

Traitement Automatique des Langues

Plan

- Introduction
- Etapes de l'analyse linguistique
- Exemple d'un pipeline d'analyse linguistique

Introduction

Langages

- Un langage est un système incluant un ensemble de symboles, une syntaxe (pour former des expressions complexes à partir des symboles) et une sémantique (définissant le sens des expressions du langage)
- Exemples : lambda calcul, logique des prédicats, langages de programmation
- Les langages formels sont généralement peu ambigus : la syntaxe et le sens de chaque expression tendent à être uniques

Langues naturelles

Parlée (et écrite) par des humains

- Anglais, français, allemand, chinois, etc.

Deux différences importantes entre langages formels et langues naturelles :

- Multidimensionalité : le décodage d'une expression fait intervenir l'analyse syntaxique et sémantique mais aussi l'analyse phonétique, phonologique, morphologique, pragmatique (interaction avec le contexte) ainsi que le raisonnement basé sur les connaissances
- Combinatoire forte
 - ▶ Ambiguïté : plusieurs analyses syntaxiques et/ou sémantiques possibles.
 - ▶ Paraphrases : le même contenu peut être exprimé de différentes façons.
 - ▶ Objectifs : gérer/réduire la combinatoire ; résoudre les ambiguïtés (analyse) ; faire les choix appropriés (génération)

Les niveaux d'analyse linguistique

Tous les niveaux d'analyse linguistique sont pertinents :

- Phonétique, Phonologie : sons / phonèmes / morphèmes
- Morphologie : morphèmes / mots
- Syntaxe : mots / constituants
- Sémantique : syntaxe / sens littéral
- Pragmatique : sens littéral, contexte / sens en contexte

Ambiguïté

L'ambiguïté est présente à tous les niveaux linguistiques.

- Phonologique: Le même *signal sonore* peut avoir plusieurs interprétations possibles :
Recognise speech ou *Wreck a nice peach ??*
- Sémantique lexicale: Le même *mot* peut dénoter différents objets.
étoile : célébrité ou astre?
- Partie du discours (catégorie morphosyntaxique): Le même *mot* peut avoir différentes catégories.
la : pronom, nom ou déterminant ?
- Syntaxe: La même *phrase* peut avoir plusieurs analyses syntaxiques.
Jean regarde (la fille avec un télescope)
Jean ((regarde la fille) avec un télescope)
- Sémantique phrastique: La même *phrase* peut avoir plusieurs analyses sémantiques.
La belle ferme la porte

Les applications du TAL

Traitent du texte et utilisent des connaissances linguistiques.

- Les interfaces vocales
- La reconnaissance de l'écriture manuelle (handwriting recognition)
- La correction orthographique
- La recherche d'information (K. Spark-Jones, 1972) e.g., les moteurs de recherche (Google, Yahoo, etc.)
- La traduction automatique e.g., Google Translate
- Les systèmes de dialogue homme-machine
- L'enseignement des langues assisté par ordinateur
- La détection d'opinions (à partir des blogs, des pages web, des réseaux sociaux)
- etc.

Etapes de l'analyse linguistique:

Analyse lexicale (morphologique)

Difficultés de la Segmentation / Normalisation

- ▶ Les écritures sans segmentation (chinois, thaï...)
- ▶ S'accomoder des ambiguïtés typographiques :
 - ▶ . : dans *etc.*, dans *20.3*, dans *enst.com*, dans ..., dans *TF.1*...
 - ▶ ' : dans *jusqu'à*, dans *aujourd'hui*, dans *3'4*, dans *Sotheby's* ou *Floc'h* ...
 - ▶ - : dans *Jean-Michel*, dans *donne-t-il*, dans *06-04-62-26-16-23*, dans *1914-1918*, dans *-1.2 %*...
 - ▶ sans parler de l'espace lui-même
- ▶ Détecter et normaliser les variantes typographiques : *France-Inter*, *France-inter* et *France Inter* ; *États-Unis* et *Etats-unis* et *Etats-Unis*...
- ▶ "Reconnaître" les chiffres, dates, durées, nombres, montants, numéros (de téléphone, de carte bleue), les scores...
- ▶ "Faire avec" les mots inconnus, les emprunts, les coquilles...

Le niveau lexical

- ▶ **But** : identifier les éléments lexicaux, leur structure et leurs caractéristiques ; regrouper les formes d'une même famille.
- ▶ **Moyen** : accès lexical direct, analyse morphologique (i.e. décomposition en *morphèmes*, à partir desquels les propriétés d'une forme sont calculées).
- ▶ **Outils** : un lexique, une description des morphèmes et des procédures de décomposition/recomposition associées.
- ▶ **Difficultés** : taille du lexique, vitesse d'accès et d'analyse, représentation du lexique, traitement des mots composés.
- ▶ **Résultat** : une représentation linéaire ou arborescente du mot, ses caractéristiques morpho-syntaxiques, une représentation de sa signification, un représentant de sa famille.

Le traitement lexical : résultat

- ▶ *le* - det. masc. sing., /lə/ ; pron. pers. masc. sing., /lə/
- ▶ *président* - vrb 3pers. plur. prés. ind./ subjonctif [présid+ent], <présider(X), présider(X,Y)>, /pʁezid+ət/ ; nom masc. sing., ← présider : action de X, <president(X)>, /pʁezidā/
- ▶ *des* - det. masc./fem. plur., /dɛ+z/ ; prep. contr. *de les*. ...
- ▶ *antialcooliques* - adj. masc./fem. plur. [anti+alcool+ique+s], ← alcoolique : s'opposer à X, antialcoolique(X), /ātiaɫkɔlikə+z/ ; nom. masc. sing. [anti+alcool+ique+s], ← antialcoolique (adj) : être X, antialcoolique(X), /ātiaɫkɔlikə+z/
- ▶ *mangeait* - vrb (1,3) pers. sing. imp. ind., [mang+e+ait], <manger(X),manger(X,Y)>, /māʒɛ+t/
- ▶ *pomme* - nom fem. sing., [pomme], <pomme(X),fruit(X),golden(X)...>, /pɔmə/
- ▶ ...

Etiquettage morpho-syntaxique

Catégories syntaxiques

- Les mots peuvent être regroupés en classes d'après leur comportement syntaxique.
- 8 grandes catégories : nom, verbe, pronom, préposition, adverbe, conjonction, adjectif et article.
- Autre catégories utilisées: eg Penn Treebank (45 étiquettes), Susanne (353 étiquettes).

Le jeu d'étiquettes du Penn Treebank (1)

CC	Coord Conjunction	<i>and, but, or</i>	NN	Nom, sing. or mass	<i>dog</i>
CD	Cardinal number	<i>one, two</i>	NNS	Nom, plural	<i>dogs</i>
DT	Article	<i>the, some</i>	NNP	Proper nom, sing.	<i>Edinburgh</i>
EX	Existential there	<i>there</i>	NNPS	Proper nom, plural	<i>Orkneys</i>
FW	Foreign Word	<i>mon dieu</i>	PDT	Prearticle	<i>all, both</i>
IN	Préposition	<i>of, in, by</i>	POS	Possessive ending	<i>'s</i>
JJ	Adjectif	<i>big</i>	PP	Personal pronom	<i>I, you, she</i>
JJR	Adj., comparative	<i>bigger</i>	PP\$	Possessive pronom	<i>my, one's</i>
JJS	Adj., superlative	<i>biggest</i>	RB	Adverbe	<i>quickly</i>
LS	List item marker	<i>1, One</i>	RBR	Adverbe, comparative	<i>faster</i>
MD	Modal	<i>can, should</i>	RBS	Adverbe, superlative	<i>fastest</i>

Le jeu d'étiquettes du Penn Treebank (2)

RP	Particle	<i>up, off</i>	WP\$	Possessive-Wh	<i>whose</i>
SYM	Symbol	<i>+, %, &</i>	WRB	Wh-adverbe	<i>how, where</i>
TO	"to"	<i>to</i>	\$	Dollar sign	<i>\$</i>
UH	Interjection	<i>oh, oops</i>	#	Pound sign	<i>#</i>
VB	verbe, base form	<i>eat</i>	"	Left quote	<i>' , "</i>
VBD	verbe, past tense	<i>ate</i>	"	Right quote	<i>' , "</i>
VBG	verbe, gerund	<i>eating</i>	(Left paren	<i>(</i>
VBN	verbe, past part	<i>eaten</i>)	Right paren	<i>)</i>
VBP	Verbe, non-3sg, pres	<i>eat</i>	,	Comma	<i>,</i>
VBZ	Verbe, 3sg, pres	<i>eats</i>	.	Sent-final punct	<i>. ! ?</i>
WDT	Wh-article	<i>which, that</i>	:	Mid-sent punct.	<i>: ; — ...</i>
WP	Wh-pronom	<i>what, who</i>			

Etiquettes morpho-syntaxiques universelles

Etiquette Penn TreeBank	Description	Etiquette Universelle
CC	conjunction, coordinating	CONJ
CD	cardinal number	NUM
DT	determiner	DET
EX	existential there	DT
FW	foreign word	X
IN	conjunction, subordinating or preposition	ADP
JJ	adjective	ADJ
JJR	adjective, comparative	ADJ
JJS	adjective, superlative	ADJ
LS	list item marker	X
MD	verb, modal auxillary	VERB
NN	noun, singular or mass	NOUN
NNS	noun, plural	NOUN
NNP	noun, proper singular	NOUN
NNPS	noun, proper plural	NOUN
PDT	predeterminer	DET
POS	possessive ending	PRT
PRP	pronoun, personal	PRON
PRPDOL	pronoun, possessive	PRON
RB	adverb	ADV
RBR	adverb, comparative	ADV
RBS	adverb, superlative	ADV
RP	adverb, particle	PRT
SYM	symbol	X
TO	infinitival to	PRT
UH	interjection	X
VB	verb, base form	VERB
VBZ	verb, 3rd person singular present	VERB
VBP	verb, non-3rd person singular present	VERB
VBD	verb, past tense	VERB
VBN	verb, past participle	VERB
VBG	verb, gerund or present participle	VERB
WDT	wh-determiner	DET
WP	wh-pronoun, personal	PRON
WPDOL	wh-pronoun, possessive	PRON
WRB	wh-adverb	ADV
.	punctuation mark, sentence closer	.
,	punctuation mark, comma	,
:	punctuation mark, colon	:
(contextual separator, left paren	(
)	contextual separator, right paren)

Étiquetage : Définition

- L'étiquetage associe une catégorie syntaxique unique à chaque mot d'un texte
- Exemple:
“ The/DT guys/NNS that/WDT make/VBP traditional/JJ hardware/NN
are/VBP really/RB being/VBG obsoleted/VBN by/IN
microprocessor-based/JJ machines/NNS ,/, said/VBD Mr./NNP
Benton/NNP ./.”

