Introduction to Natural Language Processing

Anne-Laure Ligozat

2019/2020[1]

[1] librement inspiré des cours de Xavier Tannier, Aurélien Max et Dan Jurafsky, que je remercie

What is NLP?

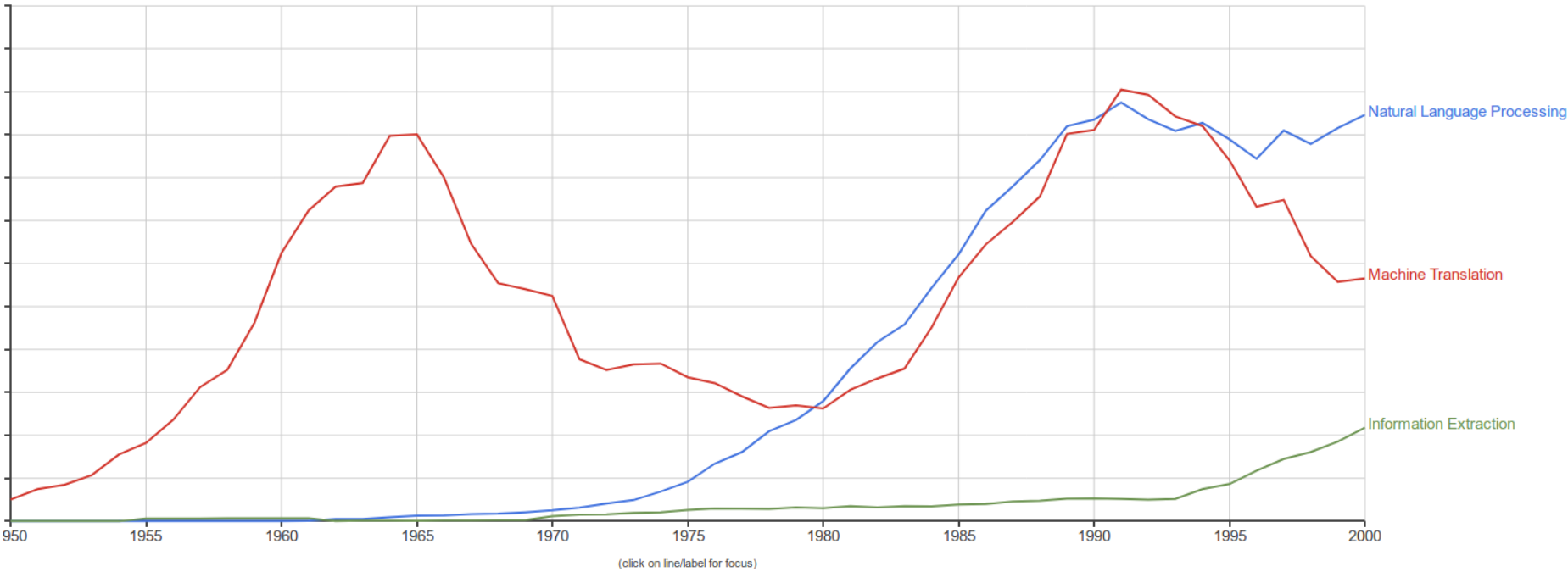
Natural Language Processing ou NLP (Traitement automatique des langues ou TAL)

- computer science and artificial intelligence discipline, that studies interactions between computers and languages
- discipline between linguistics and computer science
- born around the same time as computer science (50s)
- initial ambition: machine translation

