Langage naturel : les mots

Béatrice Daille - Université de Nantes, LINA

19 octobre 2012

⋆ **Le mot : définition**

⋆ **Racination : algorithmes de Lovins et Porter**

⋆ **Lemmatisation et analyse morphologique : automate à états finis et transducteur**

⋆ **Compter les mots : n-grammes**

⋆ **Reconnaissance bruitée : distance minimale d'édition (algorithmique du texte)**

Le mot

Linguistique

Morphologie : étude des mots et de leur construction

1. Classes de mots ;
2. Étude de ces classes (construction, variations, etc.).

Parties du discours

Distribution des mots dans différentes classes : parties du discours.
Chaque mot de la langue a une **catégorie morpho-syntaxique** ou catégorie grammaticale.

Neuf classes de mots

- les noms ou les substantifs
- les verbes
- les pronoms
- les déterminants (regroupent les articles, adjectifs possessifs, adjectifs démonstratifs, adjectifs interrogatifs, adjectifs exclamatifs, adjectifs numéraux cardinaux)
- les adjectifs qualificatifs et numeraux ordinaux
- les adverbes
- les conjonctions
- les prépositions
- les interjections

un mot simple est un ensemble de morphèmes
Morphème : unité significative minimale

Morphèmes

- **Morphème lexicaux / grammaticaux**

morphème lexical : *vent, chat*
morphème grammatical : *s* du pluriel
mot peut être composé de plusieurs morphèmes : *é|vent|é|s*

- **Morphème autonome / non autonome**
- **Différents morphèmes : racine/affixe**
  - ⋆ **affixe** : préfixe, suffixe, dédoublement

    le dédoublement marque le pluriel en indonésien

    orang (homme), *orang+orang* (hommes)
- **Flexion / Dérivation**
- **Morphotactique / Morphophonématique**

⋆ **Morphotactique** : l'ordre dans lequel les morphèmes peuvent apparaître au sein du mot.

bio, *dégrader*, *able* → *biodégradable*

[[*bio*/NOM] [[*dégrad(er)*/VBE] *able*/ADJ] /ADJ] /ADJ]

⋆ **Morphophonématique** : l'altération de la forme d'un morphème selon un contexte phonétique ou orthographique

misère, *able* → *misérable*

èCe → éC

Morphologie (2)

◦ **Paradigme flexionnel** : je *travaille*, tu *travailles*, elle/il *travaille* ... Le **lemme** est *travailler*.

◦ **Paradigme dérivationnel** : *nation*, *nationalité*, *nationaliser* ... La **racine** est *nation*.

◦ **Composition** : un *lave-vaisselle*, un *timbre poste*, un *centimètre*, *tout à fait*.

Description d'un mot

◦ **Racine et lemme** de *nationalisaient* : lemme *nationaliser* et racine *nation*.

◦ **Catégorie morphosyntaxique (ou grammaticale)** attachée au lemme : *nationaliser* est un **verbe**.

◦ **Traits morphologiques** distinguent les différentes flexions d'un paradigme flexionnel : *nationalisaient* est le verbe *nationaliser* à la **3ème personne** de l'**imparfait** de l'**indicatif**

Morphologie dérivationnelle

Affixations

◦ **Préfixation** : *construire* $\rightarrow$ *dé-construire*.

◦ **Suffixation** : *construire* $\rightarrow$ *construct-eur*

◦ **Allomorphies** :
– de l'affixe qui marque le pluriel pour les noms : *s*, *x*
– de la racine induite par le suffixe dérivationnel *-ion* :
permettre/permission,
confondre/confusion,
conduire/conduction.

◦ **Combinaison d'affixations sur une même racine**
Structure d'un mot construit :
déconstructeur = [[*dé* [*construire*]$_V$]$_V$ *eur*]$_N$.

Racination

Définition : associer une "racine" commune à un ensemble de variantes morphologiques

Algorithmes de racination : désuffixage et normalisation
- Lovins 1968
- Porter 1980
  anglais : *http ://www.tartarus.org/ martin/PorterStemmer/*
  français : *http ://snowball.tartarus.org/french/stemmer.html*

Lovins : Désuffixage et normalisation séparés

1. Terminaisons (recherche par taille décroissante)

11	-alistically -arizability -izationally	10	-antialness -arisations -arizations -entialness	9	-allically -antaneous -antiality -arisation, ...