Étiquetage et ambiguïté

- L'étiquetage vise à lever l'ambiguïté morpho-syntaxique

- ▶ The *back*/*JJ* door
- ▶ On my *back*/*NN*
- ▶ With the voters *back*/*RB*
- ▶ He promised to *back*/*VB* the bill

- Brown Corpus

1 cat	2 cat.	3 cat.	4 cat.	5 cat.	6 cat.	7 cat.
35340	3760	264	61	12	2	1

NLTK: Python Natural Language ToolKit

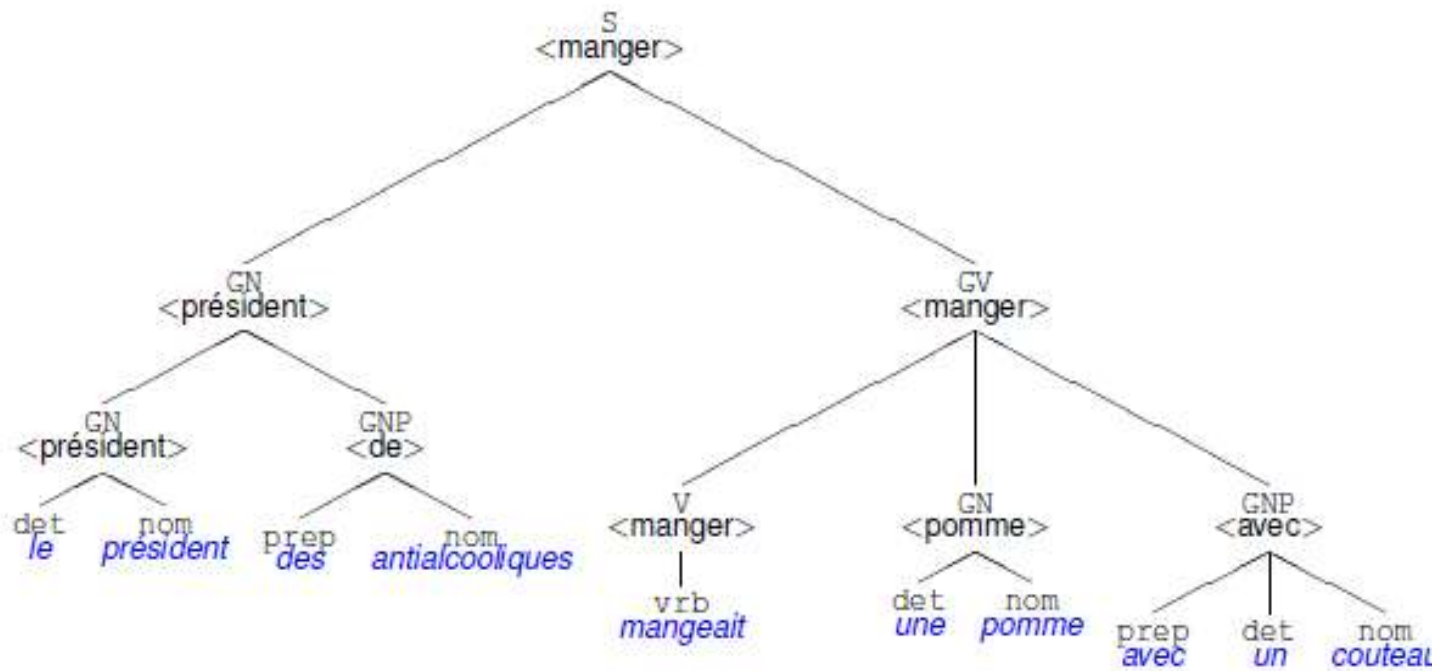
- Projet libre source
<http://nltk.sourceforge.net>
- Développeurs : Steven Bird, Ewan Klein, Ed Loper
- NLTK est un ensemble de modules python qui implémentent des algorithmes pour le TAL :
 - ▶ traitement des expressions régulières
 - ▶ extraction de phrases et de mots
 - ▶ étiquetage
 - ▶ analyse locale
 - ▶ analyse syntaxique
 - ▶ etc.

Analyse syntaxique

Le niveau syntaxique

- ▶ **But** : identifier les composants syntaxiques (syntagmes), leur fonction, et les relations qu'ils entretiennent entre eux.
- ▶ **Moyen** : analyse syntaxique, qui fournit une représentation arborescente des composants de l'énoncé.
- ▶ **Outils** : un analyseur syntaxique, c'est-à-dire un formalisme de description des règles syntaxiques, des règles valides pour un (sous)-langage donné, et un système d'analyse (un parseur) capable d'exploiter ces règles.
- ▶ **Difficultés** : compromis entre richesse de description, vitesse d'analyse, et prolifération des ambiguïtés, complexité des phénomènes à décrire, robustesse aux entrées "bruitées" (coquilles, casse...).
- ▶ **Résultat** : un (ou des) arbres syntaxiques représentant la phrase.

Le traitement syntaxique : résultat



L'ambiguïté lexicale

Un des principaux problèmes de l'analyse syntaxique est l'ambiguïté.

Ambiguïté lexicale :

- ▶ *souris* : formes verbales de *sourir*, nom féminin singulier et pluriel ;
- ▶ *petit* : adjectif ou nom masculin singulier ;
- ▶ *la* : déterminant ou pronom personnel féminin singulier, nom masculin ;
- ▶ *mousse* : formes verbales de *mousser*, nom masculin, nom féminin ;

Plus la description lexicale est précise, plus l'ambiguïté est grande : *monter* (*monter un escalier*, *monter un cheval*, *monter une pièce*, ...).

Cette ambiguïté n'est pas seulement statique, mais également *dynamique* : les phénomènes syntaxiques de *translation* rendent ambigus adjectifs et participes passés (emploi nominal) : *ces affreux se sont enfuis*

L'ambiguïté syntaxique

- ▶ *La petite brise la glace ;*
- ▶ *La troupe monte Molière vs Le jockey monte Belino ;*
- ▶ *Elle mange une pomme avec les doigts vs Elle mange une pomme avec la peau ;*
- ▶ *Elle mange une glace à la fraise vs Elle mange une glace à la plage ;*
- ▶ *C'est la fille du cousin qui boit ;*
- ▶ *Il a parlé de déjeuner avec Paul ;*

La désambiguïsation est possible au niveau sémantique ou pragmatique ;
chaque raffinement de la grammaire accroît l'ambiguïté.

Grammaire à contexte libre

Context Free Grammars

A context free grammar $G = (N, \Sigma, R, S)$ where:

- ▶ N is a set of non-terminal symbols
- ▶ Σ is a set of terminal symbols
- ▶ R is a set of rules of the form $X \rightarrow Y_1 Y_2 \cdots Y_n$ for $n \geq 0$, $X \in N$, $Y_i \in (N \cup \Sigma)$
- ▶ $S \in N$ is a special start symbol

Grammaire à contexte libre: Exemple

$N = \{S, NP, VP, Adj, Det, Vb, Noun\}$

$\Sigma = \{fruit, flies, like, a, banana, tomato, angry\}$

$S = 'S'$

$R =$

$S \rightarrow NP VP$

$NP \rightarrow Adj Noun$

$NP \rightarrow Det Noun$

$VP \rightarrow Vb NP$

$Adj \rightarrow fruit$

$Noun \rightarrow flies$

$Vb \rightarrow like$

$Det \rightarrow a$

$Noun \rightarrow banana$

$Noun \rightarrow tomato$

$Adj \rightarrow angry$

Grammaire à contexte libre: Dérivation

Left-most derivation is a sequence of strings s_1, \dots, s_n where

- ▶ $s_1 = S$ the start symbol
- ▶ $s_n \in \Sigma^*$, meaning s_n is only terminal symbols
- ▶ Each s_i for $i = 2 \dots n$ is derived from s_{i-1} by picking the left-most non-terminal X in s_{i-1} and replacing it by some β where $X \rightarrow \beta$ is a rule in R .

