

Représentation des langues

Représentation des langues et des connaissances

Damien Nouvel

Inalco

L'écriture comme dispositif **artificiel**

- **matérialisation** de la langue (orale)
- correspondance avec la langue parlée
- identification d'unités d'analyse

Double articulation (Martinet, 1961)

- **phonèmes** : unités minimales de prononciation
- **morphèmes** ou **monèmes** : unités minimales de sens

Quelques éléments linguistiques

Apparition des langues

- langues **naturelles**
- langues **construites** (conlangs, idéolangues)

Typologie des langues

- groupes
- langues
- dialectes
- variantes

Les chaînes de traitement TAL

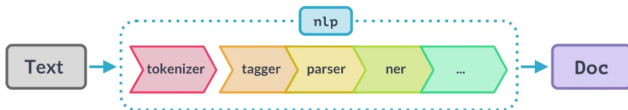
Représentation et analyse linguistique des textes

- entrée : **chaîne de caractères** (UTF8)
- sortie : analyses dans des **formats structurés**

Modules généralement implémentés

- tokenisation
- analyse morphologique
- catégorisation morpho-syntaxique
- analyse syntaxique en dépendance
- reconnaissance des entités nommées

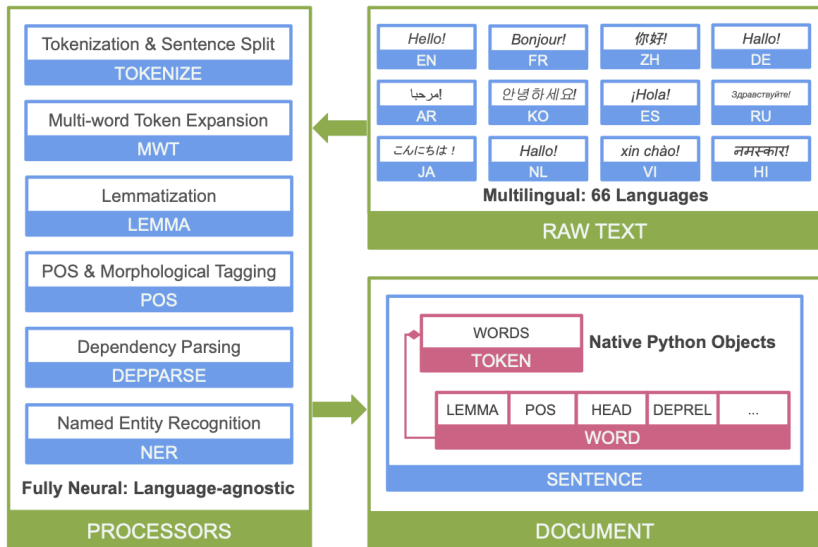
Chaîne de traitement de Spacy



NAME	COMPONENT	CREATES	DESCRIPTION
tokenizer	Tokenizer ☰	Doc	Segment text into tokens.
PROCESSING PIPELINE			
tagger	Tagger ☰	Token.tag	Assign part-of-speech tags.
parser	DependencyParser ☰	Token.head, Token.dep, Doc.sents, Doc.noun_chunks	Assign dependency labels.
ner	EntityRecognizer ☰	Doc.ents, Token.ent_iob, Token.ent_type	Detect and label named entities.
lemmatizer	Lemmatizer ☰	Token.lemma	Assign base forms.
textcat	TextCategorizer ☰	Doc.cats	Assign document labels.
custom	custom components	Doc._.xxx, Token._.xxx, Span._.xxx	Assign custom attributes, methods or properties.

Chaîne de traitement de Spacy

Chaîne de traitement de Stanza



Chaîne de traitement de Stanza

À la recherche des unités minimales

Nécessité d'identifier des unités minimales

- pour les caractères (lettres, symboles de l'alphabet, glyphes)
- **segmentation** des textes en tokens ou **tokenisation**

Construction de représentation par appel à des **ressources**

- détermination des unités minimales
 - morphèmes (cognats, racines, préfixes, suffixes)
 - tokens
 - formes (dictionnaires, lexiques, terminologies, thesaurus)
 - expressions polylexicales
- catégorisation morpho-syntaxique des unités minimales
- recherche de relations entre unités minimales
 - composition (locutions)
 - règles de syntaxe (grammaire, dépendances)

Encodage des symboles minimaux

Encodage informatiques des caractères avec **Unicode** (UTF8)

- normalisation **universelle** (généralisée depuis 2000)
- capacités
 - nombreuses langues
 - affichage contextuel
 - composition de caractères
 - évolutif

Segmentation en tokens

Découpage d'un texte en unités minimales : les **tokens**

- texte comme **séquence** de caractères (codes UTF8)
- recherche de **positions** pour découper le texte
- tokenisation sans perte : capacité à reconstituer le texte original