2. Normalisation des terminaisons (recherche dans l'ordre)
   a suppression des doubles : bb-, dd-, gg, ll-, mm-, nn-, pp-, rr-, dd-, tt-, ...
   b iev- $\rightarrow$ ief-
   c uct- $\rightarrow$ uc-
   d umpt- $\rightarrow$ um-
   e rpt- $\rightarrow$ rb-
   f ...

Racineur de Porter

Désuffixage et normalisation simultanés

Algorithme :

Consonne (c) une lettre autre que A, E, I, O, U et autre que Y si Y est précédé d'une consonne.

Voyelle (v) une lettre qui n'est pas une consonne

C suite de consonnes (au moins 1)

V suite de voyelles (au moins 1)

mot CVCV...C, CVCV...V, VCVC...C, VCVC...CV $\rightarrow$ [C]VCVC...[V]

mesure (m) [C]VC$\{m\}$...[V]

règle (condition) S1 $\rightarrow$ S2

condition m>1, *S, *v*, *d, *o + combinaisons logiques (et, ou, non)

Étapes :

Step1a	-SSES $\rightarrow$ -SS -IES $\rightarrow$ -I -SS $\rightarrow$ -SS -S $\rightarrow$	careSSES $\rightarrow$careSS ponIES $\rightarrow$ ponI careSS $\rightarrow$ careSS catS $\rightarrow$ cat
Step1c	-Y $\rightarrow$ -I -ANT $\rightarrow$ -EMENT $\rightarrow$ -MENT $\rightarrow$	happY $\rightarrow$ happI irritANT $\rightarrow$ irrit replacEMENT $\rightarrow$ replac adjustMENT $\rightarrow$ adjust
Step2	(m>0) -ATIONAL $\rightarrow$ -ATE (m>0) -TIONAL $\rightarrow$ -TION	relATIONAL$\rightarrow$ relATE condiTIONAL $\rightarrow$ condiTION

Exemples de racinisation

Racines obtenues par Lovins

Chaîne initiale	Chaîne après désuffixage	Chaîne normalisée
magnesia	magnes	magnes
magnesite	magnes	magnes
magnesian	magnes	magnes
magnetize	magnet	magnet
magnetometer	magnetometer	magnetometer
magnetometric	magnetometr	magnetometer
magnetometry	magnetometr	magnetometer

Erreurs produites par Porter

Mauvais regroupement (faux positifs)	organization doing generalization policy university	organ doe generic police universe
Regroupement non effectué (faux négatifs)	European matrices noise sparse explain	Europe matrix noisy spasity explanation

Lemmatisation

Définition : associer un lemme à une forme fléchie

- **Lemme :** une forme choisie conventionnellement pour représenter un paradigme flexionnel

- **Paradigme flexionnel** : je *travaille*, tu *travailles*, elle/il *travaille* ... Le **lemme** est *travailler*.

Tache qui s'effectue aisèment dès que la catégorie grammaticale de la forme fléchie est connue

Étapes :

1. Reconnaissance de la forme fléchie
2. Calcul de la racine
3. Calcul des flexions (identification des affixes flexionnels)
4. Génération de la forme neutre

Reconnaissance de la forme fléchie

Automates finis

M= (Q, $\sum$, δ, $\mathbf{q}_1$, F)