For example: $[S], [NP VP], [Adj Noun VP], [fruit Noun VP], [fruit flies VP], [fruit flies Vb NP], [fruit flies like NP], [fruit flies like Det Noun], [fruit flies like a], [fruit flies like a banana]$

Grammaire à contexte libre: Arbre syntaxique

$S \rightarrow NP VP$

$NP \rightarrow Adj Noun$

$NP \rightarrow Det Noun$

$VP \rightarrow Vb NP$

-

$Adj \rightarrow fruit$

$Noun \rightarrow flies$

$Vb \rightarrow like$

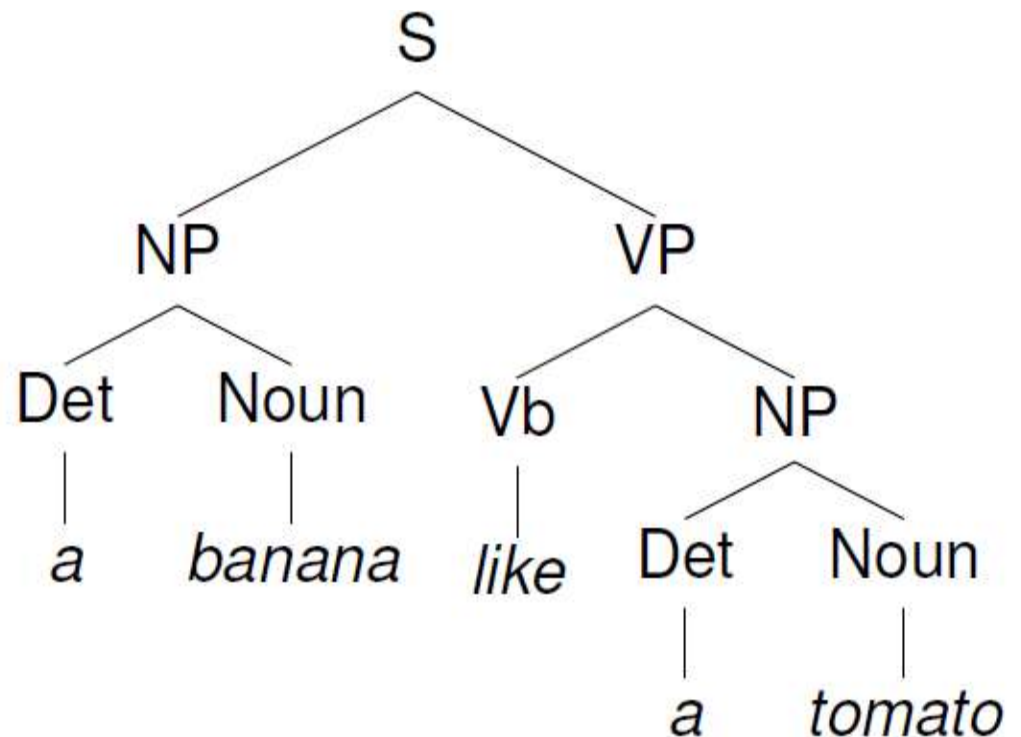
$Det \rightarrow a$

$Noun \rightarrow banana$

$Noun \rightarrow tomato$

$Adj \rightarrow angry$

...



Grammaire à contexte libre: Arbre syntaxique

$S \rightarrow NP VP$

$NP \rightarrow Adj Noun$

$NP \rightarrow Det Noun$

$VP \rightarrow Vb NP$

-

$Adj \rightarrow fruit$

$Noun \rightarrow flies$

$Vb \rightarrow like$

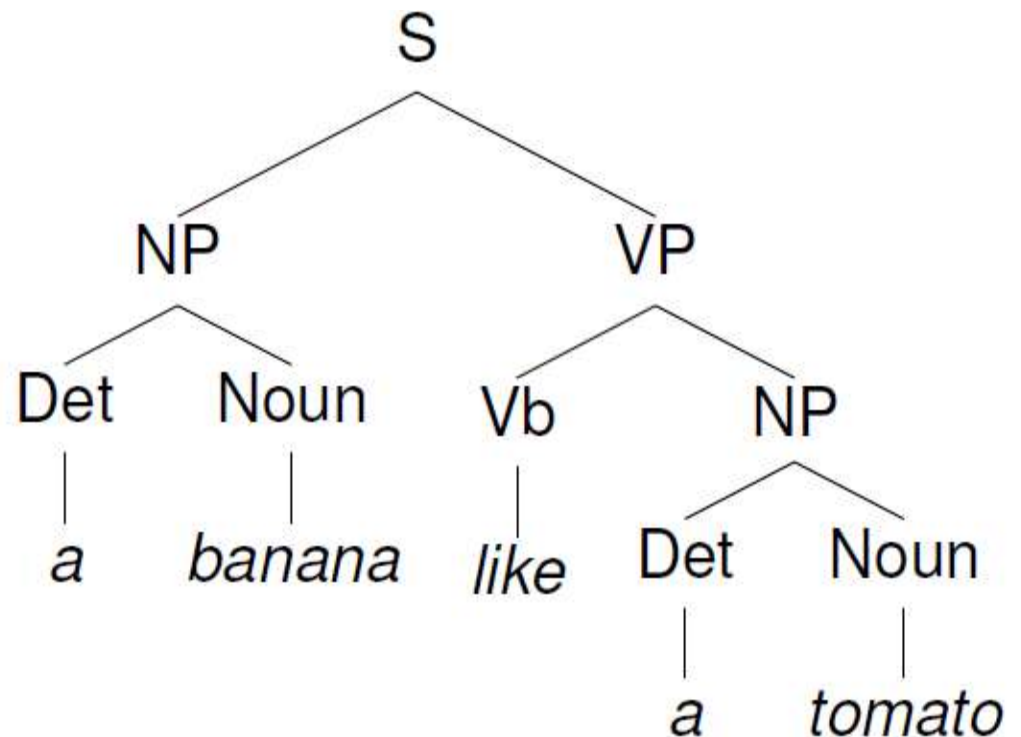
$Det \rightarrow a$

$Noun \rightarrow banana$

$Noun \rightarrow tomato$

$Adj \rightarrow angry$

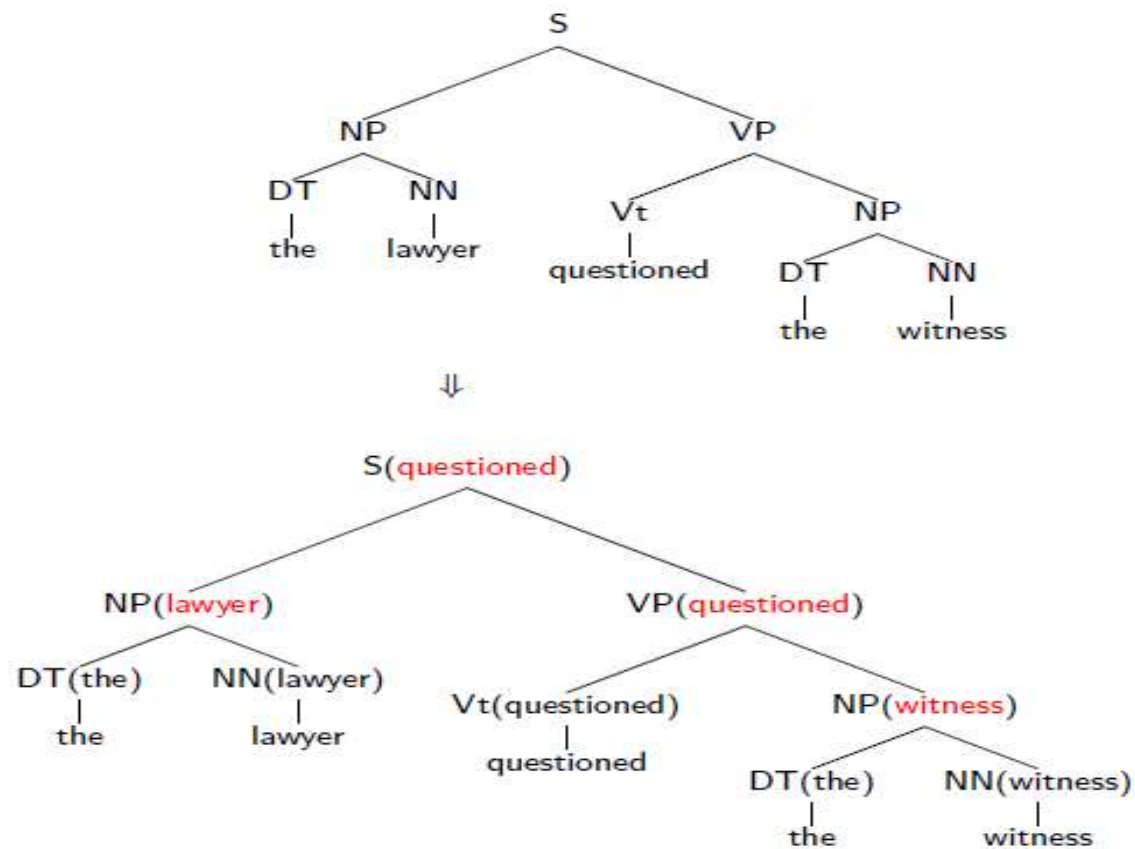
...



