Traitement Automatique des Langues
pour les
Systèmes d'Information

Morphologie, terminologie, lexiques

Damien Nouvel
Nathalie Friburger
Jean-Yves Antoine

Morphologie, terminologie, lexiques
Plan

- Mots, tokens, formes, lemmes
- Morphologie flexionnelle
- Morphologie dérivationnelle
- Expressions composées
- Terminologie
- Lexiques, dictionnaires, thesaurus

Morphologie, terminologie, lexiques
Plan

- Mots, tokens, formes, lemmes
- Morphologie flexionnelle
- Morphologie dérivationnelle
- Expressions composées
- Terminologie
- Lexiques, dictionnaires, thesaurus

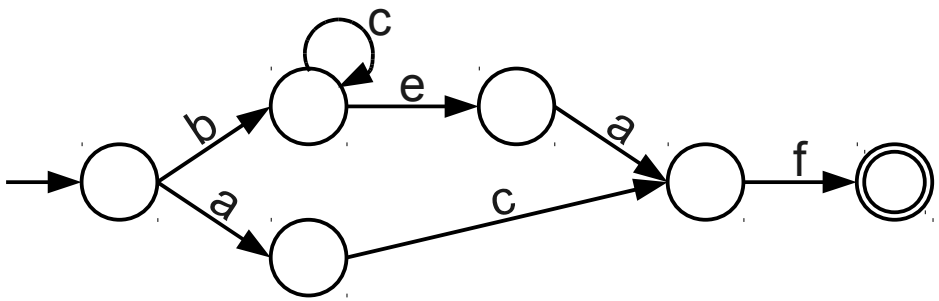
Morphologie, terminologie, lexiques
Mots, tokens, formes, lemmes

- Unités « logiques » pour le traitement de textes :
  - Document ⊃ paragraphe ⊃ phrase ⊃ « mot » ⊃ « caractère »
- Mais un « **mot** » n'est pas un unité bien définie :
  - Exemples : avion, mangée, très, Robert, SNCF, 42...
  - **Forme** : notion graphique du mot (Igor Mel'čuk)
  - **Lemme** : intersection entre une **forme** (graphique) et un **sens**, parfois par composition de morphèmes
    - **Morphème** : plus petite unité porteuse de sens (par ex. « re »)
  - **Token** (jeton, identificateur) : unité **minimale** d'information détectée lors de l' « analyse lexicale » ou « **tokenization** »
    - En français, souvent nommée « lexème »

Morphologie, terminologie, lexiques
Mots, tokens, formes, lemmes

- Théorie du langage, quelques rappels :
  - **Alphabet**, ensemble de symboles atomiques :
    - Lettres : {a, b, c, d, e, é, è, ê, f... z} ∪ { α, β, γ...} ∪ ...
    - Chiffres : {0, 1, 2...9 }
    - Ponctuation : {. , !, ?, (,)...}
    - Symboles mathématiques : {+, -, *, /, =, (,)...}
    - Symboles monétaires : {$, €, £, ¥...}
    - ...
  - **Opérateurs du langage** :
    - **Concaténation** : « . » ou « ø »
    - **Union** : « + » ou « | »
    - **Répétition** : « * », « + », « ? » ou « {n, m} »

- **Automates**, quelques rappels :
  - Représentation sous forme de **graphe** :



  - Représentation par **expressions régulières** :

    $(bc^*ea|ac)f$

- **Grammaires**, quelques rappels :
  - Régies par des « **règles de production** »
  - Hiérarchie de **Chomsky** : FG > CG > CFG > RG

Morphologie, terminologie, lexiques
Mots, tokens, formes, lemmes

- **Tokenization** :
  - Segmenter un texte en « **unités minimales** » pour le traiter
  - Ensemble d'**automates** qui reconnaissent les **tokens** en acceptant des chaînes, éventuellement en les typant :
    - Lexème : -?[A-Z] ?[a-z]*
    - Ponctuation : .|...|,|!|?
    - Nombre : -?[0-9]*(,|.)[0-9]*
    - ...
  - Exemple :

Les étudiants, ceux du M2-SIAD, n'ont-ils pas tous 15,3 de moyenne ?

