Morphologie

Béatrice Daille - Université de Nantes, LS2N

16 novembre 2020

⋆ **Le mot : définition et propriétés**

⋆ **Racination : règles désuffixation-recodage**

Le mot

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.

Douze classes de mots
Jeu de catégories grammaticales validées sur 22 langues
Tables de correspondance pour 27 jeux d'étiquettes

VERB verbes (tous les temps et modes)
NOUN noms (commum et propre)
PRON pronoms
ADJ adjectifs et numeraux ordinaux
ADV adverbes
ADP adpositions (prépositions and postpositions)
CONJ conjonctions
DET déterminants (regroupent les articles et les adjectifs possessifs, démonstratifs, interrogatifs, exclamatifs ou numéraux cardinaux)
NUM chiffres
PRT particules ou autres mots fonctionnels
X autres : mots étrangers, typos, abréviations, interjections
. ponctuation

`http://code.google.com/p/universal-pos-tags/`

"A Universal Part-of-Speech Tagset" by Slav Petrov, Dipanjan Das and Ryan McDonald LREC 2012

Morphèmes

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

- **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** du **pluriel** 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*]$_N$

Dérivations régulières

humain	humanité	humanitaire	humanitarisme
collecte	collecter	collection	collectionner
industrie	industriel	industrialiser	industrialisation

Ambiguïtés

◦ **homographes/homophones** : *fils/foie*, *fois*, *foi*

◦ **homonymes** : *produits*, *vase*, *avocat*

◦ **Dérivations et homonymie**

Deux verbes homonymes avec chacun leur dérivé nominal :

griller $\rightarrow$ grillage
griller $\rightarrow$ grillade

◦ **Dérivations et polysémie**

Deux verbes polysémiques dont chaque sens a ses dérivés nominaux :

raffiner $\rightarrow$ raffinage
$\rightarrow$ raffinement

◦ **Dérivations multiples**

Un verbe dont on dérive plusieurs noms :

coller $\rightarrow$ collage
$\rightarrow$ colleur

Composition

Définition : mécanisme de formation des mots qui consiste à combiner deux ou plusieurs éléments lexicaux autonomes pour former une unité de sens

Langues : allemand, néerlandais, grec, suédois, danois, russe, etc.
EN *toolbar*, *waterproof*
DE *Staatsfeind*, *Natriumdampfniederdrucklampe*

Composition dans des langues différentes

◦ **Concaténation** :
EN *kilowatthour*, FR *poisson-pilote*

◦ **Interfixe** (morphème de liaison) :
DE *Staatsfeind* [ennemi d'état] = *Staat* [état] + *s* + *Feind* [ennemi]

◦ **Modification du radical** :
DE *Gänseklein* [abats d'oie] = *Gans* PL [oie] + *Klein* [petit] ;

◦ **Composition néoclassique** :
FR *biomasse*, DE *Turbomaschine*

Composition vs. Incorporation, agglutination (langues polysynthétiques)
Inuit *angyaghillangyugtug* (il veut acheter un grand bateau)
angya.ghilla.ng.yug.tug = bateau.grand.acquérir.volonté.3SING

Problémes de la composition morphologique en TALN

- **Composé ou non ?**

- **Identifier les frontières**
DE *Traktionsbatterie*
traktion + *batterie* ?
trakt + *ion* + *batterie* ?

- **Trouver les lemmes des composants**
RU *vetrogenerator* = *veter* [vent] + *generator* [générateur]
vetro (composant) → veter (lemme)

- **Désambiguïser**
SW *bildrulle* = *bild* + *rulle* ("bobine de film")
bildrulle = *bil* + *drulle* ("mauvais conducteur")

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.ht*

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 sparsity 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