Analyse en dépendance syntaxique

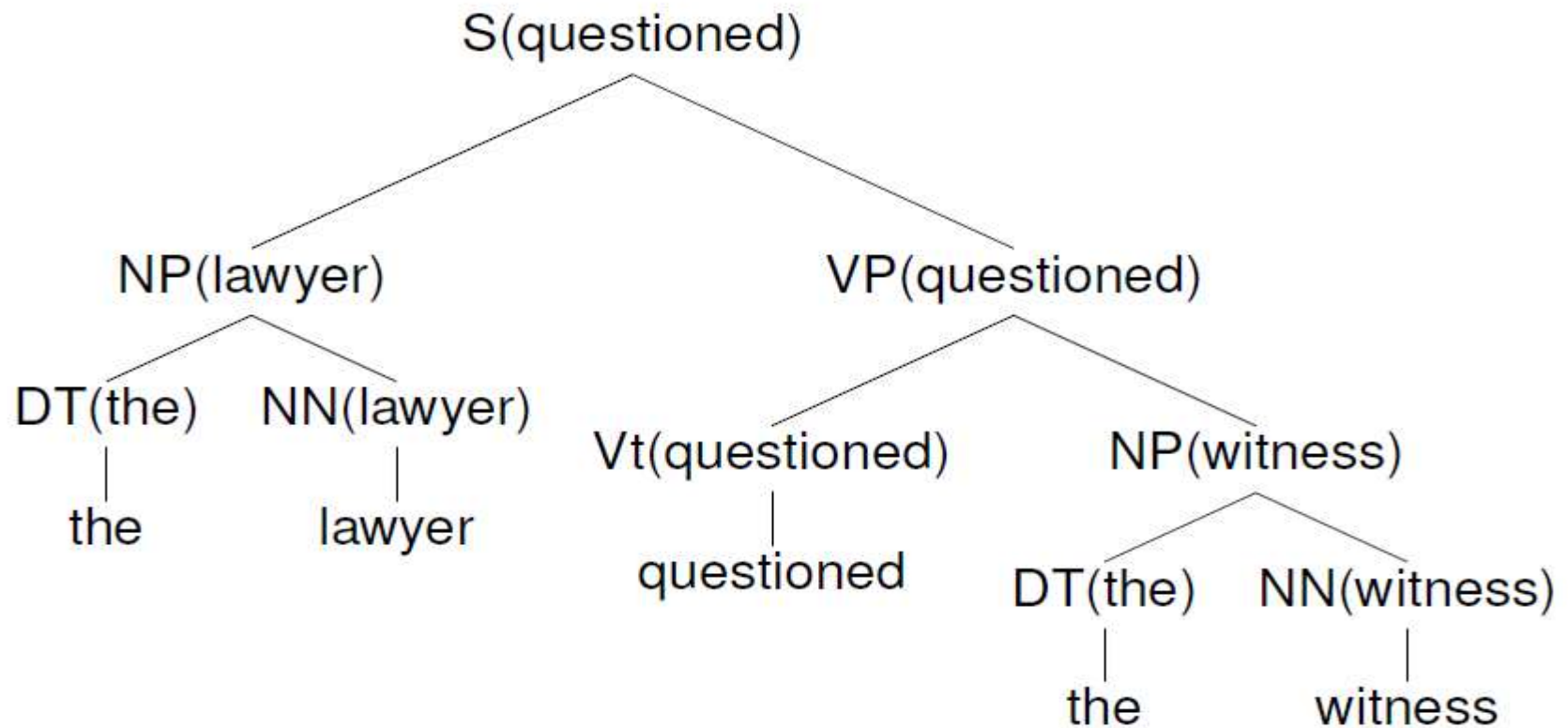
Dependency Parsing

→ Capturer la relation entre les mots dans une phrase

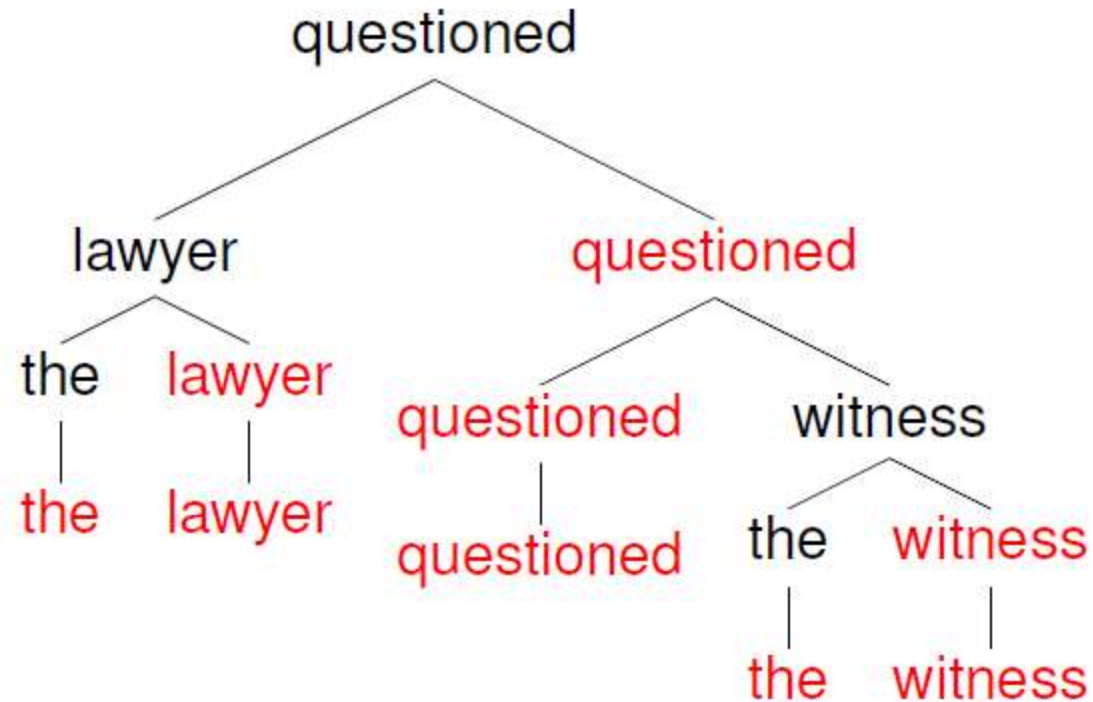


Analyse en dépendance syntaxique

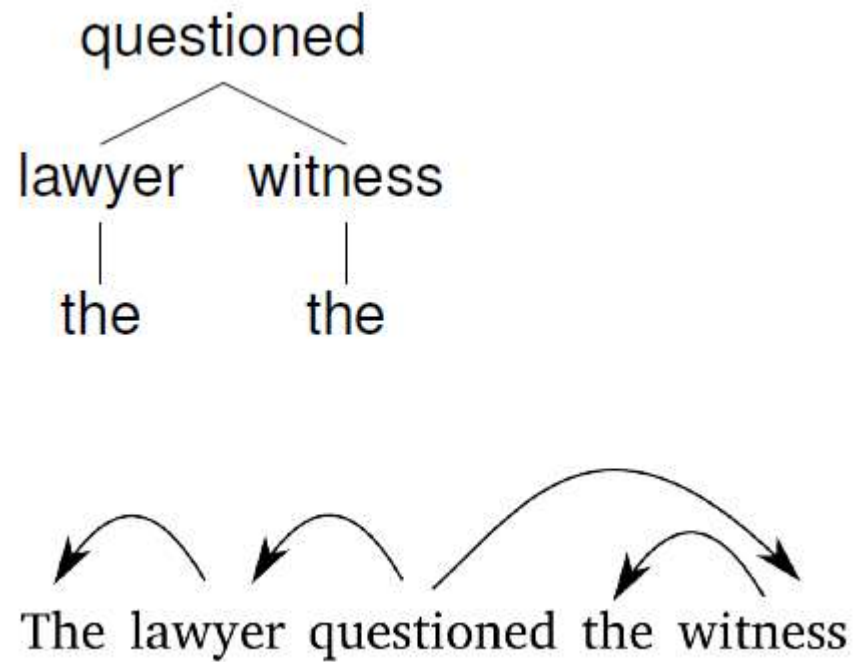
Dependency Parsing



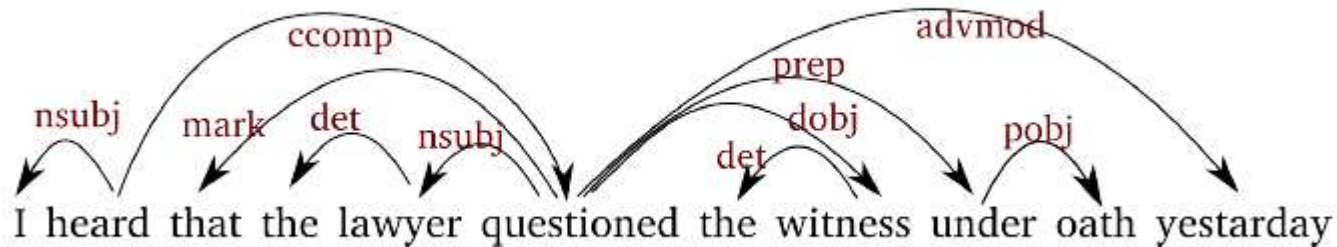
Représentation en dépendance syntaxique



Représentation en dépendance syntaxique



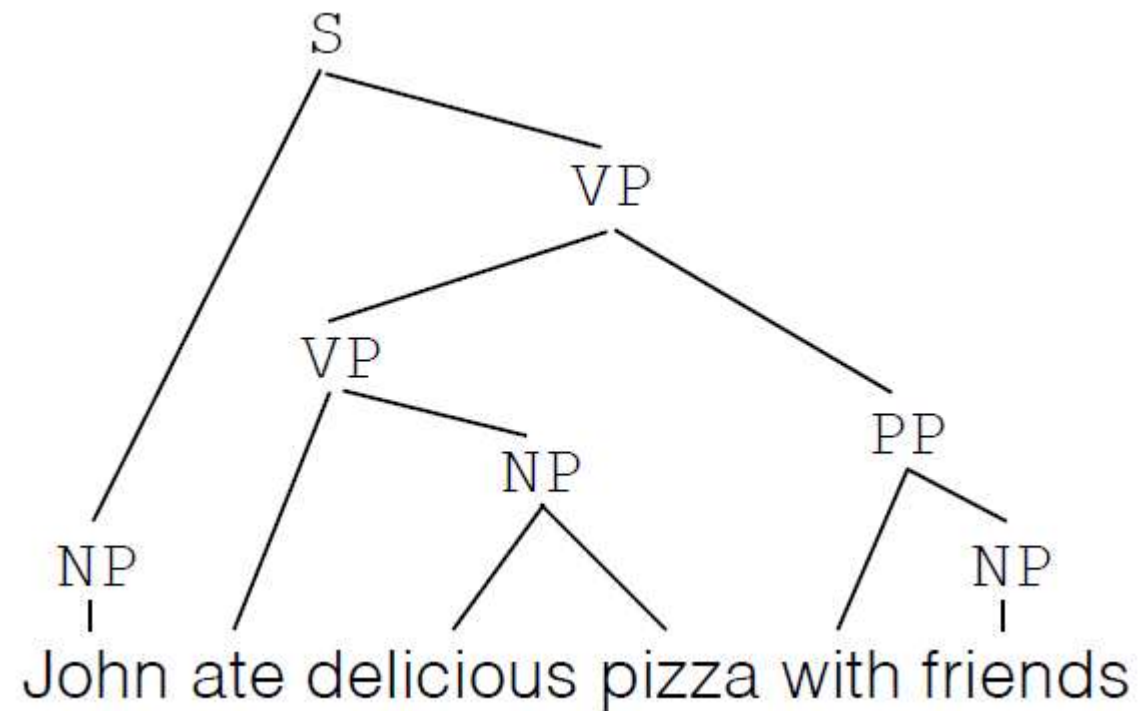
Représentation en dépendance syntaxique



Dependency relations (Stanford Dependencies):

- **nsbj**: nominal subject
- **mark**: subordinating conjunction
- **ccomp**: clausal complement
- **det**: determiner
- **prep**: preposition prepositional modifier
- **dobj**: direct object
- **advmod**: adverbial modifier
- **pobj**: object of a preposition

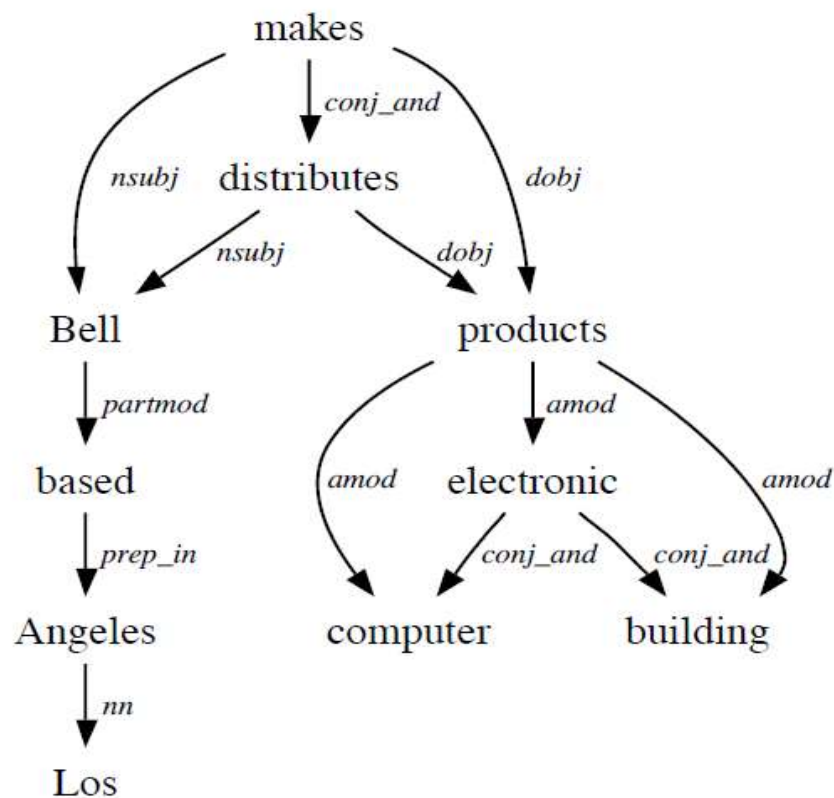
Analyse syntaxique: Arbre syntaxique



Analyse syntaxique: Graphe de dépendance

Phrase:

Bell, based in Los Angeles, makes and distributes electronic, computer and building products.



nsubj (makes-8, Bell-1)

nsubj (distributes-10, Bell-1)

vmod (Bell-1, based-3)

nn (Angeles-6, Los-5)

prep in (based-3, Angeles-6)

root (ROOT-0, makes-8)

conj and (makes-8, distributes-10)

amod (products-16, electronic-11)

conj and (electronic-11, computer-13)

amod (products-16, computer-13)

conj and (electronic-11, building-15)

amod (products-16, building-15)

dobj (makes-8, products-16)

dobj (distributes-10, products-16)

Etiquettes universelles pour les relations de dépendance syntaxique

Etiquette Universelle	Description
root	the head of a sentence
nsubj	nominal subject
nsubjpass	passive nominal subject
csbj	clausal subject
csbjpass	clausal passive subject
dobj	direct object
iobj	indirect object
ccomp	clausal complement
xcomp	open clausal complement
nmod	nominal modifier
advmod	adverbial modifier
advcl	adverbial clause modifier
neg	negation
appos	apposition
amod	adjectival modifier
acl	clausal modifier of a noun (adjectival clause)
det	determiner
case	case marking
vocative	addressee
aux	auxiliary verb
auxpass	passive auxiliary
cop	copula verb
mark	subordinating conjunction
expl	expletive