Q un ensemble fini d'états q

$\sum$ un alphabet fini de lettres ou de morphèmes σ de L

δ**(Q,** $\sum$**)** un ensemble de règles de transition

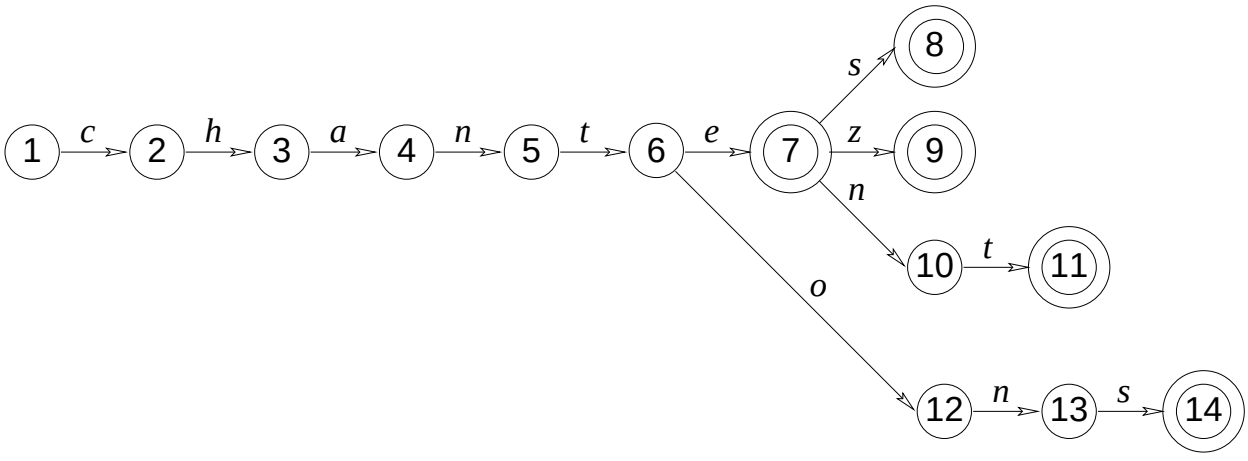
$\mathbf{q}_1$ état initial

F ensemble états finals

Une chaîne est acceptée ssi il existe un chemin allant de l'état initial à un état final étiqueté par cette chaîne.

Exemple

Reconnaissance des formes fléchies du verbe *chanter* au présent de l'indicatif



Transducteurs finis

Un transducteur fini est un automate dont les transitions portent des couples d'étiquettes : une étiquette d'entrée et une étiquette de sortie.

M= (Q, K, δ, q_1, F)

Q un ensemble fini d'états q

K un alphabet fini de symboles complexes : couples d'étiquettes entrée/sortie avec les étiquettes d'entrée $\in \sum$ et les étiquettes de sortie $\in$ O

δ(Q, $\sum$:O) un ensemble de règles de transition,

q_1 état initial

F ensemble états finals

Une chaîne est acceptée ssi il existe un chemin C allant de l'état initial à un état final étiqueté par cette chaîne. La chaîne émise est obtenue en concaténant les symboles émis sur les transitions du chemin C.

Trois opérations sur les transducteurs

Union

Si T1 et T2 sont deux transducteurs, il existe un transducteur T1 U T2 tel que l'image de toute chaîne par T1 U T2 soit l'union des images par T1 et T2.

Inversion

Si T est un transducteur, il existe un transducteur T-1 tel que l'image de toute chaîne C par T-1 est l'union des chaînes dont l'image par T est C.