Difficulties of machine translation

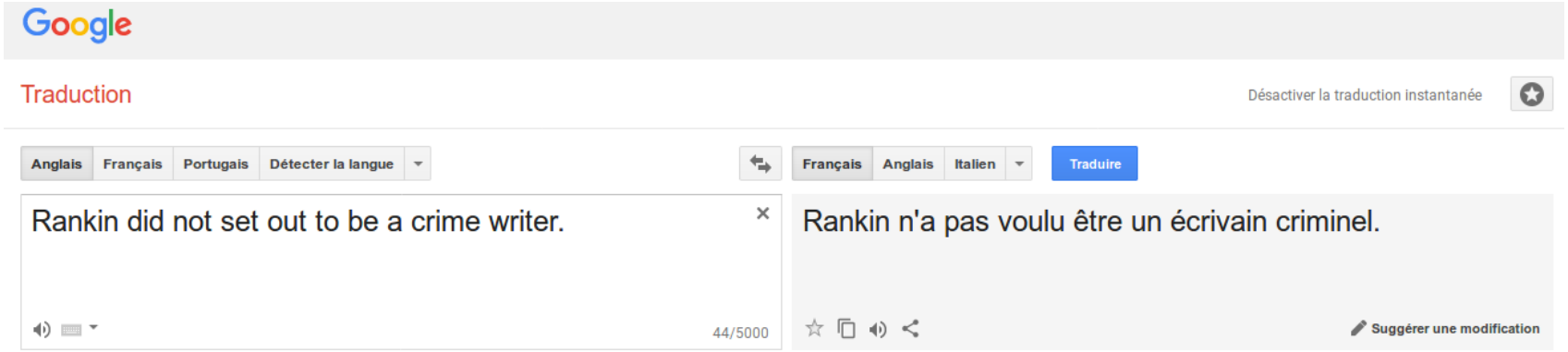
- The spirit is willing but the flesh is weak.
- The vodka is strong but the meat is rotten.

History


Natural Language Processing
Machine Translation
Information Extraction
950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
(click on line/label for focus)

What is NLP used for?

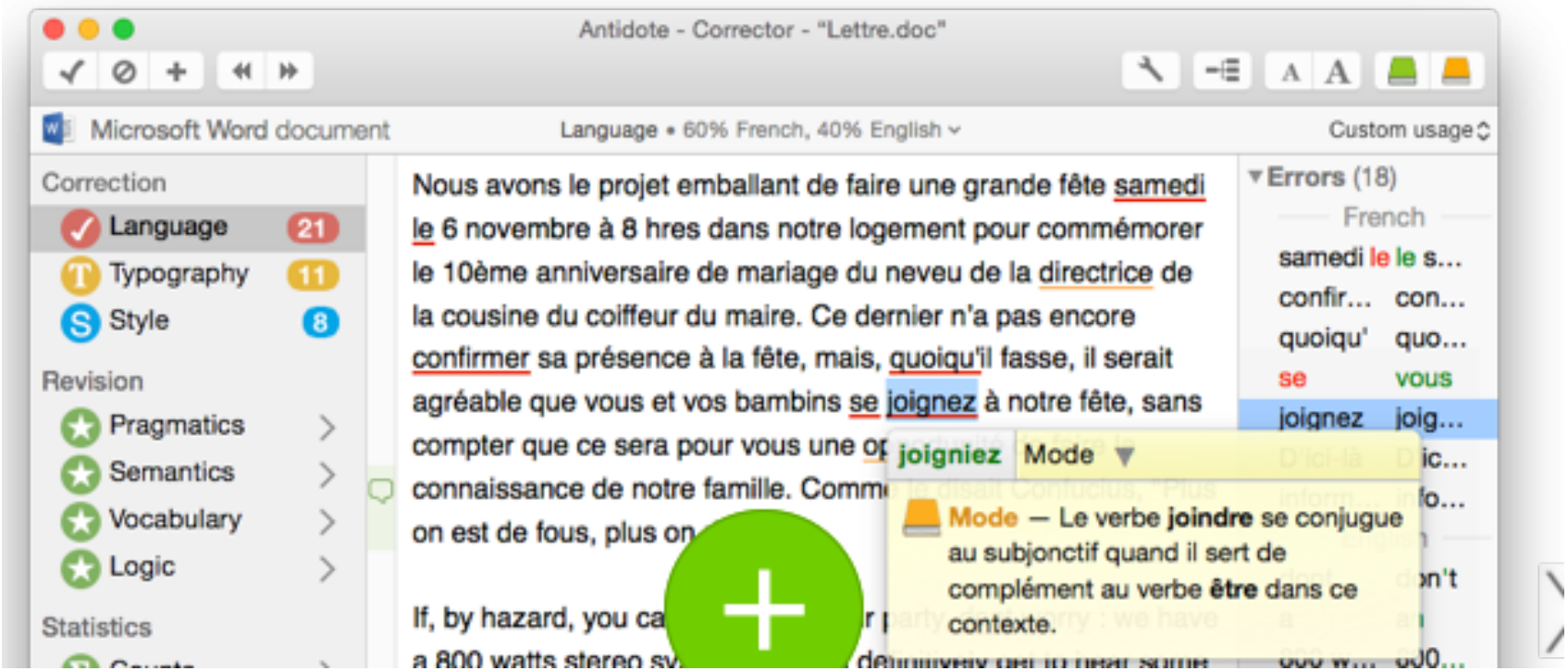
- machine translation
- spelling correction
- information extraction
- text simplification
- conversational agents
- ...