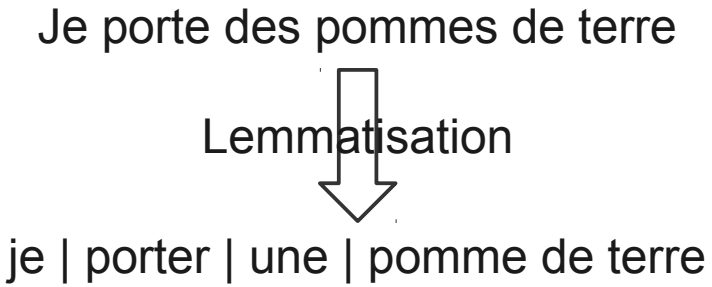
Tokenization

Les | étudiants | , | ceux | du | M2-SIAD | , | n' | ont | -ils | pas | tous | 15,3 | de | moyenne | ?

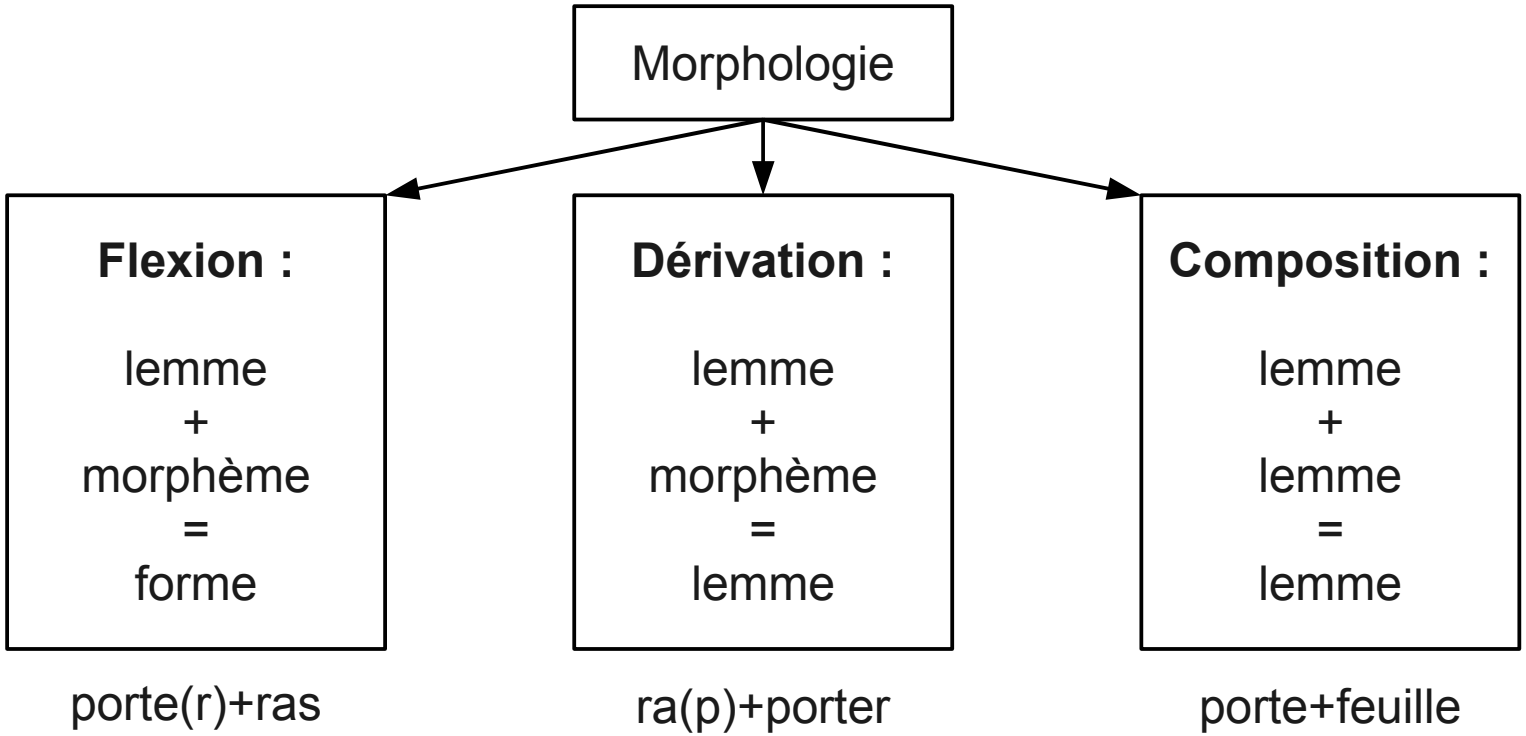
- **Racines** et **formes**
  - A partir d'un ensemble de **formes** (token) apparentées il est possible de retrouver sa **racine** (son radical) par « **racinisation** » : la plus petite / ancienne partie commune :
    - La racine de « déplaira » et « déplaisent » est « déplai »
    - Si l'on ajoute « déplu », la racine devient «dépl »
    - Avec « plaira », la racine est maintenant « pl »
  - Inversement, des mécanismes permettent de construire les **formes** possibles à partir de racines (flexion, dérivation, composition) :
    - Par exemple « place » peut donner : « places », « placerions », «déplacer », « déplacements », « remplaceras », etc.
    - Ces formes sont **plus ou moins prévisibles**