conj	conjunct
cc	coordinating conjunction
discourse	discourse element
compound	relation for marking compound words
name	names
mwe	multiword expressions that are not names
foreign	text in a foreign language
goeswith	two parts of a word that are separated in text
list	used for chains of comparable elements
dislocated	dislocated elements
parataxis	parataxis
remnant	remnant in ellipsis
reparandum	overridden disfluency
punct	punctuation
dep	unspecified dependency

Reconnaissance d'Entités Nommées

Extraction d'Entités Nommées

- La tâche d'Extraction d'Information a mis en évidence l'intérêt de reconnaître les Entités Nommées :

- **Qu'est-ce que c'est une Entité Nommée ?**

...tous les éléments du langage qui font référence à une entité unique et concrète, appartenant à un domaine spécifique (humain, économique, géographique, etc.)

...noms propres mais aussi expressions de temps et de quantité

(MUC-7, Chinchor 1998)

Extraction d'Entités Nommées

De manière générale, il s'agit de noms propres pouvant être classés

- dans des catégories prédéfinies
 - ENAMEX : organisation, lieu, personne
 - TIMEX : dates, expressions temporelles
 - NUMEX : valeurs monétaires, pourcentage, ...
- dans des catégories spécifiques à un domaine
 - biologie : espèces, protéines, gènes, etc.
 - médecine : médicaments, conditions médicales, etc.
 - mais aussi noms de bateau, modèles d'avion, etc.

Extraction d'Entités Nommées

Exemple

Né à Paris[lieu] le 21 octobre 1944[date], Jean-Pierre Sauvage[personne] a effectué sa thèse à l'Université de Strasbourg[lieu][organisation,lieu] sous la direction de Jean-Marie Lehn[personne]. Après un post-doctorat à Oxford[organisation,lieu], il revient en France[lieu] et effectue sa carrière au CNRS[organisation] qu'il intègre en 1971[date] et devient directeur de recherche au CNRS[organisation] en 1979[date]. Jean-Pierre Sauvage[personne] travaille à l'Institut de science et d'ingénierie supramoléculaire[organisation] (CNRS[organisation]/Université de Strasbourg[lieu][organisation,lieu]). Il a également reçu la médaille de bronze en 1978[date] et celle d'argent du CNRS[organisation] en 1988[date].

