Analyse syntaxique
et
application aux langues naturelles

Jacques Farré

et

Sylvain Schmitz

Plan du cours

2. Introduction, généralités, rappels
3. Analyse ascendante, LR, LALR, …
4. N-SLR, N-LALR et LALR-R
5. GLR, Earley, CYK
6. TP d’utilisation d’outils

Bibliographie sommaire (analyse syntaxique)

- Parsing Techniques – A Practical Guide, 2ème edition, D. Grune & C.J.H. Jacobs, Springer Verlag, 2006
- Parsing Theory, S. Sippu & E. Soisalon-Soinnen, Springer Verlag, 1988
- R. Grishman, Computational Linguistic: an introduction, Cambridge University Press, 1986
- L'intelligence artificielle et le langage naturel, G. Sabah, Hermès, 1988-1989 (2 volumes)
- Les nouvelles syntaxes : grammaires d'unification et analyse du français, A. Abeillé, Colin, 1993

Principaux journaux

- Computational Linguistics
- Computers and the Humanities
- Computer, Speech & Language
- Computer Assisted Language Learning
- Grammars
- Journal of Language and Computation (logic, linguistics, formal grammar, and computational linguistics)
- Linguistics
- Literary and Linguistic Computing
- Natural Language Engineering
- Machine Translation
- …

Quelques conférences

- Sous l'égide de l' (European) Association for Computational Linguistic : ACL
- COLING (Int'l Conf on Comp. Ling.)
- CICLing
- LATIN (Latin American Theoretical Informatics)
- ANLP (Applied Nat. Lang. Processing)
- IEEE Int'l Conf on Natural Language Processing and Knowledge Engineering
- CIAA (Implementation & Application of Automata)
- …

Domaines d'application (1)

- Traitement documentaire
  - Traduction automatique (cf société Systran)
    - Bons traducteurs dans un domaine spécialisé pour préparer le terrain à une traduction par un humain
  - Recherche de documents
    - Veille scientifique
    - Routage, indexation de documents
  - Analyse de documents
    - Graphe de relations entre termes d'un document
  - E-learning

Domaines d'application (2)

- Production de documents
  - Correcteurs d'orthographe, de syntaxe, …
  - Correcteurs stylistiques (bonnes pratiques rédactionnelles dans un domaine donné)
  - Génération automatique à partir de spécifications plus ou moins formelles (documents techniques, juridiques, …) de documents finalisés par des humains

Domaines d'application (3)

- Interfaces homme-machine
  - Interrogation de bases de données
    - Traduction langage naturel → SQL
  - E-learning (encore)
  - Interfaces vocales
    - Téléphonie
    - Ordinateurs (et autres machines) mains libres
    - Marché de plusieurs milliards de dollars

Un peu d'histoire (1)

- A l'origine, les militaires (comme souvent)
  - Années 50 : traduction de documents russes (nucléaire, recherche spatiale, …)
  - Concentré sur l'élaboration de dictionnaires bilingues
  - Traduction mot à mot pour l'essentiel
    - Exemple célèbre : *The spirit is willing but the flesh is weak* donne traduite en russe puis retraduite en anglais *The vodka is strong but the meat is rotten*

Un peu d'histoire (2)

- Travaux de Noam Chomsky (fin des 50) sur les grammaires formelles et leurs relations avec les langues naturelles
- Parallèlement, travaux sur l'intelligence artificielle (Mc Carthy, Minsky, …)
  - Système ELIZA (MIT, 1966) : simulation d'un dialogue entre un psy et son patient (application de modèles reprenant des mot-clés du patient)

Un peu d'histoire (3)

- Travaux sur la représentation des connaissances dans les années 70 (Minsky, Shank, …)
  - La sémantique prime, la syntaxe est jugée secondaire
- Mais aussi développement de l'analyse syntaxique (dans le cadre des langages de programmation, mais récupéré en partie pour les langues naturelles)