Morphologie, terminologie, lexiques
Mots, tokens, formes, lemmes

- **Lemme** : unité autonome (composée de morphèmes) permettant de constituer le **lexique** d'une langue :
  - **Morphèmes** : les « parties » du lemme
  - **Autonome** : peut-être utilisé tel quel dans une phrase
  - Elément du **lexique** : éviter les redondances dans un lexique
- **Lemmatisation**, trouver les **lemmes** pour chaque **token** au sein d'une phrase :



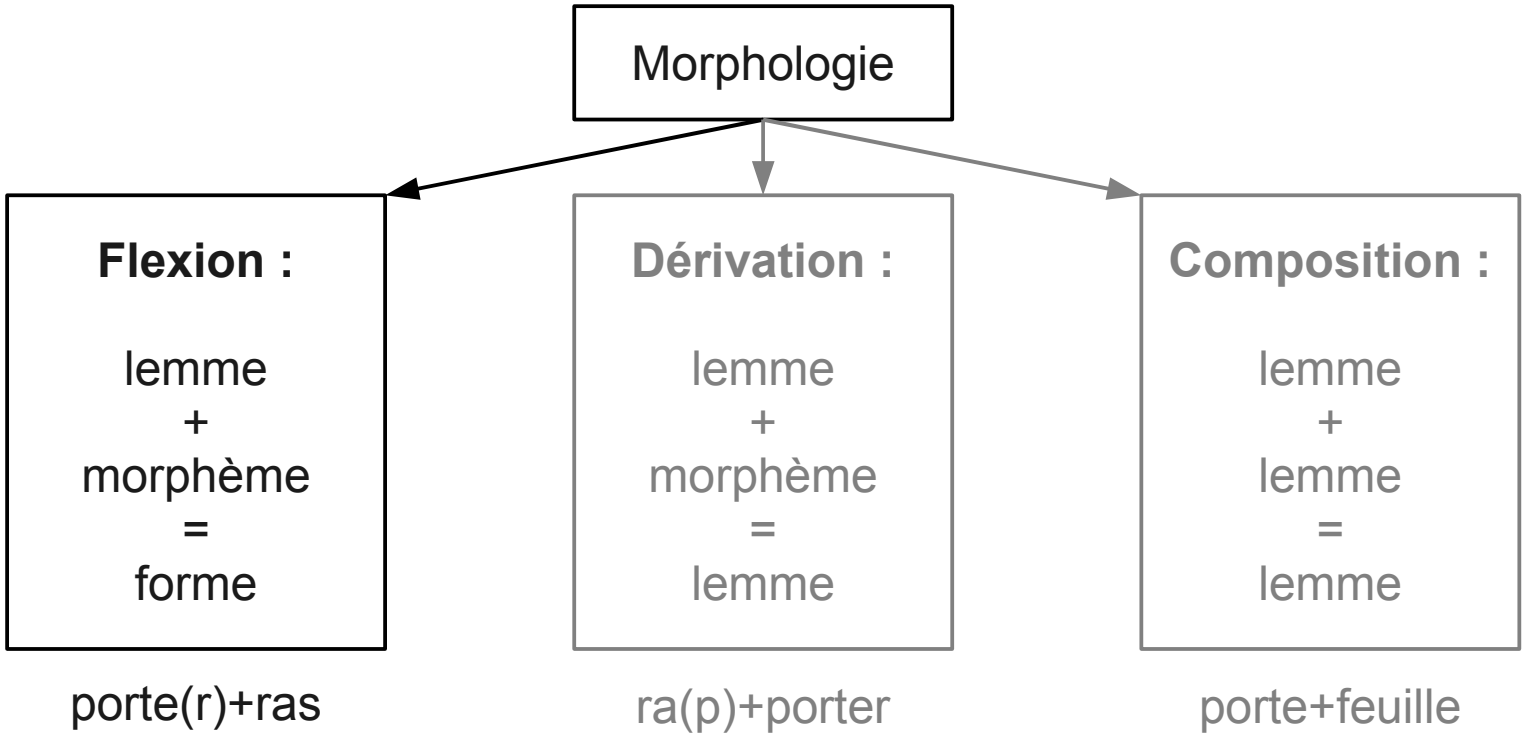
- Règles de morphologie :