Google translate

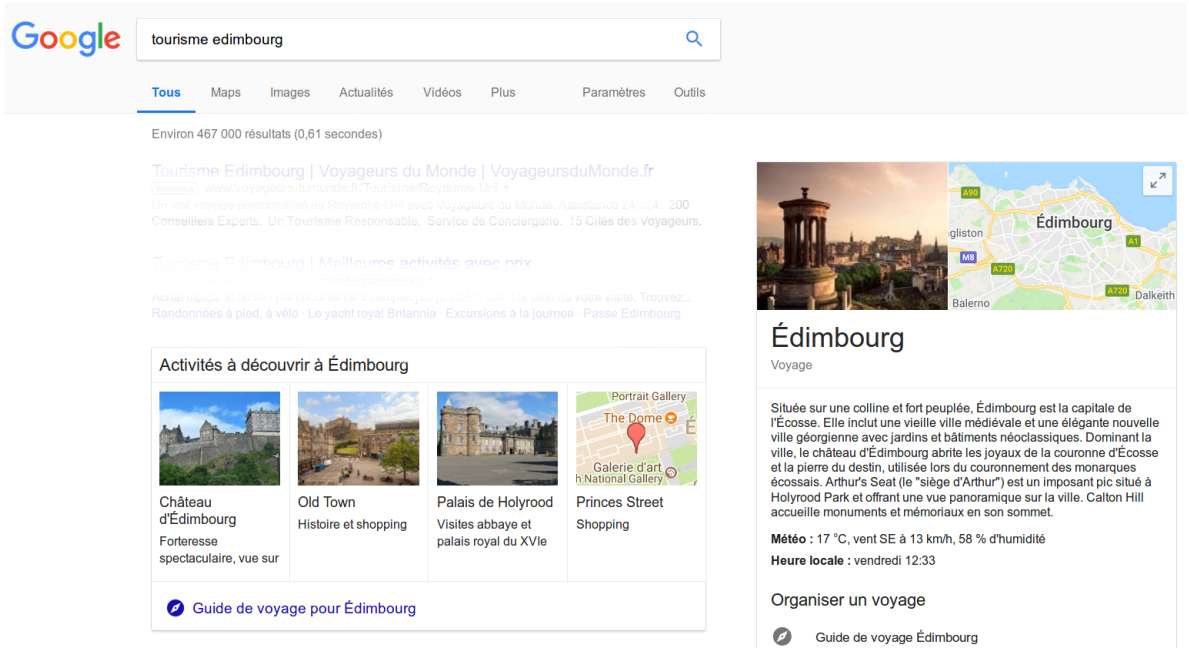
What is NLP used for?