Problématique

- Difficultés de plusieurs ordres, notamment
  - Ambiguïtés
    - Des terminaisons : que marque un *s* final ?
    - Des lexèmes : *les poules du couvent couvent*
    - Des formes grammaticales : *il poursuit les filles à vélo*
  - Implicite/contextuel : à qui/quoi réfère *il*
    - Le prof a saqué cet élève parce
      - qu'*il* ne peut pas le sentir (il, prof)
      - qu'*il* lui a cassé les pieds toute l'année (il, élève)

Niveaux de traitement et outils

- Lexical : découper le texte en mots et calculer leur genre (adjectif/verbe/nom/préposition…)
  - Dictionnaires, automates à états finis
- Syntaxique : trouver la structure de la phrase (quelque chose comme *sujet verbe complément)*
  - Grammaires, arbres, forêts (partagées)
- Sémantique : “donner” un sens
  - Graphes conceptuels, prédicats logiques : *ce chien est curieux, il poursuit les [filles à vélo]*
- Pragmatique : juger la pertinence selon le contexte
  - *J’ai froid → ferme la fenêtre/serre moi dans tes bras*

Traitement lexical : buts (1)

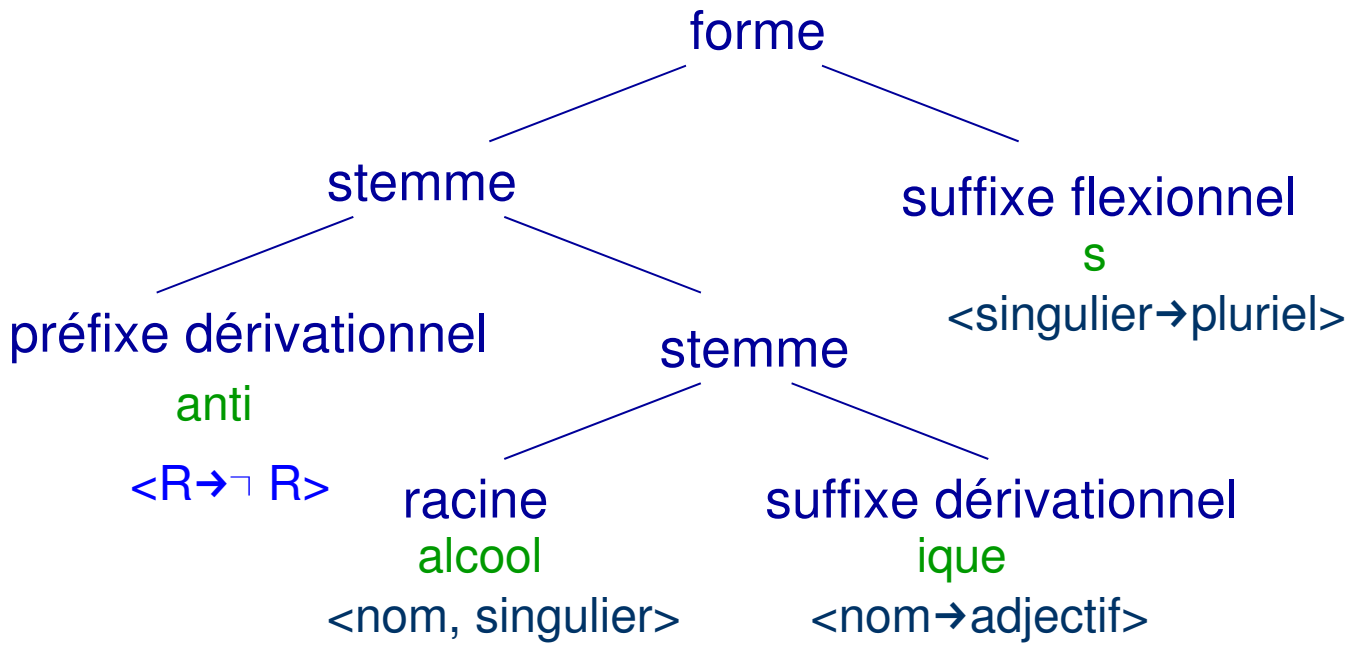
- Décomposition en unités lexicales (lexème)
  - Trouver les *mots* : ambiguïté des séparateurs, par ex. le point (point final, abréviation, sigles comme S.N.C.F.)
  - Un segmenteur (tokenizer) doit connaître les règles d'usage des signes de ponctuation de la langue
  - Une forme enrichie des textes (HTML…) peut être une aide
  - Taux d'erreur des meilleurs segmenteurs pour reconnaître la fin d'une phrase ≈ 1%

