

Cours sur le traitement automatique des langues

Violaine Prince
Université de Montpellier 2
LIRMM-CNRS

Introduction

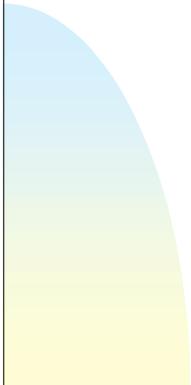
- Les outils
 - ◆ Analyseurs
 - ◆ Bases de connaissance
- Les applications
 - ◆ Ingénierie linguistique
 - ◆ Aux autres domaines de l'informatique
 - ◆ Aide à la recherche linguistique

Ingénierie linguistique :

- Aide à la traduction automatique
- Correcteurs grammaticaux et orthographiques
- Dictionnaires
- Alignement de corpus multilingues
- Résumés automatiques

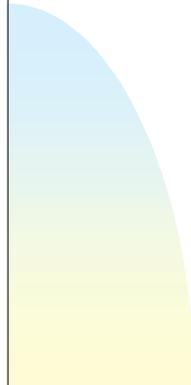
Application aux autres domaines de l'informatique

- Moteurs de recherche d'information
- Interrogation de bases de données
- Tuteurs intelligents
- Informatique documentaire
- Reconnaissance de la parole continue



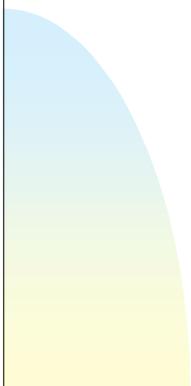
Aide à la recherche linguistique

- Recherche de fréquences
- Aide à l'analyse de textes
- Typage de données textuelles



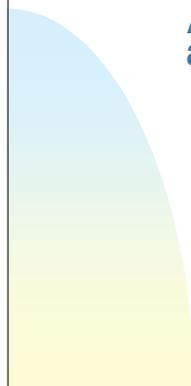
Thématiques et domaines

- Les différents types de « TAL »
 - ◆ Informatique linguistique
 - ✦ Algorithmique et info théorique
 - ✦ Intelligence Artificielle
 - ✦ Systèmes à base d'agents
 - ◆ Linguistique informatique
 - ✦ Statistiques
 - ✦ Logique



Éléments traités

- Analyse automatique
 - ◆ Modèles et outils
- Terminologie
- Présentation des options



Analyse automatique

- Dimensions d'analyse
 - ◆ Morphologique
 - ◆ Syntaxique
 - ◆ Sémantique
 - ◆ Pragmatique

Analyse morphologique

- Objectif :
 - ◆ Reconnaissance de mots dans un texte
 - ◆ Reconnaissance de la ponctuation
 - ◆ Affectation d'une catégorie grammaticale au mot
- S'appelle LEMMATISATION ou ETIQUETAGE

Exemple

- Ajouter du texte
 - ◆ Reconnaissance de la frontière des unités lexicales (ul)
 - ◆ Reconnaissance de l'ul comme « motif » présent dans un thésaurus : catégorie « verbe », forme « infinitif »
 - ◆ Lettre majuscule A : reconnaissance du début du texte

AJOUTER

Quelques difficultés

- J'ajoute du texte
 - ◆ Reconnaître une forme de « je » pronom personnel
 - ◆ Reconnaître une forme du motif « ajouter » ou le reconnaître comme motif : catégorie « verbe », forme « première personne du singulier ».

La multiplicité des signes

- Les signes spéciaux :
 - ◆ Qui interviennent dans une unité lexicale :
 - + - , exemple : porte-manteau
 - + ' , exemple : aujourd'hui
 - ◆ qui marquent la contraction :
 - + ' , exemple : j'arrive
 - ◆ Qui marquent un début ou une fin d''unité composée :
 - + « », (), majuscule et point, — —.

- Les signes de ponctuation :
 - ◆ , ; :
- Les signes d'énumération :
 - ◆ 1) nombre suivi d'une parenthèse fermante
 - ◆ [] , - , *
- Le symbole du dialogue
 - ◆ _
- Les signes d'annotation (*), (1)
- Les signes arithmétiques et les nombres inclus dans un texte

L'ambiguïté

- Des signes :
 - ◆ l'apostrophe, le tiret, la parenthèse fermante
- Des catégories affectables à une ul :
 - ◆ une texture (ferme) ← adjectif
 - ◆ je (ferme) la porte ← verbe
 - ◆ la (ferme) de Jean ← nom

- De la majuscule : début de texte ou emphase
- ambiguïté de forme précise
 - ◆ je ferme la porte
 - + ferme
 - catégorie : VERBE
 - forme : 1ere personne du singulier (FORME FLECHIE)
 - ◆ Il ferme la porte
 - + ferme
 - catégorie : VERBE
 - forme : 3ème personne du singulier

Le côté « multiplicatif » de l'ambiguïté de catégorie

- Je ferme la porte

La combinatoire théorique

- pronom verbe pronom verbe
- pronom verbe article verbe
- pronom verbe pronom nom
- pronom verbe article nom
- pronom nom pronom verbe
- pronom nom article verbe
- pronom nom pronom nom
- pronom nom article nom
- pronom adjectif pronom verbe
- etc. soit 12 combinaisons alors qu'il n'y en a qu'une seule de bonne...