Exemple avec « Le chat boit du lait »

- la séquence comporte 20 caractères (espaces compris)
- on identifie les tokens $((0, 1), (3, 6), (8, 11), \dots)$
- le deuxième token $(3, 6)$ est (c_4, c_5, c_6, c_7) soit « chat »
- la chaîne est reconstruite avec les tokens (modulo les espaces)

Segmentation en tokens

La tokenisation détermine les autres traitements TAL

- les tokens seront les **unités minimales** de traitement
- les **représentations** sont construites à partir des tokens

Deux exigences contradictoires

- associer un **sens** à chaque segment
- **limiter** raisonnablement la taille du vocabulaire

Problèmes linguistiques de la tokenisation

- **morphologie** et **agglutination**
 - possibilité de calcul par dérivation
 - exemple : « anticapitalisme »
- **unités polylexicales**
 - impossible de construire le sens à partir de sous-segments
 - exemple : « cordon bleu »

Recherche de la taille **optimale** de segments

Segmentation par repérage de séparateurs

Recherche des séparateurs entre tokens

- focalisée sur la recherche des éléments entre les tokens
- modélisation à partir d'**expressions régulières**
- avec `grep` utilisation de `\b`

Avantages

- **simple** à implémenter
- très **rapide** (automates)
- relativement **robuste**

Inconvénients

- inopérante pour les **langues sans séparateurs** (chinois)
- pas d'**analyse** des tokens
- erreurs pour les tokens avec **punctuations** (sigles, etc.)

Segmentation par modélisation linguistique

Approches **descriptives** de la segmentation

- exploitation d'un **lexique** de la langue
- **paradigmes** de flexions et de conjugaisons
- utilisation d'**expressions régulières**
- avec `grep`, utilisation de `\w`

Avantages

- segmentation motivée **linguistiquement**
- **informations** sur les mots (catégorie, racine, famille, etc.)

Inconvénients

- mise en œuvre difficile pour la segmentation en **morphèmes**
- **ressources** à constituer pour chaque langue
- limitées au **vocabulaire** du lexique
- spécifier ou détecter la langue est nécessaire

Segmentation statistique en sous-mots

Algorithmes BPE (Byte Pair Encoding, Gage 1994) et WordPiece

- exploitation d'un **corpus volumineux**
- initialisation, chaque caractère est un token
- jusqu'à une **taille de vocabulaire** souhaitée
 - recherche dans le corpus des « **paires** » les plus
 - fréquentes (BPE)
 - probables (WordPiece)
 - **fusion** et **remplacement** des paires trouvées

Avantages

- capacité à traiter des **tokens hors-vocabulaire**
- modélisation (empirique) de la **morphologie**

Inconvénients

- pas de justification **linguistique**
- pas d'**informations linguistiques** explicites

Utilisés par les LLMs (BERT: WordPiece 30K, GPT4 : BPE 200K)

Comparaison des tokenizers

Outil	Tokens
NLTK	Charles de Gaulle n' était pas anticapitaliste
Spacy	Charles de Gaulle n' était pas anticapitaliste
Stanza	Charles de Gaulle n' était pas anticapitaliste
BERT	Charles de Gaulle n' était pas antica pita iste
GPT	Charles de Gaulle n' était pas ant icap ital iste

Quelques critiques

- traitements très différents des **ponctuations**
- aucune **expression polylexicale**
- pas d'analyse **morphologique** bien construite

Construction de représentations

Associer des informations à chaque token

- nature du token : **catégories** morfo-syntaxiques
- informations linguistiques : **traits** associés aux tokens

Relation entre les tokens

- décrits par les grammaires (descriptions)
- analyse **syntaxique**
- grammaires de **dépendances**

Exploitation par représentations mathématiques

- calcul de **plongements**
 - statiques
 - contextuels
- inférence **empirique** à partir de **corpus**

Ressources lexicales

Construction par description

- établir la liste des entrées (lemmes, formes)
- **paradigmes** linguistiques (déclinaison, conjugaison, etc.)
- automatisation de la liste de formes

Construction empirique

- collecte de **corpus représentatif**
- **tokenisation**
- **statistique** sur les tokens
- filtrages
 - par expressions régulières
 - par calculs statistiques (fréquence, corrélation)

Ressource **descriptive** et **prescriptive**

Caractéristiques principales

- liste de **mots** (lemmes) utilisés
- généralement **monolingues**, parfois bilingues
- tri par **ordre alphabétique** (lexicographique)
- **définition** pour chaque entrée de la liste
- informations linguistiques variées (nature, catégorie)
- peut contenir des noms propres
- construits essentiellement par des humains