Traitement lexical : buts (2)

- Caractériser les lexèmes
  - Nom, adjectif, verbe, …
  - Singulier, pluriel, diminutif, abréviation, …
- Difficile de ranger tous les mots possibles d'une langue dans un dictionnaire :
  - conjugaison des verbes, composition de noms, néologismes, …
- Avoir une forme de dérivation des mots à partir d'une racine

Traitement lexical : moyens (1)

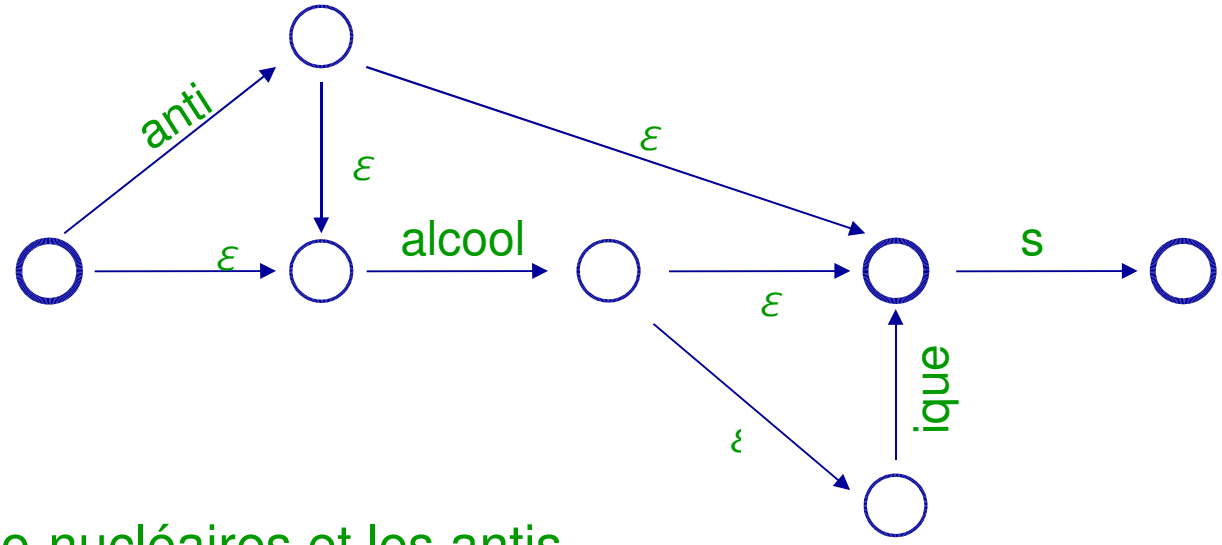
- Exemple de règle de formation des mots (français)



antialcooliques : <adjectif, pluriel, (idée d'opposition à alcool)>

Traitement lexical : moyens (2)

- Par un automate d'état fini (non déterministe)