Morphologie, terminologie, lexiques
Plan

- Mots, tokens, formes, lemmes
- Morphologie flexionnelle
- Morphologie dérivationnelle
- Expressions composées
- Terminologie
- Lexiques, dictionnaires, thesaurus

- Règles de morphologie :



- Flexion - modification de l'**affixe** d'un lemme selon son rôle dans la phrase, sans en modifier le sens :
  - **Déclinaison** (noms, articles, adjectifs, pronoms) :
    - **Genre** (masculin / féminin) : « heureu-*x* » → « heureu-*se* »
    - **Cas** (nominatif / accusatif, etc.) : « ros-*a* » → « ros-*ae* »
    - **Nombre** (singulier / pluriel) : «chev-*al* » → « chev-*aux* »
  - **Conjugaison** (verbes) :
    - **Mode** et **temps** (indicatif / subjonctif / présent / passé / futur) : « chant-*er* » → « chant-*ais* », « chant-*ai* », « chant-*erai* »
    - **Personne** (1e / 2e / 3e) : « march-*er* » → « march-*ai* », « march-*as* », « march-*a* »
    - **Nombre** (singulier / pluriel) : « parl-*er* » → « parl-*es* », « parl-*ez* »

Morphologie, terminologie, lexiques
Mots, tokens, formes, lemmes

- Importance des mécanismes **lemmatisation** / **racinisation** pour la représentation du langage :
  - Eviter des dictionnaires trop volumineux comportant toutes les formes possibles (en français, ~60 000 lemmes « courants », mais >500 000 formes fléchies « courantes »)
    - Gain en **espace** de stockage
    - Moindre complexité pour l'encodage du lexique
  - Parvenir à une représentation **structurée** du texte, à partir de laquelle on peut faire des traitements :
    - Le « sens » est plutôt lié au lemme qu'à la forme (par exemple : temps verbaux, masculin / féminin, etc.)
    - Lier les catégories grammaticales aux lemmes plutôt qu'aux formes : désambiguisation morpho-syntaxique (à venir)

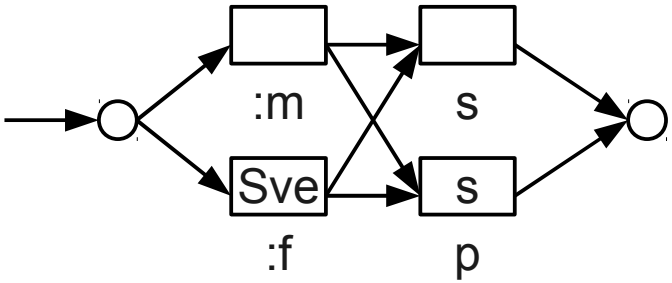
- Systèmes flexionnels pour quelques langues :

	Allemand	Anglais	Arabe	Serbe
Genres	Masculin Féminin Neutre		Masculin Féminin	Masculin Féminin Neutre
Cas	Nominatif Accusatif Datif Génitif		Nominatif Accusatif Génitif	Nominatif Accusatif Datif Génitif Instrumental Locatif Vocatif
Nombres	Singulier Pluriel	Singulier Pluriel	Singulier Duel Pluriel	Singluier Pluriel
Modes	Indicatif Impératif Conjonctif	Indicatif Participe Infinitif Conditionnel	Accompli Inaccompli	Infinitif Indicatif Conditionnel Impératif Participe Adjectival
Personnes	1 / 2 / 3	1 / 2 / 3	1 / 2 / 3	1 / 2 / 3