- machine translation
- spelling correction
- information extraction
- text simplification
- conversational agents
- ...



What is NLP used for?

- machine translation
- spelling correction
- information extraction
- text simplification
- conversational agents
- ...



Google

What is NLP used for?

- machine translation
- spelling correction
- information extraction
- text simplification
- conversational agents
- ...

Having Nyong'o, with her darker skin and natural short crop, on the cover of Porter magazine's "Desire Issue" or putting trans actress and activist Laverne Cox on Variety's cover would have been unheard of years ago.

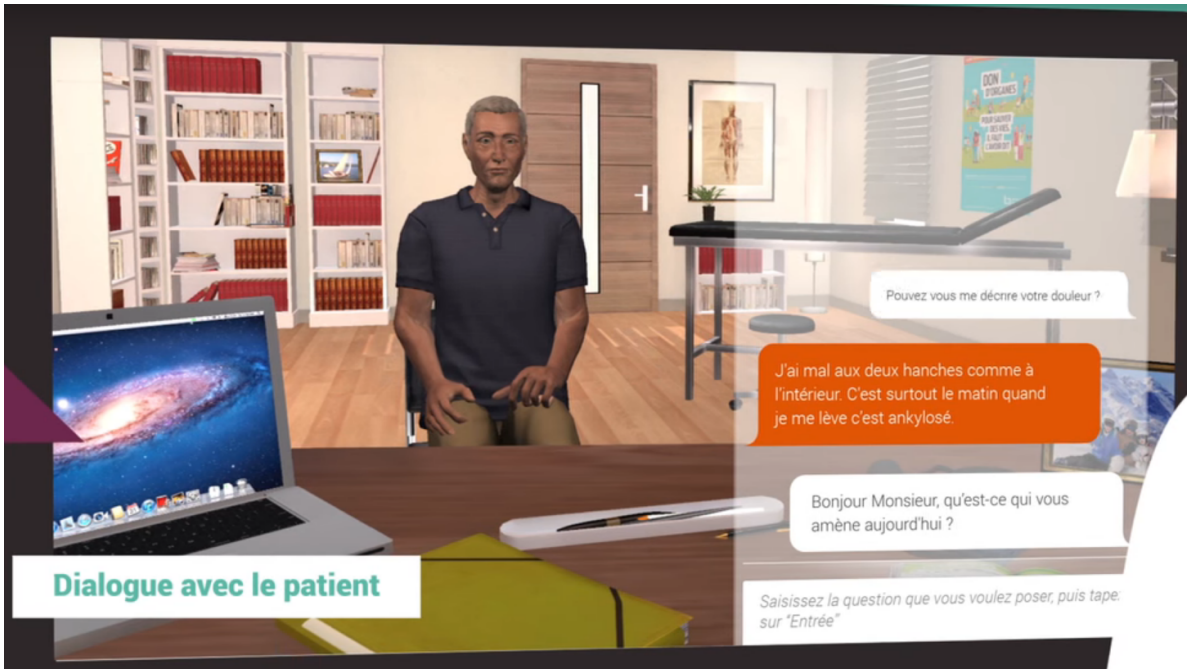
→

Nyong'o has dark skin and short hair on one magazine's cover. Laverne Cox is a transgender actress who appears on another cover. These women would not have been on magazine covers years ago.

Newsela

What is NLP used for?

- machine translation
- spelling correction
- information extraction
- text simplification
- conversational agents
- ...



PatientGenesys

NLP actors

- big editors
  - Facebook, IBM, Microsoft, Xerox, Apple, Toshiba, Sony, Google, Yahoo, Orange, etc.
- integrators/users
  - Ford, Symantec, EADS, Thalès/Arisem, BBN, SRI, EC, etc
- French SMEs
  - Exalead, Temis, ACapella, Lingway, Sinequa, Synapse, Systran, Reverso/Softissimo, Vecsys, Pertimm, Mondeca, etc.
- research labs
  - John Hopkins, Stanford, Berkeley, MIT, U. Maryland, Columbia, NYU, Cambridge, Edimbourg, AixlaChapelle, Stuttgart, Paris Diderot/INRIA, Paris Sud/LIMSI etc.