Les pro-nucléaires et les antis

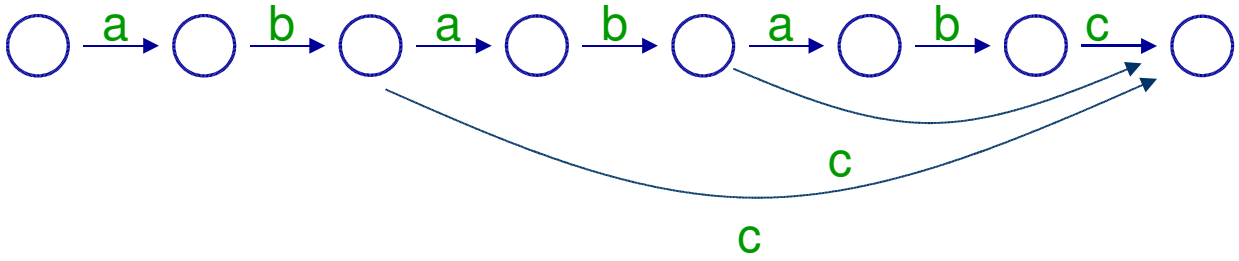
Antimilita(i)r(e)iste  → règles d'ajustement phonologique

Traitement lexical : moyens (3)

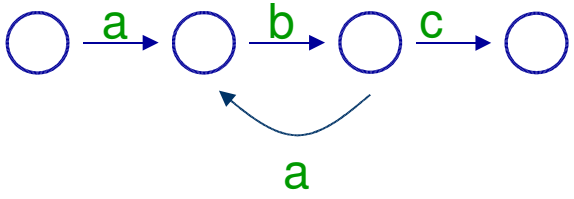
- Relations entre concepts
  - Synonymie : X est équivalent à Y
  - Antonymie : X est l'opposé de Y
  - Hyponymie : X est une spécialisation de Y
  - Hyperonymie : X est une généralisation de Y
- Mécanismes de composition de mots
  - Nom + adjectif (*système distribué*)
  - Nom + préposition + nom (*réseau de neurones*)

Traitement lexical : quelques problèmes

- Minimisation des automates
  - Par un automate reconnaissant un langage plus grand (une *couverture*)

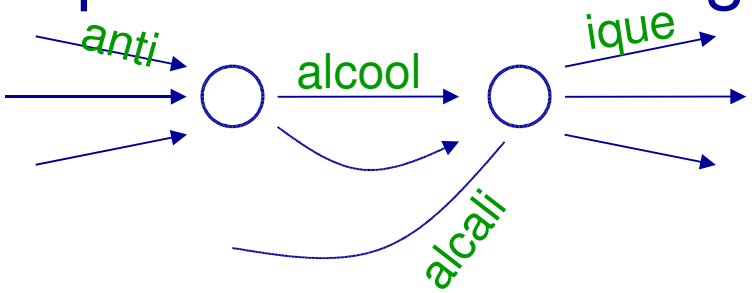


donne (en $O(n^2)$, peut-on faire mieux ?)



Traitement lexical : quelques problèmes

- Tolérance aux erreurs (fautes d'orthographe)
  - Notion de région dans laquelle on cherche à corriger l'erreur : pour antialcalique, chercher à corriger par un mot de la région



  - Essai Google :
    Essayez avec cette orthographe : ***antialcoolique***
    Aucun document ne correspond aux termes de recherche spécifiés (**antialcalique**)

Traitement syntaxique : buts

- Vérifie la validité d'une suite de lexèmes
  - Permet de lever des ambiguïtés lexicales comme dans *les poules du couvent couvent*