- Un système de flexion, pour une langue, peut-être décrit à l'aide de **règles**
- Généralement, la **flexion automatique** (par catégories flexionnelles) est réalisée à partir de la **racine** ou du **lemme** :
  - Tableau de **suffixes** (racine)

:ms	:mp	:fs	:fp
f	fs	ve	ves

  - **Transducteurs** de flexion (lemme)



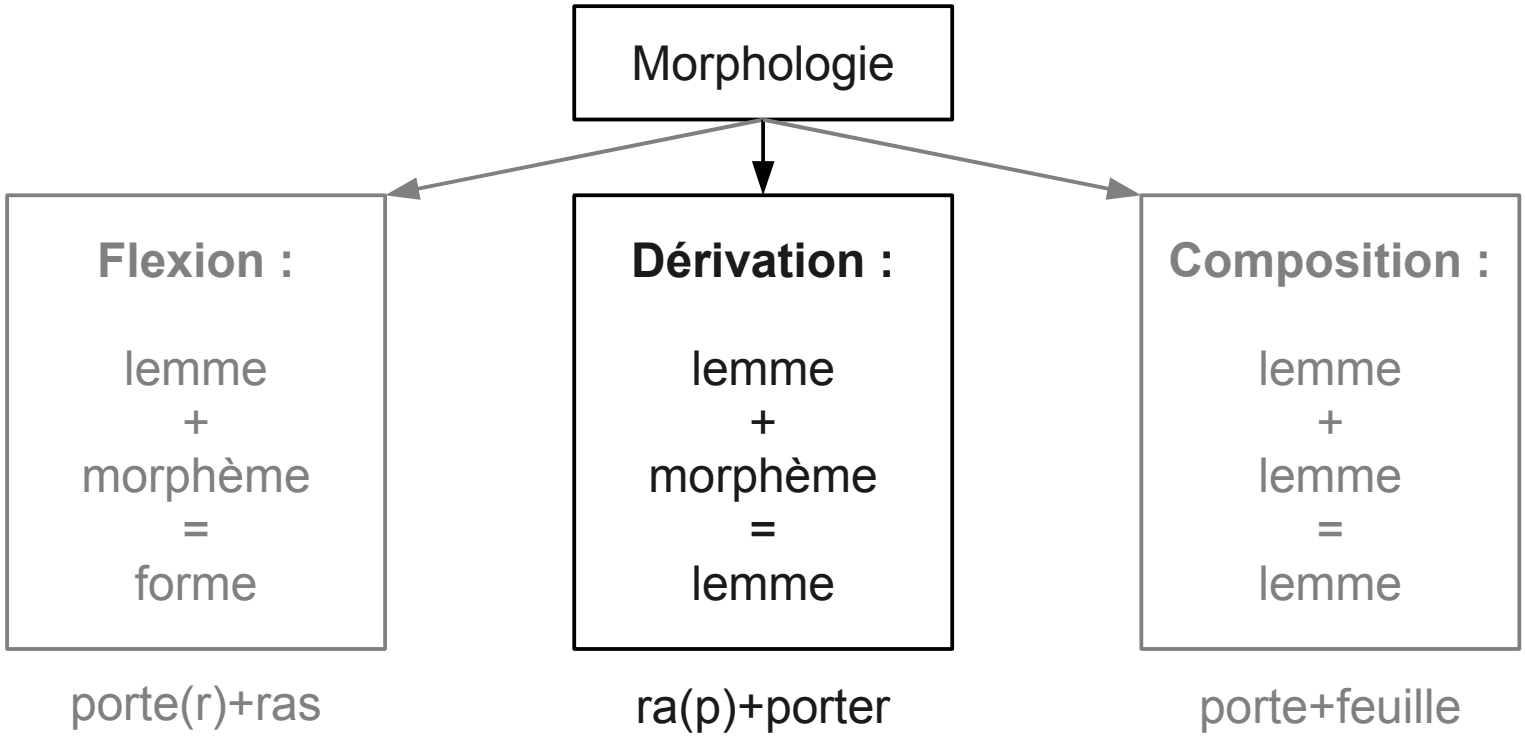
- Exemple : neuf, veuf, actif, bref

- **Analogie**, principe général :
  - A est à B ce que C est à D, « A : B :: C : D »
    - Avec A = a_1a_2, D = d_1d_2 alors B = a_1d_2 et C = d_1a_2
- Pour la morphologie, on **décompose** les termes :
  - « pensais » : « penser » :: « marchais » : « marcher »
    - Avec a_1 = « pens », a_2 = « ais », d_1 = « march » et d_2 = « er »
  - Le principe s'applique pour **déduire de nouvelles formes** :
    - Trouver le plus grand préfixe commun entre A et B : a_1
    - En déduire les suffixes a_2, d_2 puis d_1 pour déterminer C
    - Par exemple, « grandissait » : « grandir » :: « ? » : « gémir »