LA BONNE COMBINAISON

Les différentes techniques d'analyse morphologique

- Soit une u dans un texte T
 - ◆ Etiquetage
 - ✦ affectation d'une catégorie grammaticale et/ou d'une forme à U
 - ◆ Lemmatisation
 - ✦ étiquetage et reconnaissance de U comme élément de dictionnaire

Exemples

- Je ferme la porte
 - ◆ Etiquetage :
 - ✦ (« je », pronom personnel), (« ferme », verbe), (« la » article), (« porte », nom)
 - ✦ étiquetage en bi-grammes
 - (« U », C_U)
 - ◆ Lemmatisation
 - ✦ Etiquetage plus
 - ✦ (« ferme », verbe : FERMER)
 - (« U », C_U , LEXEME)

- Etiquetage tri-gramme
 - ◆ (« U », C_U , F_U)
 - ✦ où F est la forme prise par U (forme fléchie)
- Lemmatisation avec étiquetage tri-gramme
- (« U », C_U , F_U , LEXEME)
- Un **lexème** est une unité lexicale significative.
 - ◆ Exemples : FERMER, JE, LA, PORTE, PORTER...

Quelques éléments de terminologie

- **Entrée lexicale :**
 - ◆ Unité lexicale qui sert d'entrée du dictionnaire. Elle est généralement représentée par :
 - la chaîne de caractères X qui la définit
 - le lexème L auquel elle est associée
 - la catégorie grammaticale associée
 - la ou les forme(s) fléchie(s) du lexème catégorisé prise par la chaîne de caractères.
- † (X, L, C, {F_x})

Exemples

- Il existe trois entrées lexicales pour l'ul « ferme »
 - (« ferme », FERMER, verbe, { 1ère personne du singulier, 3ème personne du singulier})
 - (« ferme », FERME, nom commun, féminin singulier)
 - (« ferme », FERME, adjectif qualificatif, {masculin singulier, féminin singulier})
- Remarque : les lexèmes peuvent être ambigus.

Les dictionnaires

- Dictionnaires de lexèmes uniquement : thesaurii lexicographiques
 - † FERMER : verbe
 - † FERME-1 : nom commun
 - † FERME-2 ; adjectif qualificatif
- Dictionnaires de formes fléchies : toutes les entrées lexicales de type (X, L, C, {F_x})

- Dictionnaires sémantiques de formes fléchies:

- ◆ on ajoute le sens du mot pour augmenter la discrimination
 - (« ferme », FERMER, verbe, { 1ère personne du singulier, 3ème personne du singulier}, *FERMER)
 - ici, on met un pointeur sur la forme infinitive fermer, qui va elle, porter le ou les sens.
 - (« ferme », FERME-1, nom commun, féminin singulier, *bâtiment agricole*)
 - (« ferme », FERME-1b, nom commun, féminin singulier, *poutre de toit*)
 - etc.

Comment réaliser la lemmatisation

- Pour chaque ut U d'un texte T
- Si on a un dictionnaire de formes fléchies de type (X, L, C, {F_x}) alors
 - ◆ appairer U et X
 - ◆ Récupérer toutes les sous-listes (L, C, {F_x}) correspondantes.

Qualité de la lemmatisation

- La qualité de la lemmatisation est l'adéquation réelle entre ce que doit valoir U dans le texte T et la sous-liste (L, C, {F_x}) sélectionnée.
- A priori, plus il existe de listes différentes avec la même tête de liste, plus la qualité de la lemmatisation est mauvaise. Il faut donc désambigüiser.

