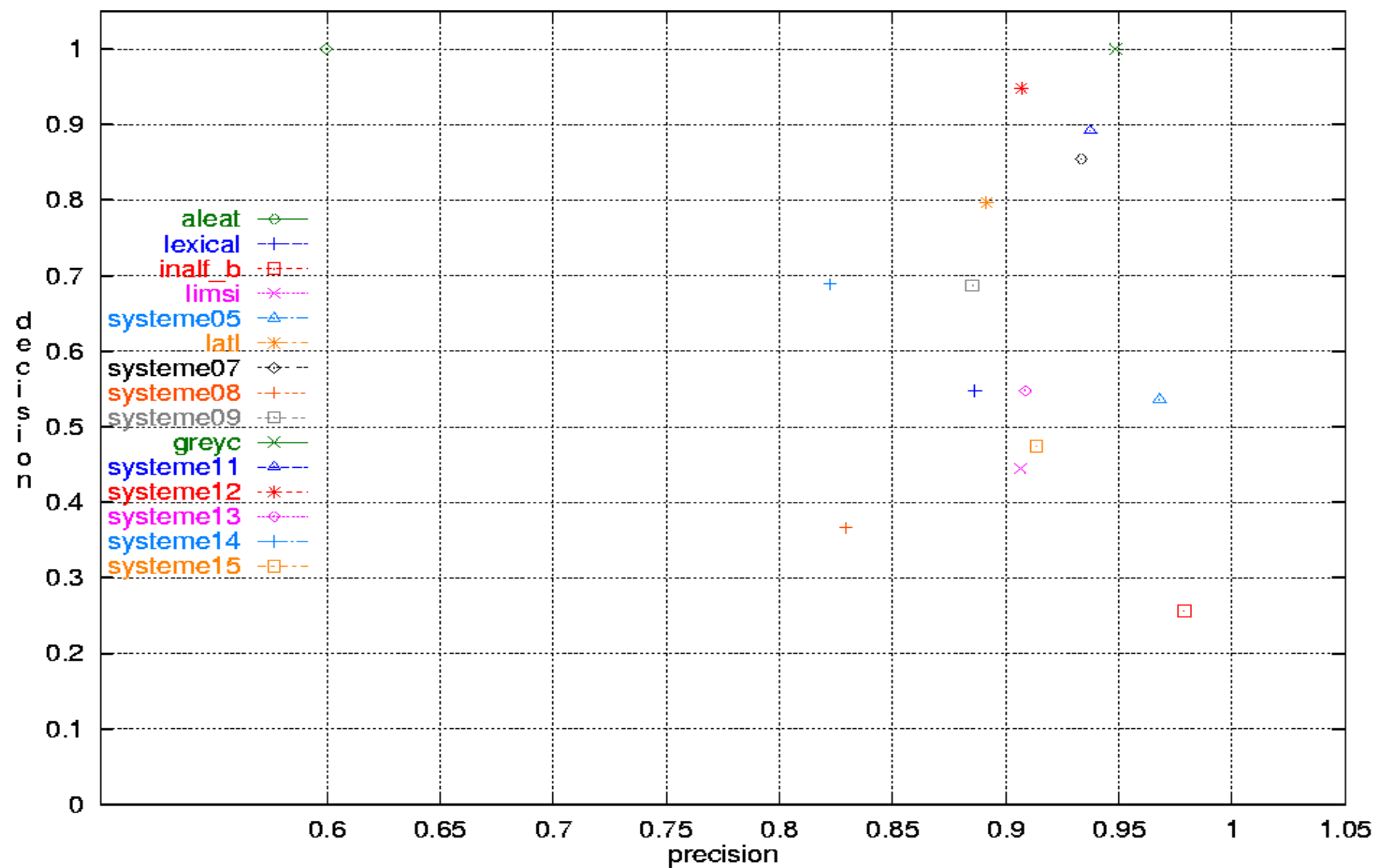
•Meilleur (P, Dmax): score(P, D): (**0.948489 , 1.000000**)
intervalle[Pmin, Pmoy, Pmax]: [0.948489 , 0.948489 ,
0.948489]

• Meilleur P: score(P, D): (**0.978802 , 0.256331**)
intervalle[Pmin, Pmoy, Pmax]: [0.251084 , 0.404534 ,
0.952951]

•Vote 15 systèmes: score(P, D): (**0.936202 , 0.961558**)
intervalle[Pmin, Pmoy, Pmax]: [0.903685 , 0.917102 ,
0.933155]

•Vote 5 meilleurs P: score(P, D): (**0.966567 , 0.928952**)
[Pmin, Pmoy, Pmax]: [0.902195 , 0.925850 , 0.961424]

Evaluation absolue



Mise en pratique:

Réaliser en commandes bash shell un étiqueteur en parties du discours à accès lexical.

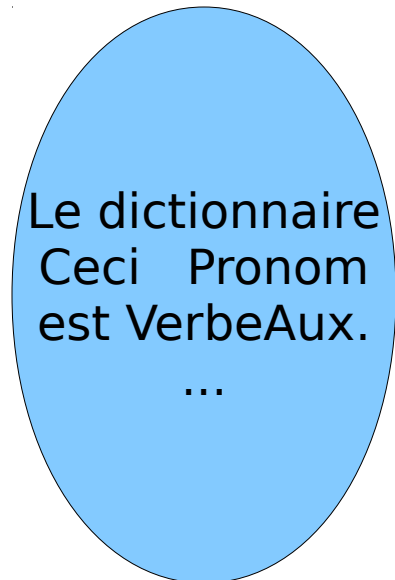
Un tel programme prendra en entrée un fichier contenant du texte et produira en sortie un fichier avec pour chaque mot l'étiquette morpho-syntaxique que le programme aura trouvé dans un dictionnaire.

Pour vous aider:

a) identifier combien de fichiers seront lus en entrée et combien seront écrits en sortie.


b) sur un exemple “jouet” quelques mots d'un texte et un petit Lexique contenant quelques mots avec leurs étiquettes possibles par ex. porte → N, V, simuler le travail de l'étiqueteur à la main sur ordinateur.

Entrées



tagger

Sortie



Ceci Pronom
est VerbeAux
Un ?
...

```
#!/bin/bash
```

```
# les entrées et sorties du programme d'étiquetage morpho-syntaxique
```

```
TEXT=$1
```

```
DICO=$2
```

```
RESULT=$3
```

```
# pour lire tous les mots du texte (boucle for)
```

```
cat $TEXT | tr '\040' '\012' > /tmp/text_mots.txt
```

```
for w in `cat /tmp/text_mots.txt`
```

```
Do
```

```
    echo " Le mot à étiqueter est $w"
```

```
done
```