On peut reconnaître

- les entités nommées, imbriquées ou non
- les types associés aux entités, parfois ambiguës

Extraction d'Entités Nommées

Obstacle important en TAL :

- Majorité des mots inconnus d'un corpus
- Porteurs d'informations importantes
- Similaires aux groupes nominaux complexes avec beaucoup de variation
 - (Wikipedia EN) Carl XVI Gustaf of Sweden, Carl XVI Gustaf, Carl Gustaf Folke Hubertus, King Carl Gustaf, His Majesty Carl XVI Gustaf, King of Sweden, Carl Gustaf
 - (Wikipedia FR) Barack Obama, Barack Hussein Obama II, Barack Obama Jr., Obama, président Obama, président Barack Obama
- Acronymes peuvent être similaires aux mots : *OTAN*, *Laser*, *Radar*
- Nécessitent plusieurs analyses

Difficultés dans l'Extraction d'Entités Nommées

- La portée des classes : Clint Eastwood, l'épouse Chirac, les frères Cohen, les démocrates, les Boeings, Bison futé
- La coordination : Barack et Michelle Obama, M. et Mme Obama
- L'imbrication : Université de Strasbourg
- Les frontières : l'équipe de Nantes, le Palais Bourbon, monsieur Hollande/le président Hollande, le couple Obama
- Les variantes : l'équipe de Nantes/le stade nantais/les canaris/les nantais/Nantes/FCN
- La polysémie : Clint Eastwood (acteur, réalisateur, producteur, mais aussi chanteur jamaïcain, chanson, personne de film), Leclerc (maréchal, homme d'affaire, Char, supermarché)

Approches symboliques pour l'Extraction d'Entités Nommées

Exemple de Règles

- Rule format

`<trigger> : <preceding context> : <following context> : <type> : =><action>`

- regular expressions based on linguistic information

- Examples for entity LOCATION

```
@RegionsAndCountries=(Afghanistan,Africa,Albania,Alberta,...)
@City=(Aaccra,Aalborg,Aarhus,Ababa,Abadan,Abakan,Aberdeen,...)
@GeographicalPrecision=(South,West,North,East,south,west,north,east,Southern,
Western,Northern,Eastern,southern,western,northern,eastern)
@locationKey=(sea,ocean,Village,Valley,Trail,Station,Stadium,Square,River,...)

@GeographicalPrecision::(@City|@CountryOrRegion):LOCATION:
$L_NP:(@GeographicalPrecision)?:@locationKey:LOCATION:
```

Approches symboliques pour l'Extraction d'Entités Nommées

Exemple d'extraction de Règles à partir de corpus annoté

■ Example of rule composition from annotated text

Bob Boulton, born in 1950, has been appointed President of Minelco, effective December 1, 2010. He succeeds Markus Petäjänemi, who has been appointed LKAB's Senior Vice President.

@function=(President, Director,...)

@functionModifier=(Senior, Junior, Executive)

@functionModifier2=(Vice)

@function:@functionModifier{0-2} @functionModifier2? @function::FUNCTION:

[<FUNCTION>]::[of] \$PROPER_NOUN (\$PROPER_NOUN){0-2}:ORGANIZATION:

[<FUNCTION>]:\$PROPER_NOUN (\$PROPER_NOUN){0-2} [s]::ORGANIZATION:

Identification des Entités Nommées

Problèmes

- conflit entre indices internes et externes
La société Yves Saint-Laurent, le groupe Hugo Boss, la société Hughes Aircraft
→ On privilégie l'indice externe
- Ambiguïté du contexte :
 - *All American Bank* vs. *All State Police*
 - *JFK* (mais aussi *Charles De Gaulle*)→ Un contexte plus large doit être utilisé
- Ambiguïté de la coordination
 - *C&A, H&M, Pratt & Whitney* vs. *Apple et Samsung*

Approches statistiques pour l'Extraction d'Entités Nommées

Utilisation de méthodes d'étiquetage séquentiel (par apprentissage)

- ① Données annotées selon la représentation BIO(/IOB)
- ② Apprentissage d'un modèle (HMM, CRF, etc.) sur les données annotées
- ③ Utilisation du modèle pour étiqueter les données selon la représentation BIO
- ④ Post-traitement pour interpréter la représentation BIO

Approches statistiques pour l'Extraction d'Entités Nommées

Représentation BIO

- Chaque mot est associé à une classe
 - **B** (Begin), **I** (Inside), **O** (Outside)
 - ou en prenant en compte la catégorie sémantique :
 - Personne : **B-PERS** (Begin), **I-PERS** (Inside)
 - Organisation : **B-ORG** (Begin), **I-ORG** (Inside)
 - ...
 - **O** (Outside)

Approches statistiques pour l'Extraction d'Entités Nommées

Exemple

Né	
à	
Paris	LOC
le	
21	DATE
octobre	DATE
1944	DATE
,	
Jean-Pierre	PERS
Sauvage	PERS
a	
effectué	
sa	
thèse	

Né	O
à	O
Paris	B-LOC
le	O
21	B-DATE
octobre	I-DATE
1944	I-DATE
,	O
Jean-Pierre	B-PERS
Sauvage	I-PERS
a	O
effectué	O
sa	O
thèse	O

Approches statistiques pour l'Extraction d'Entités Nommées

Modèle d'apprentissage avec des CRF

- Définition d'un modèle probabiliste décrivant des caractéristiques de surface spécifiques aux entités nommées comme les CRF
- CRF - Conditional Random Field :
 - Objectif : maximiser $p(t|w)$ sans calculer de modèle $p(w)$ permet l'utilisation d'un ensemble de *features* plus important
 - Modèle graphique (CRF linéaire)

$$p(t|w) = \frac{\prod_{i=2}^N \exp(\sum_k \lambda_k f_k(t_{i-1}, t_i, w, i))}{\sum_{t'} \exp(\prod_{i=2}^N \exp(\sum_k \lambda_k f_k(t'_{i-1}, t'_i, w, i)))}$$

- Les *features* f_k doivent être définies par l'utilisateur
- Les paramètres du modèle (λ_k) sont estimés sur des données d'entraînement

t: Observations (Annotations), w: Variables (tokens)

Approches statistiques pour l'Extraction d'Entités Nommées

Exemple d'utilisation de patrons de features

Né	VER:pper	naître	O
à	PRP	à	O
Paris	NAM	Paris	B-LOC
le	DET:ART	le	O
21	NUM	@card@	B-DATE
octobre	NOM	octobre	I-DATE
1944	NUM	@card@	I-DATE
,	PUN	,	O
Jean-Pierre	NAM	Jean-Pierre	B-PERS
Sauvage	NAM	Sauvage	I-PERS
a	VER:pres	avoir	O
effectué	VER:pper	effectuer	O
sa	DET:POS	son	O
thèse	NOM	thèse	O
à	PRP	à	O
l'	DET:ART	le	O
Université	NOM	université	B-ORG
de	PRP	de	I-PERS
Strasbourg	NAM	Strasbourg	I-ORG
sous	PRP	sous	O
la	DET:ART	le	O
direction	NOM	direction	O
de	PRP	de	O
Jean-Marie	NAM	Jean-Marie	B-PERS
Lehn	NAM	Lehn	I-PERS
.	SENT	.	O

Patron de features :
forme fléchie, étiquette morpho-syntaxique,
lemme, et étiquette EN du mot courant

Approches statistiques pour l'Extraction d'Entités Nommées

Exemple d'utilisation de patrons de features

Né	VER:pper	naître	O
à	PRP	à	O
Paris	NAM	Paris	B-LOC
le	DET:ART	le	O
21	NUM	@card@	B-DATE
octobre	NOM	octobre	I-DATE
1944	NUM	@card@	I-DATE
,	PUN	,	O
Jean-Pierre	NAM	Jean-Pierre	B-PERS
Sauvage	NAM	Sauvage	I-PERS
a	VER:pres	avoir	O
effectué	VER:pper	effectuer	O
sa	DET:POS	son	O
thèse	NOM	thèse	O
à	PRP	à	O
l'	DET:ART	le	O
Université	NOM	université	B-ORG
de	PRP	de	I-PERS
Strasbourg	NAM	Strasbourg	I-ORG
sous	PRP	sous	O
la	DET:ART	le	O
direction	NOM	direction	O
de	PRP	de	O
Jean-Marie	NAM	Jean-Marie	B-PERS
Lehn	NAM	Lehn	I-PERS
.	SENT	.	O

Patron de features :
 étiquette morpho-syntaxique, lemme et
 étiquette EN du mot courant
 étiquette morpho-syntaxique et étiquette EN
 du mot précédent

Approches statistiques pour l'Extraction d'Entités Nommées

Modèle d'apprentissage avec des CRF

Processus:

- Données annotées utilisés comme exemple $((w, y))$
- Définition des *features* ou des patrons de *features* ($f_k(\dots)$)
- Apprentissage des poids du CRF permettant d'obtenir un modèle (λ_k)
- Application du modèle sur de nouvelles données en cherchant la séquence d'annotations y qui maximise $p(t|w)$

Bilan:

- CRF : meilleurs résultats pour les tâches correspondant à des annotations sur des séquences
- Autres possibilités :
sans étiquetage séquentiel : arbres de décision, SVM, etc.