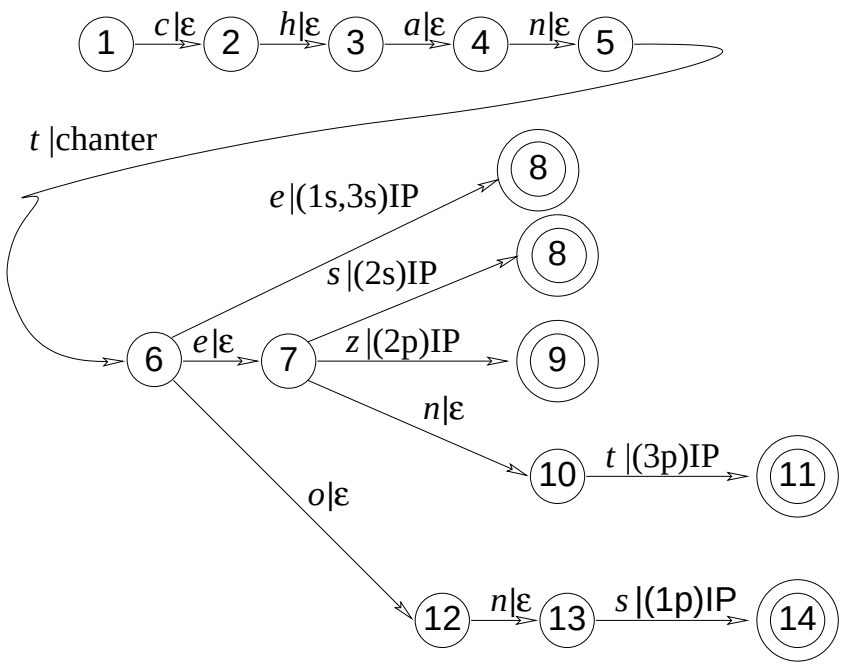
Composition

Si T1 et T2 sont deux transducteurs, il existe un transducteur T2 o T1 tel que l'image de toute chaîne C par T2 o T1 soit l'image par T2 de l'image de C par T1.

Inversion et composition sont les deux propriétés les plus importantes. Elles permettent :
(1) d'inverser un transducteur passant, par exemple, d'analyse en génération ;
(2) de composer autant de transducteurs élémentaires que l'on souhaite en une machine complexe.

Les transducteurs ne sont pas fermés pour l'intersection.

Exemple de transducteur effectuant une lemmatisation

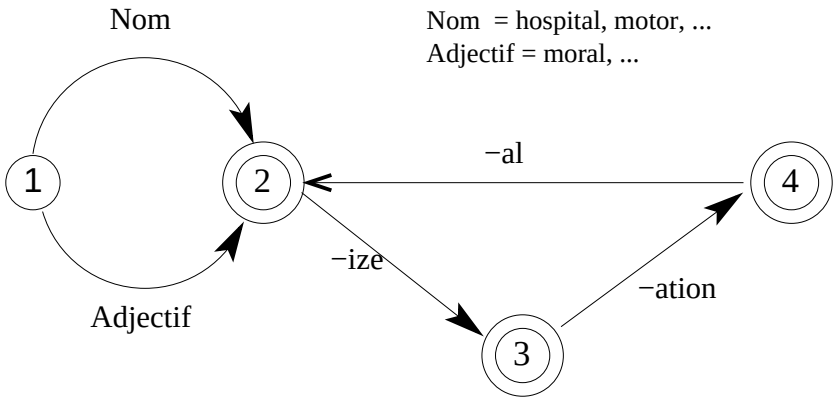


Analyse morphologique

Définition : analyser un mot en racine et affixes dérivationnels

- **Paradigme dérivationnel** : *nation*, *nationalité*, *nationaliser* ... La **racine** est *nation*.

Automates à états finis



Tables des transitions

	Entrée				
État	Nom	Adjectif	ize	ation	al
1	2	2	0	0	0
2 :	0	0	3	0	0
3 :	0	0	0	4	0
4 :	0	0	0	0	2

Analyse morphologique : transducteurs

Morphologie à deux niveaux

Représentation lexicale

	c	h	a	t	+N	+Pl		

Représentation de surface

	c	h	a	t	s		

$\sum = \{$ c :c, h :h, a :a, t :t, +N :ϵ, +Pl :ϵ, ϵ :s$\}$

Représentation lexicale

	l	i	o	n	+N	+F		

Représentation intermédiaire

	l	i	o	n	$\diamond$	e	#

Représentation de surface

	l	i	o	n	n	e	

Analyse morphologique : transducteurs

Doublement des consonnes : ϵ :n $\Leftarrow$ n$\diamond$__e#

Formalisme des règles : C **op** CG __ CD

C, CG, CD expressions régulières

op

1. Règle d'exclusion : transformation interdite dans le contexte
   a :b /$\Leftarrow$ CG __ CD
2. Règle de restriction contextuelle : tranformation s'applique uniquement dans le contexte (la transformation dans un autre contexte est interdite)
   a :b $\Rightarrow$ CG __ CD
3. Règle de contrainte surfacique : transformation s'applique toujours dans le contexte (une autre transformation est interdite dans le contexte)
   a :b $\Leftarrow$ CG __ CD
4. Règle de composition : transformation qui s'applique uniquement et toujours dans le contexte
   a :b $\Leftrightarrow$ CG __ CD