Difficulties in natural language

Natural language

- ambiguous
  - *Hospitals are sued by 7 foot doctors.*
  - *Teacher strikes idle kids.*
- implicit
  - *London is a famous writer.*
- redundant

In practice...

- non standard language
  - *Ouuuui c'est fini les horaires d'été #RERB*
  - *s est pas de votre faute après s est les siège pourris que j aime pas*
- expressions
  - *avoir les pieds sur terre*
  - *retourner sa veste*
- neologisms
  - *uberisation*
  - *songwriteuse*
- entity names
  - *il crée sa plus célèbre chorégraphie, Le Sacre du printemps*
  - *quatre années après le premier Rebus*
  - *elle sort le single Formation*

...

1 Introduction

2 A few linguistic notions

3 NLP processing

Language levels

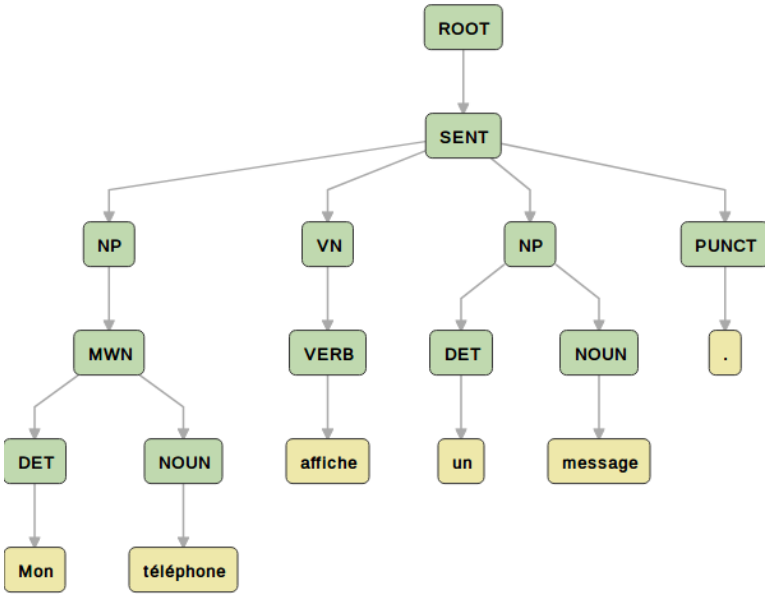
- *phonetics et phonology, study of sounds*
- morphology, study of words
- syntaxy, study of how words are arranged
- semantics, study of sense
- *pragmatics, study of the use of language in context*

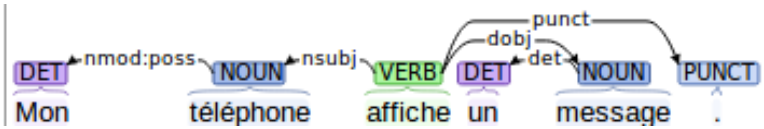
Morphology

- word = word form chantais, chantons... or lexical unit chanter
- notions of radical and affixes (prefixes et suffixes) a-ton-al
- major processes:
  - derivation: creation of new lexical units by adding affixes banal → banaliser
  - flexion: variation of a lexical unit according to grammatical criteria oeil → yeux
    - uninflected form of a word = lemma
  - composition: creation of new lexical units from existing units portefeuille
- (morpho-syntax) parts-of-speech La/DET mer/NOUN est/VERB bleue/ADJ ./PUNCT
  - possible addition of morphological information La/DET-fs mer/NOUN-fs est/VERB-3ppi bleue/ADJ-fs ./PUNCT