Performances des différentes approches pour l'Extraction d'Entités Nommées

- Utilisation de règles
 - Règles lisibles, évolution des systèmes par ajout de lexique, mais coût de la description
 - Surtout adaptés à la langue écrite
 - Rappel & précision $> 90\%$
- Apprentissage de modèles
 - Modèles numérique, arbre de décision,... difficilement modifiables, mais coût de la description faible (nécessite un corpus d'apprentissage)
 - Surtout adaptés à la langue orale, mais aussi bonne performances sur les textes de spécialité
 - Rappel entre 50 et 90%
- Systèmes mixtes : avantages et inconvénients des deux

Mais performances variables suivant les entités nommées et le nombre de catégories

Analyse sémantique

Le niveau sémantique

- ▶ **But** : résoudre les problèmes de *référence* ; obtenir une *représentation conceptuelle* de l'énoncé dans un langage formel (formules de la logique du premier ordre, graphes conceptuels) ; *articuler* cette représentation conceptuelle avec le monde « physique » de la scène ;
- ▶ **Moyen** : calcul sémantique couplé à l'analyse syntaxique ou traduction ex-post de la représentation arborée dans un langage formel
- ▶ **Outils** : une description sémantique au niveau lexical (relations de synonymie, méronymie, hyper/hyponymie, etc), des règles de composition, des outils de représentation du monde physique ;
- ▶ **Difficultés** : explicitation partielle de l'implicite (problèmes de co-référence) ; ambiguïtés sémantiques (portée des quantifieurs) ; taille et précision de la connaissance nécessaire ; choix du formalisme de représentation (temporalité, croyances, etc).
- ▶ **Résultat** : un ensemble de représentations formelles de la scène dans lesquelles les objets et les relations qu'ils entretiennent sont identifiés ;

Le traitement sémantique : résultat

L'arbre syntaxique permet directement d'extraire les propositions (1) à (5), dont on peut déduire, compte-tenu d'une représentation du sens commun, (6), (7), (8) et (9) :

- ▶ $\exists X, \text{president}(X)$: il existe une entité X qui est président (et dont le référent est déjà connu) ;
- ▶ $\exists Y, \text{pomme}(Y)$: il existe une entité Y qui est une pomme ;
- ▶ $\exists Z, \text{couteau}(Z)$: il existe une entité Z qui est un couteau ;
- ▶ $\text{manger}(X, Y)$: cette entité X mange Y ;
- ▶ $\text{moyen}(\text{manger}(X, Y), Z)$: l'opération de manger s'effectue au moyen de Z ;
- ▶ $\text{president}(X) \Rightarrow \text{humain}(X) \Rightarrow \dots$;
- ▶ $(\text{pomme}(Y) \wedge \text{manger}(X, Y)) \Rightarrow \text{aliment}(Y)$
- ▶ $(\text{pomme}(Y) \wedge \text{manger}(X, Y)) \Rightarrow (\text{golden}(X) | \text{granny}(X) | \dots)$;
- ▶ $\text{manger}(X, Y) \Rightarrow \text{manger}(X), \text{est_ingere}(Y)$;

Chez l'humain, ces déduction se font de manière inconsciente et quasi-réflexe.

Analyse pragmatique

Le niveau pragmatique

- ▶ **But** : achever la *désambiguïsation* de l'énoncé en prenant en compte la dynamique de l'interaction (ou narration) en y intégrant ce qui est implicite ; comprendre la *fonction argumentative* de l'énoncé dans le contexte plus général de l'interaction (ou de la narration) : quelle information nouvelle apporte-t-il, au sujet de quoi dit-il quelque chose, sous quel mode...
- ▶ **Moyen** : une théorie des activités humaines ; une théorie des interactions langagières (la pertinence, les conditions de félicité) ; une théorie des structures discursives...
- ▶ **Outils** : représentation des actions humaines (scripts), « grammaire » des interactions, logique
- ▶ **Difficultés** : taille de la connaissance à représenter, spécification de la « grammaire » des interactions
- ▶ **Résultat** : une représentation formelle contextualisée de l'énoncé, une connaissance de sa fonction argumentative, des connaissances nouvelles...

Evaluation des Tâches de TAL

Métriques:

- Precision
- Recall
- F-Measure

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

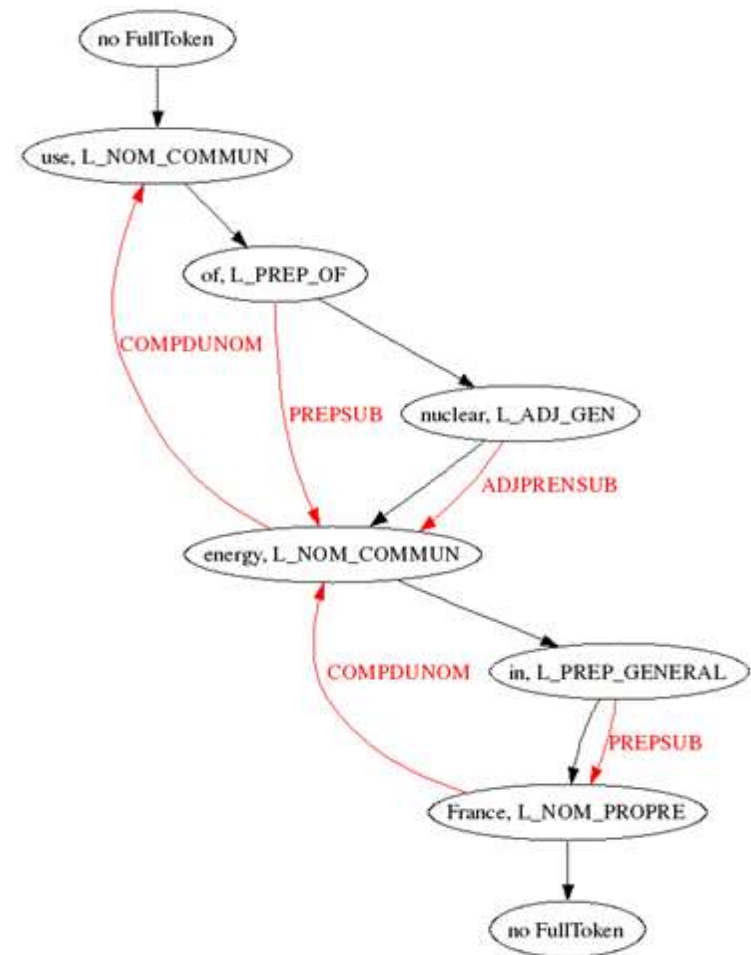
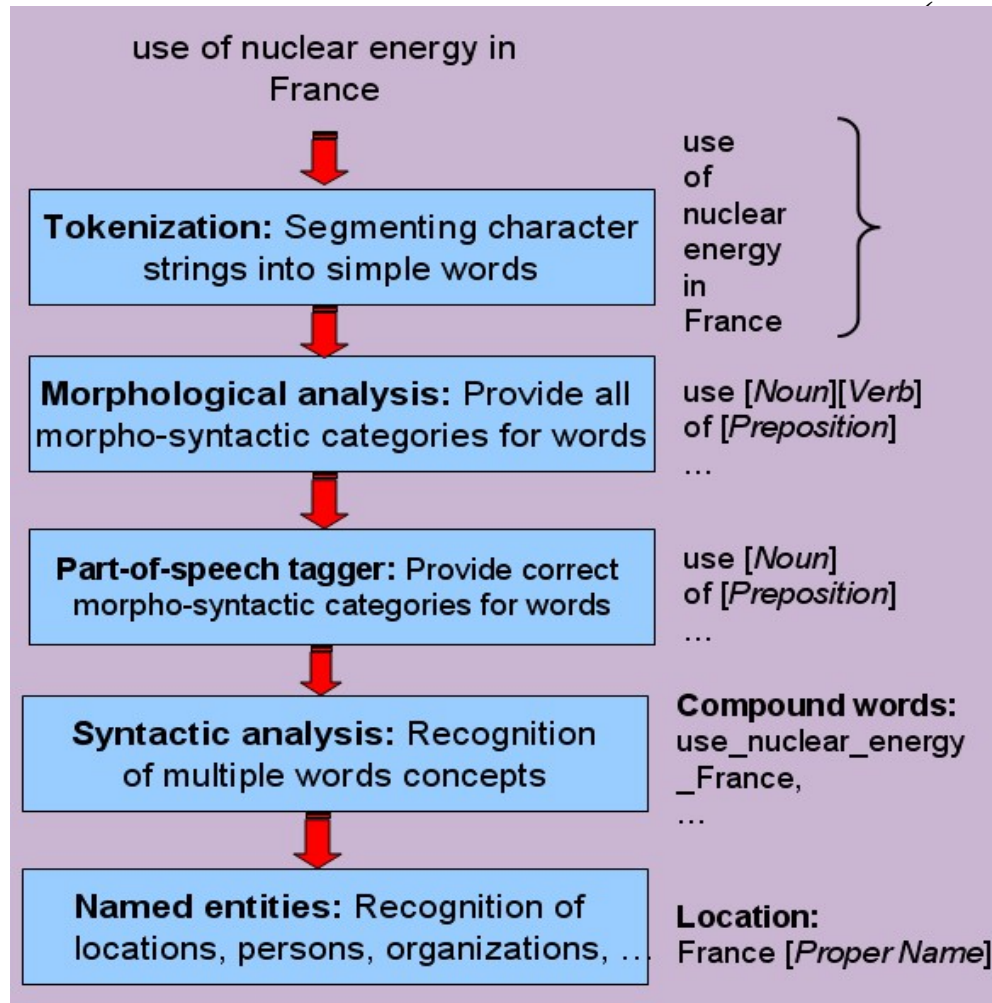
$$F = 2 * \frac{Precision \times Recall}{Precision + Recall}$$

- TP correspond au nombre de tokens (simples ou composés) correctement identifiés
- FP correspond au nombre de tokens (simples ou composés) incorrectement identifiés
- FN correspond au nombre de tokens (simples ou composés) non identifiés