Lexiques

Ressource plutôt **descriptive**

Caractéristiques principales

- liste de formes attestées dans une langue
- génération par paradigmes (conjugaison, flexion)
- présence d'informations linguistiques très variable
- contient souvent des noms propres
- génération automatique (paradigmes)
- objectif de **couverture**
- construits essentiellement automatiquement

Terminologies

Vocabulaire spécifique à un domaine

Caractéristiques principales

- recension de **termes** dans un domaine
 - par exemple : biologie, nucléaire, écologie, etc.
- définitions des termes
- généralement peu d'entrées pour la langue courante
- description de **relations** plus ou moins formelles entre termes
- curation par un humain
- peut permettre de générer le lexique (formes) d'un domaine
- objectif de **sémantique**
- majorité de noms (communs ou propres)

Thesaurus

Notions et termes mis en réseau par des **relations**

Caractéristiques principales

- s'appuie sur une terminologie
- organisation par **notions** (sémantique)
- structure hiérarchique (taxonomie)
- formalisation explicite des **relations** entre termes
 - synonymes, antonymes
 - hyperonymes, hyponymes

Catégories morpho-syntaxique

Détermination des **rôles** linguistiques des mots en **contexte**

- en anglais **Part Of Speech** (POS)
- catégories déterminées pour chaque mot par
 - la morphologie (et les lexiques)
 - le rôle syntaxique (analyse grammaticale)
- **ambiguïté** des catégories du lexique
 - plusieurs catégories
 - aucune catégorie pour les mots inconnus
- processus de désambiguïsation par la syntaxe

Jeux de catégories prédéfinies

- FrenchTreeBank
 - catégories : N, A, Adv, P, D, CL, PRO, PREF, C, I, V, ET, PONCT
 - sous-catégories et traits morphologiques (genre, nombre, etc.)
- classes Universal POS TAG (Universal Dependencies)
 - ouvertes : ADJ, ADV, INTJ, NOUN, PROPN, VERB
 - fermées : ADP, AUX, CCONJ, DET, NUM, PART, PRON, SCONJ
 - autres : PUNCT, SYM, X

Étiquetage morpho-syntaxique

Tâche d'**annotation** (étiquetage, catégorisation, etc.)

- réalisée par
 - un humain (plus ou moins expert)
 - un algorithme (automatique)
- en anglais : **POS Tagging**

Étiquetage par un **humain**

- sur des **corpus** sélectionnés (langue, domaine)
- selon un **guide** (catégorie, schéma, format)
- mesure du **taux d'accord**

Étiquetage par un **algorithme**

- par règles
- par méthodes statistiques

Algorithmes statistiques d'étiquetage

Ressources linguistiques (par langue)

- lexiques avec informations morpho-syntaxique
- corpus annotés

Vraisemblance des étiquettes pour un énoncé

Estimation de **probabilités**

- **classes majoritaires** (par mot) selon
 - mot courant
- probabilités de **dépendances** (Bayes) selon
 - mot courant et ses caractéristiques
 - mot précédent et catégorie précédente
- modèles **génératifs** (Markov) selon
 - génération du mot par catégorie
 - transitions entre catégories

Évaluation des **performances** (taux, précision, rappel, f-mesure)

Analyse syntaxique

Détermination des **fonctions** des éléments dans un énoncé

- lien avec la **nature** des éléments (catégories)
- **relations** entre les tokens
- **rôles** des tokens au sein de l'énoncé
- contraintes sur l'**ordre** d'apparition

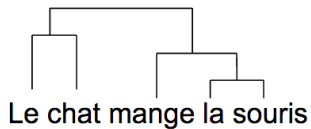
Modélisations

- **constituants** par arbres syntaxiques
- **dépendances** par graphes de relations

Grande **variabilité** entre les langues

- **typologie** des langues
- pas de jeu de relation universelle

Analyses par constituants ou dépendances



Analyses par constituants ou dépendances

Syntaxe par constituants

Description par un arbre

- **racine** comme nœud principal de l'énoncé (en général verbal)
- enfants par description des éléments
- contrainte forte d'**ordre** des constituants
- modélisation par **grammaire hors contexte**

Exemple de grammaire pour le français

$E \rightarrow GN \ GV \ GN$

$GV \rightarrow V$

$GN \rightarrow DET \ NC$

$GN \rightarrow NP$

Syntaxe par dépendances

Graphe de relations

- **arcs** entre les tokens de l'énoncé
- graphe **orienté sans cycle**
- contraintes moins forte sur l'ordre des tokens

Universal Dependency Relations

- core : nsubj, obj, iobj, csubj, ccomp, xcomp
- non-core, nominal
- coordination, headless, loose, special, other

Entités nommées