Syntax

- base unit = sentence (phrase) = list of words
- parse trees: constituents or dependencies
  - nature = nominal phrase, verb phrase...
  - fonction = subjet, object, attribute...
- notion of grammar

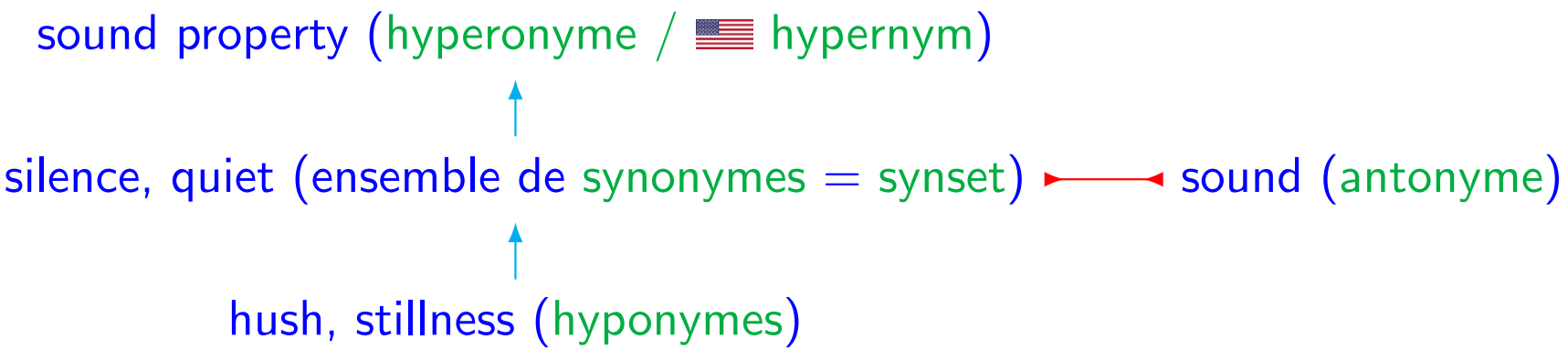




Semantics

Lexical semantics

- lexical relations: synonymy, antonymy, hyponymy
- homonymy, polysemy



relations WordNet

Grammatical semantics

- denotation (object of the world) vs sense la licorne d'en face
- construction of the sentence meaning

1 Introduction

2 A few linguistic notions

3 NLP processing

NLP processing - word level

Tokenization

Tokenization (segmentation en mots)

- very complex for some languages
- word (and sentence) separators are ambiguous
  - *etc., T.A.L, 21,3, aujourd'hui, l'illusion, jusqu'à, Jean-Louis, donne-t-il, États-Unis, France Inter...*

NLP processing - word level

Lemmatization/Stemming

Lemmatization

- find canonical form of a word or lemma (lemme)
  - chevaux → cheval
  - penseras → penser

Stemming (racinisation)

- find stem (racine) of a word
  - chevaux, cheval → cheva-
  - penseras, pensais → pens-

NLP processing - word level

POS tagging

POS tagging (étiquetage morpho-syntaxique/en parties du discours)

- attribute a grammatical category to words

Je/PRON suis/VERB content/ADJ que/SCONJ ce/PRON soit/VERB clair/ADJ entre/ADV nous/PRON ./PUNCT

Universal POS tags

NLP processing - word level

Disambiguation

Word Sense Disambiguation or WSD (désambiguïsation du sens)

- link a word occurrence to a given (or not) sense

fenêtre = menuiserie ou interface graphique ?

NLP processing - word level

Disambiguation

Word Sense Disambiguation or WSD (désambiguïsation du sens)

- link a word occurrence to a given (or not) sense

fenêtre = menuiserie ou interface graphique ?