  - Cela ne fonctionne qu'en supposant un mécanisme de flexion

Morphologie, terminologie, lexiques
Plan

- Mots, tokens, formes, lemmes
- Morphologie flexionnelle
- Morphologie dérivationnelle
- Expressions composées
- Terminologie
- Lexiques, dictionnaires, thesaurus

- Règles de morphologie :



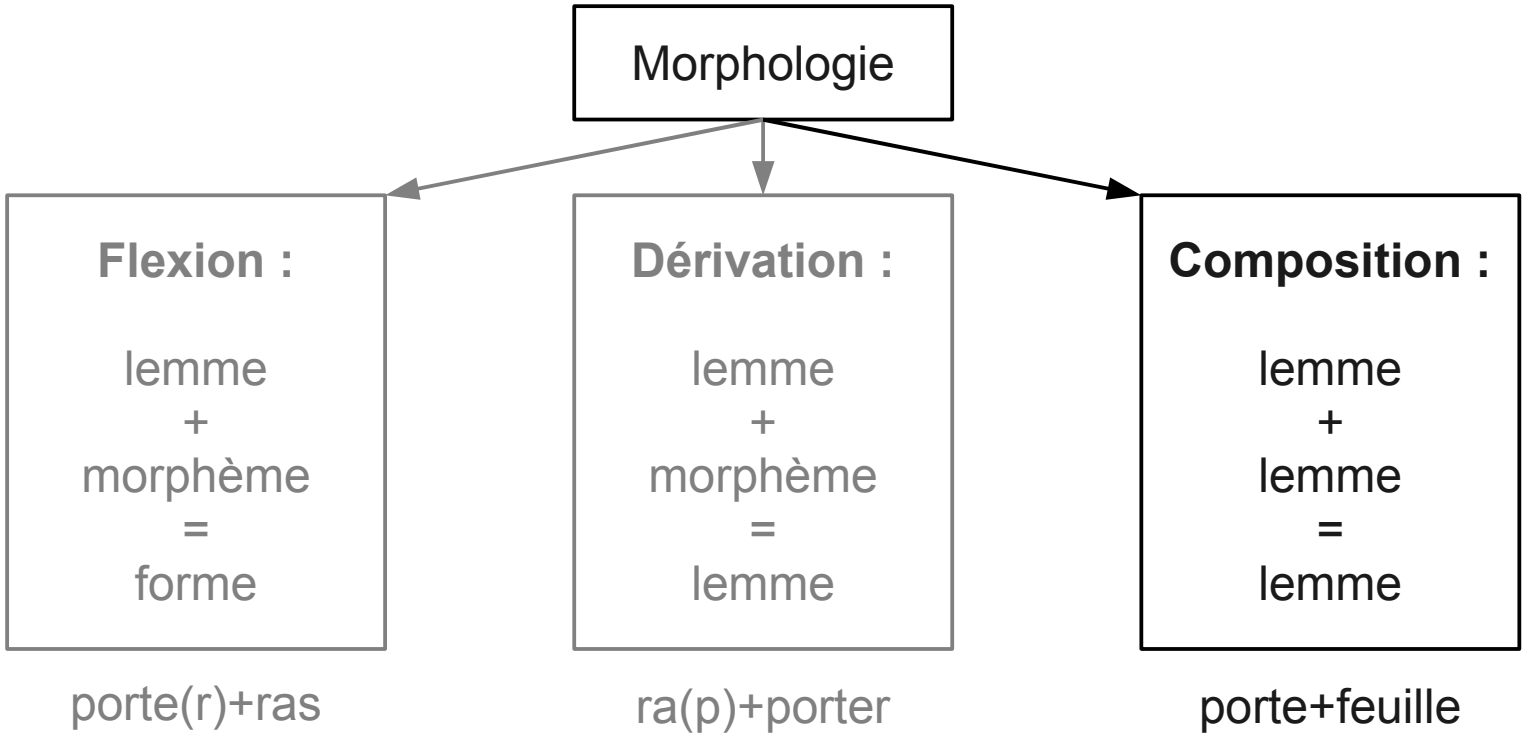
- **Dérivation** : création d'un nouveau lemme par combinaison (composition) entre un **lemme** existant et un **morphème** qui en altère le sens :
  - **Agglutination** de préfixes / suffixes avec le lemme :
    « constituer » → « constitution » → « constitutionnel » →
    « anticonstitutionnel » → « anticonstitutionnellement »
- Certaines dérivations sont plus ou moins **acceptables** (usages attestés) :
  - « changeage », « affichable », « tissement », « bravitude », « antifillonisme », « cultivation »...
- **Affixes** courants en français : -able, -ité, -et(te), -is(er), -ifi(er), -eur, -ure, -ment, -tion, -oir, dé-, in-, re-, a-, en-...