nom

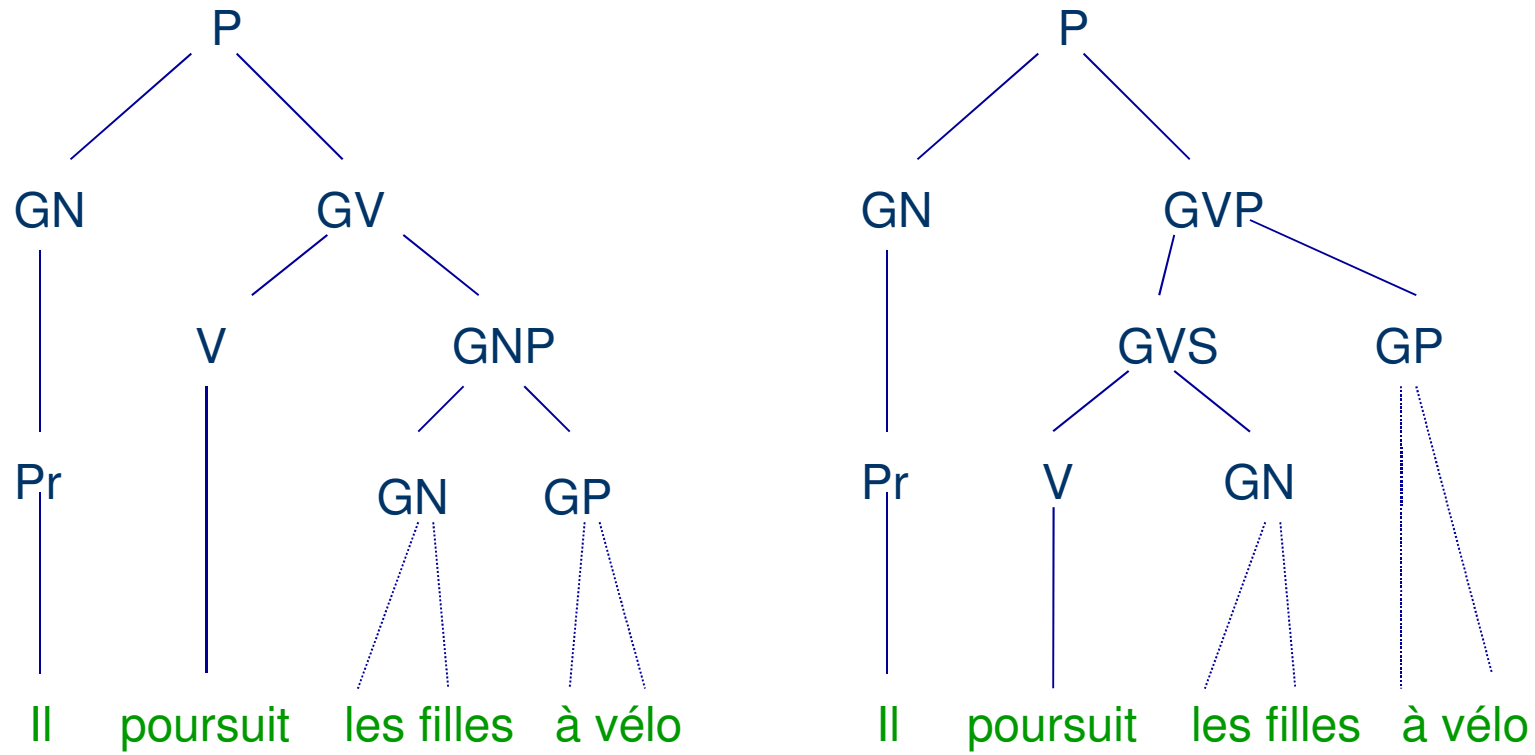
sujet *verbe*

  - Vérification de fautes d'accord :
    groupe-nominal → *article-pluriel nom-pluriel*
    | …
  - *Dégage une organisation hiérarchique : sujet, verbe, complément, …*

Traitement syntaxique : outils

- Grammaires non contextuelles (context free)
  - Plusieurs (dizaines de) milliers de règles
  - Nécessairement ambiguës, et reconnaissent un sur-langage
    - Quelle que soit la méthode d'analyse employée, construction d'une forêt partagée d'arbres syntaxiques : les phases sémantique/pragmatique choisiront l'arbre final
- Variante : grammaires probabilistes

Forêt partagée : arbres de dérivation



Ambiguîté non soluble en l'absence d'informations sémantiques contextuelles

Traitement syntaxique : problèmes

- Analyseurs courants basés sur LR(0), SLR (1) ou LALR(1)
  - Provoquent des conflits là où il n'y a pas d'ambiguïté
  - Élargir la taille de la fenêtre ?