- S: (n) **window** (a framework of wood or metal that contains a glass windowpane and is built into a wall or roof to admit light or air)
- S: (n) **window** (a transparent opening in a vehicle that allow vision out of the sides or back; usually is capable of being opened)
- S: (n) **window** (a transparent panel (as of an envelope) inserted in an otherwise opaque material)
- S: (n) **window** (an opening that resembles a window in appearance or function) *"he could see them through a window in the trees"*
- S: (n) **window** (the time period that is considered best for starting or finishing something) *"the expanded window will give us time to catch the thieves"; "they had a window of less than an hour when an attack would have succeeded"*
- S: (n) windowpane, **window** (a pane of glass in a window) *"the ball shattered the window"*
- S: (n) **window** (an opening in a wall or screen that admits light and air and through which customers can be served) *"he stuck his head in the window"*
- S: (n) **window** ((computer science) a rectangular part of a computer screen that contains a display different from the rest of the screen)

NLP processing - word level

Word embeddings

Word embeddings (plongements lexicaux)

- represent word forms in a continuous space of (relatively) low dimension and (hopefully) semantically relevant

(Ghannay et al., 2015)

NLP processing - sentence level

Segmentation

Sentence segmentation (segmentation en phrases)

- ambiguous sentence separators
- problem of sentence definition
  - bulleted lists
  - quotations
  - ...

NLP processing - sentence level

Parsing (analyse syntaxique)

Chunking

- identify group frontiers

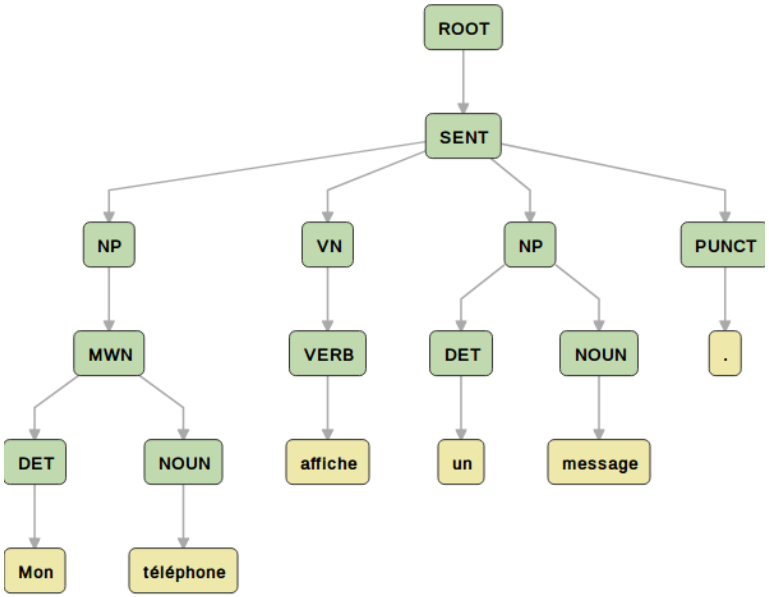
[Il] [le] [supplie] [de garder] [son nez] [anormalement grand] [en dehors] [des affaires] [des autres].

NLP processing - sentence level

Parsing (analyse syntaxique)

Constituents

- identify groups and relations between them



Grammar	Lexicon
$S \rightarrow NP\ VP$	$Det \rightarrow that \mid this \mid the \mid a$
$S \rightarrow Aux\ NP\ VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$Proper\text{-}Noun \rightarrow Houston \mid NWA$
$NP \rightarrow Det\ Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal\ Noun$	
$Nominal \rightarrow Nominal\ PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb\ NP$	
$VP \rightarrow Verb\ NP\ PP$	
$VP \rightarrow Verb\ PP$	
$VP \rightarrow VP\ PP$	
$PP \rightarrow Preposition\ NP$	