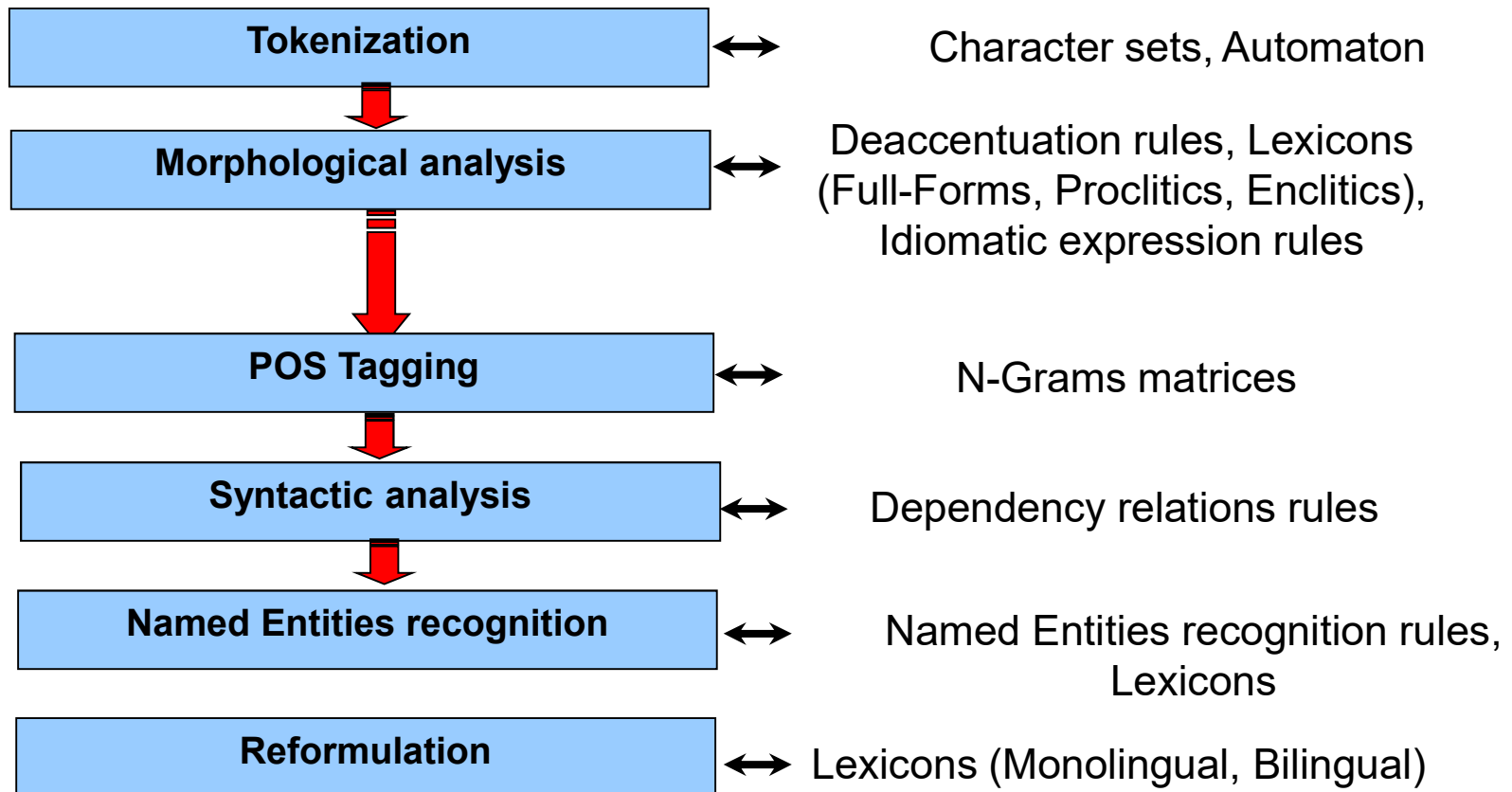
**Exemple d'un pipeline d'analyse
linguistique:**

LIMA: CEA LIST Multilingual Analyzer

Linguistic Processing



Linguistic resources



Demo

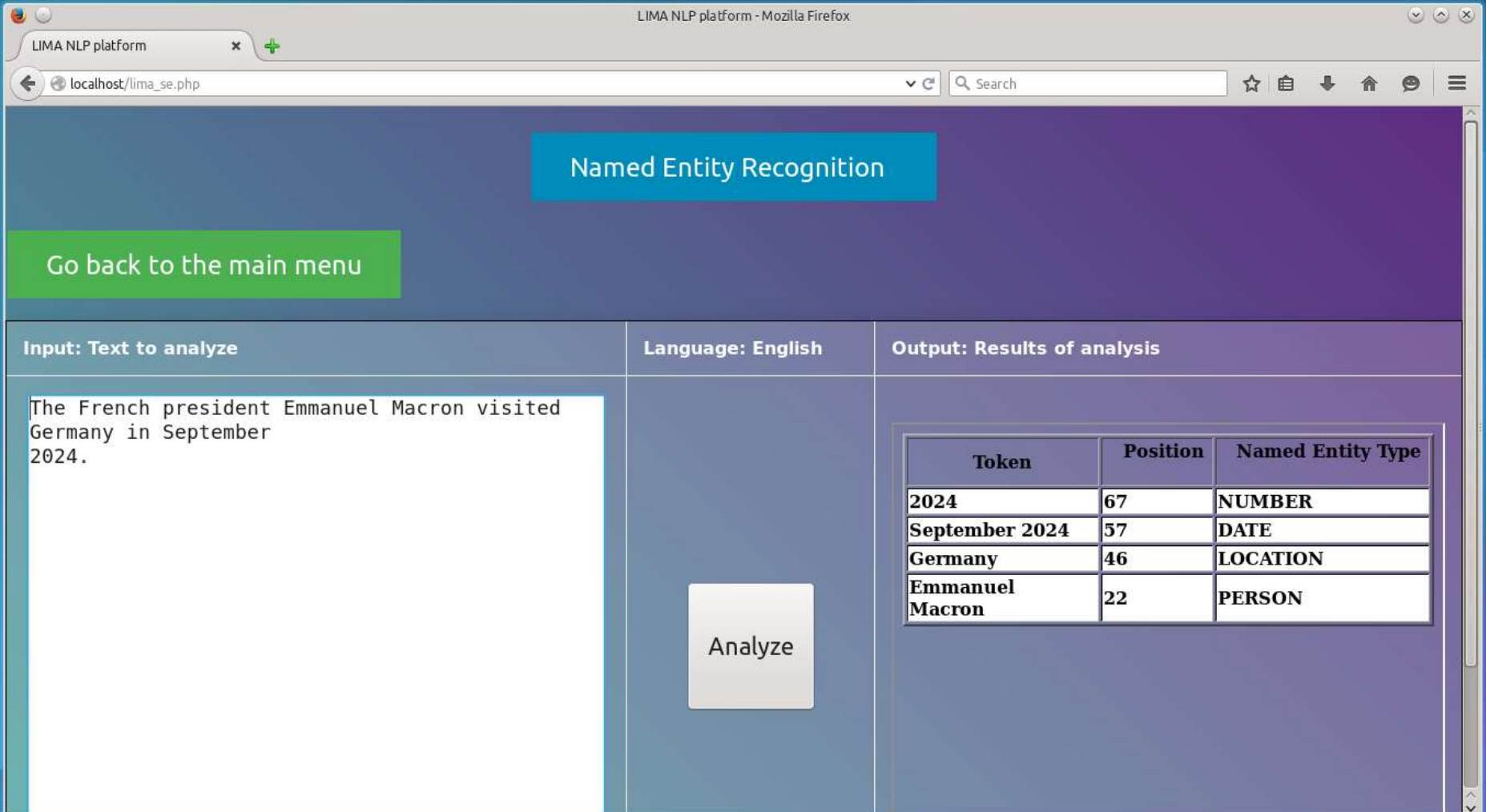
- CEA LIST Multilingual Analyzer (LIMA) – English
 - Part-Of-Speech (POS) Tagger
 - LIMA with python API: <https://github.com/aymara/lima/wiki#tldr>

The screenshot shows a web browser window titled "LIMA FAST - Mozilla Firefox" with the address bar showing "localhost/lima_sa.php". The page has a dark blue header with a "Part Of Speech Tagging" button and a "Go back to the main menu" button. Below the header, there are three main sections: "Input: Text to analyze", "Language: eng", and "Output: Results of analysis". The "Input" section contains a text box with the sentence "The French president Emmanuel Macron visited Germany in September 2024." and an "Analyze" button. The "Output" section displays a table of results.

Position	Token	Lemma	POS Tag
1	The	the	DT
5	French	French	JJ
12	president	president	NN
22	Emmanuel Macron	Emmanuel	NNP
38	visited	visit	VBD
46	Germany	Germany	NNP
54	in	in	IN
57	September 2024	September 2024	NNP
71	.	.	SENT

Demo

- CEA LIST Multilingual Analyzer (LIMA) – English
 - Named Entity (NE) Recognizer
 - NER pipelines: <https://github.com/aymara/lima/wiki/NER-pipelines>



The screenshot shows a web browser window titled "LIMA NLP platform - Mozilla Firefox" with the address bar showing "localhost/lima_se.php". The page has a dark blue header with a "Named Entity Recognition" button. Below the header is a green button labeled "Go back to the main menu". The main content area is divided into three columns: "Input: Text to analyze", "Language: English", and "Output: Results of analysis". The input column contains the text "The French president Emmanuel Macron visited Germany in September 2024." and an "Analyze" button. The output column displays a table of Named Entity Recognition results.

Token	Position	Named Entity Type
2024	67	NUMBER
September 2024	57	DATE
Germany	46	LOCATION
Emmanuel Macron	22	PERSON

Outils de TAL

Quelques liens pour des outils de TAL

- **Stanford CoreNLP**
<https://stanfordnlp.github.io/CoreNLP/>
- **Apache OpenNLP**
<https://opennlp.apache.org/>
- **FlairNLP**
<https://github.com/flairNLP/flair>
- **Stanford Stanza**
<https://stanfordnlp.github.io/stanza/>
- **Spark NLP**
<https://sparknlp.org/>
- **NLTK**
<https://www.nltk.org/>
- **Spacy**
<https://spacy.io/>