- La dérivation provoque souvent un **changement de catégorie grammaticale** (verbe, nom, adjectif, etc)
- Quoiqu'il soit théoriquement possible de trouver le sens d'un mot selon ses morphèmes, c'est de manière générale assez peu utilisé automatiquement
- Les lemmes dérivés sont entrés manuellement dans le lexique

Morphologie, terminologie, lexiques
Plan

- Mots, tokens, formes, lemmes
- Morphologie flexionnelle
- Morphologie dérivationnelle
- Expressions composées
- Terminologie
- Lexiques, dictionnaires, thesaurus

- Règles de morphologie :



Morphologie, terminologie, lexiques
Expressions composées

- **Expression composée** : lemmes juxtaposés dont le sens a une **signification différente** des lemmes qui le composent, « expression figée »
- Divers niveaux d'**agglutination** :
  - **Locutions** (tokens séparés) : « pomme de terre », « cordon bleu », « garde fou », « petit pois »...
  - Tokens séparés par un symbole de ponctuation : « coupe-gorge », « abat-jour », « aujourd'hui », « presqu'île »
  - Tokens **unifiés** (collés) : « gentilhomme », « monsieur », « lorsque », « toutefois », « vinaigre », « autobus »
- Peuvent être fléchies (« il se rendit compte ») et parfois dérivés (« mise en oeuvre », « avant-gardisme »)

Morphologie, terminologie, lexiques
Expressions composées

- Les expressions composés représentent un défi :
  - Les lemmes doivent être **groupés lors de la tokenisation**
  - Elles introduisent de l'**ambiguïté** (plus ou moins figées)
  - De nouvelles apparaissent tous les jours
  - Contrairement à la dérivation, par définition le sens ne peut être déduit des lemmes dont elles sont composées
- Elles sont généralement traitées comme des **unités particulières du lexique**
- Elles exigent la plupart du temps un prétraitement lors de la tokenization

Morphologie, terminologie, lexiques
Plan

- Mots, tokens, formes, lemmes
- Morphologie flexionnelle
- Morphologie dérivationnelle
- Expressions composées
- Terminologie
- Lexiques, dictionnaires, thesaurus

Morphologie, terminologie, lexiques
Terminologie

- **La** terminologie : science des termes qui appartiennent à un **domaine de spécialité** :
  - Par ex. : biologie, chimie, aéronautique, finance, musicologie...
  - Relatif à l'**usage** de termes au sein d'un domaine
  - Fait appel à la **sémantique** pour enrichir un lexique
- **Une** terminologie : thesaurus, dictionnaire, lexique, banque de termes...
- Peut-être extrêmement spécialisé :
  - Par ex. les « Affaires indiennes et du Nord Canada » définissent « Inuit », « Nunavut », « Droits ancestraux » mais aussi « Bande », « Cession », « Nord », « Projet de loi C-31 »