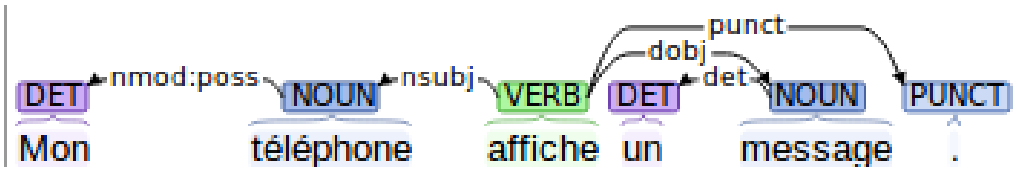
Speech and Language Processing. Daniel Jurafsky et James H. Martin

NLP processing - sentence level

Parsing (analyse syntaxique)

Dependencies

- identify relations between words



NLP processing - sub-sentence level

Named entity recognition

Named entity recognition (🇫🇷 reconnaissance d'entités nommées)

- typically: persons, locations (countries, cities, regions...), organizations (firms, universities, NGOs...)

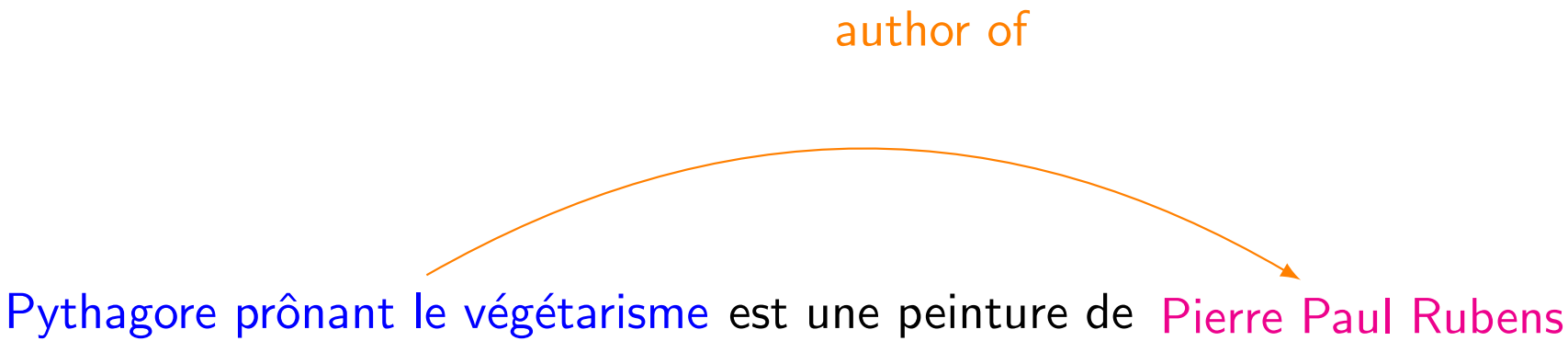
Le footballeur ivoirien Yaya Touré$_{PERSON}$ a déclaré que les joueurs noirs pourraient boycotter la Coupe du monde en 2018$_{DATE}$ si la Russie$_{LOCATION}$ n'aborde pas le problème du racisme dans les stades de football.

NLP processing - sub-sentence level

Relation extraction

Relation extraction

- relations from evaluation campaigns or knowledge bases



NLP processing - document level

Coreference resolution

Coreference resolution

- link each entity mention to the entity it refers to

Sophie Marceau est une actrice et réalisatrice française. Elle a été révélée à l'âge de 14 ans par le film La Boum, qui lui a permis de devenir d'emblée une vedette du cinéma français. Avec La Boum 2, elle obtient le César du meilleur espoir féminin. Le travail de la comédienne avec Mel Gibson sur l'épopée guerrière Braveheart en 1995 lui ouvre en grand les portes de Hollywood et marque le début d'une carrière internationale.

NLP today

- Robust approaches
  - large amount of textual documents
  - statistical approaches
- Evaluation
- Many tools and resources available
  - many languages but english largely prevalent

Language technologies

Dan Jurafsky, Introduction to NLP