Morphologie, terminologie, lexiques
Terminologie

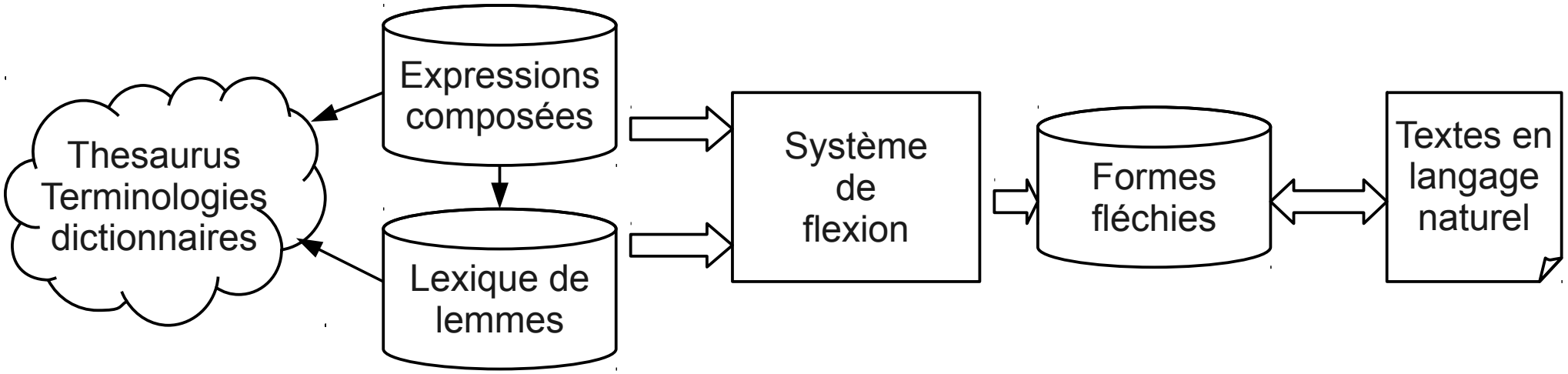
- Recenser et définir des **termes** :
  - « carte-mère », « carte graphique », « bus », « mémoire »
  - « palonnier », « manche à balais », « boîte noire », « enregistreur de vol », « dérive », descente »
  - « marché », « cours », « CAC40 », « NYSE », «NASDAQ »
- Fait appel à l'étude :
  - **Sémantique**
  - **Ontologies** (organisation de concepts : Protégé, WordNet)
    - Paraphrases, reformulations, synonymes, homonymes
    - Hyponymes, hyperonymes, méronymes, antonymes...
  - **Collocations** (expressions composées)
  - Méthodes d'extraction de connaissances (biomédical)

Morphologie, terminologie, lexiques
Plan

- Mots, tokens, formes, lemmes
- Morphologie flexionnelle
- Morphologie dérivationnelle
- Expressions composées
- Terminologie
- Lexiques, dictionnaires, thesaurus

- Différence entre :
  - **Lexique** (glossaire) : recensement d'expressions linguistiques
  - **Dictionnaire** : définitions, par lemmes
  - **Thesaurus** : synonymes entre lemmes
- Ce sont des ressources pour des **traitements linguistiques** :
  - Génération
  - Analyse
  - Classification
  - Recherche d'information
  - Désambiguisation

- Système basique TAL morphologique et terminologique :



- Le **lexique de formes fléchies** permet de :
  - **Lemmatiser** : trouver pour des **tokens** le **lemme** et **la catégorie syntaxique**, éventuellement le(s) sens associé(s)
  - Pour un **lemme** donné de donner les **formes** possibles selon la **catégorie syntaxique**
- Nécessité d'algorithmes rapides sur plusieurs centaines de milliers d'entrées : **automates**

Forme	Lemme	Cat.	F.	L.	C.	F.	L.	C.
marchais	marcher	:Vi2s	beau	beau	:Ams	action	action	:Nms
marchait	marcher	:Vi3s	belle	beau	:Afs	actions	action	:Nmp
marchions	marcher	:Vi1p	beaux	beau	:Amp	cheval	cheval	:Nms
marchiez	marcher	:Vi2p	belles	beau	:Afp	chevaux	cheval	:Nmp
marchaient	marcher	:Vi3p	bleu	bleu	:Am	bétail	bétail	:Nmi