Techniques de désambigüisation

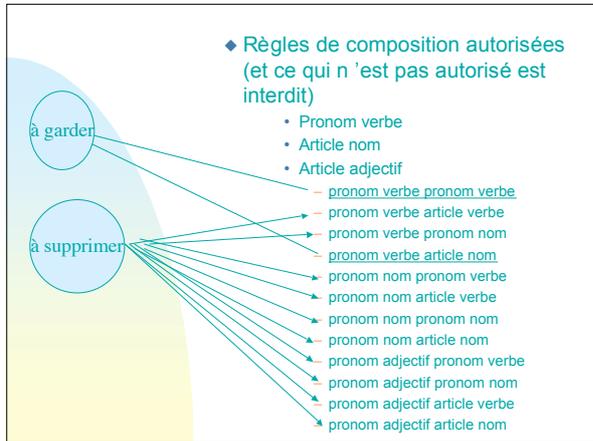
- Par l'analyse syntaxique
- Par apprentissage sur un corpus
- On reste dans l'hypothèse d'un dictionnaire de formes fléchies

Désambigüisation par l'analyse syntaxique

- ◆ Tous types d'analyse depuis l'adjonction de quelques règles de syntaxe jusqu'à l'analyse complète.
 - ◆ Présentation de règles d'interdiction
 - un article ne peut pas être suivi d'un verbe
 - pronom verbe article verbe
 - pronom nom article verbe
 - pronom adjectif article verbe

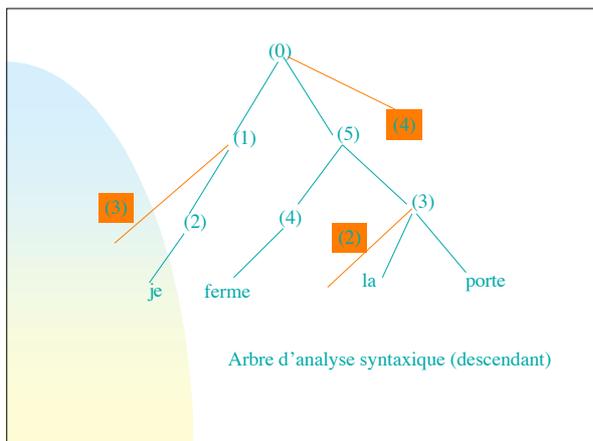
Je ferme la porte

à supprimer



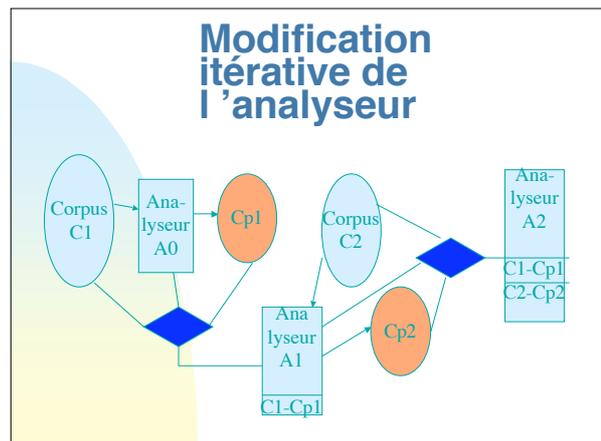
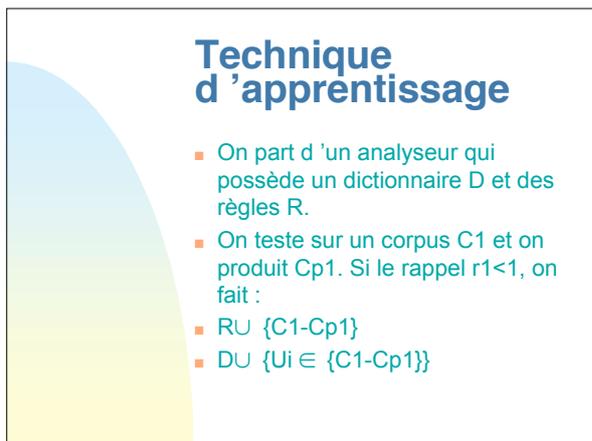
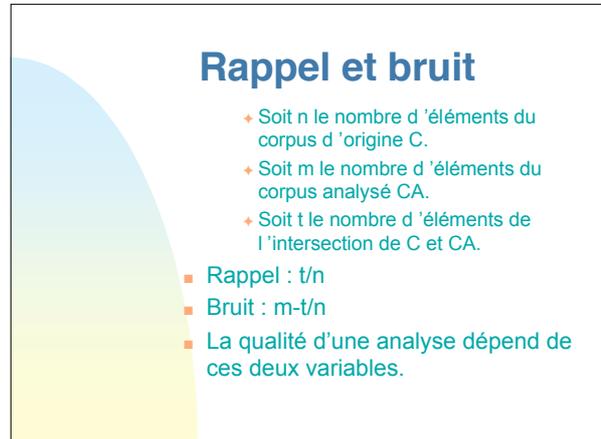
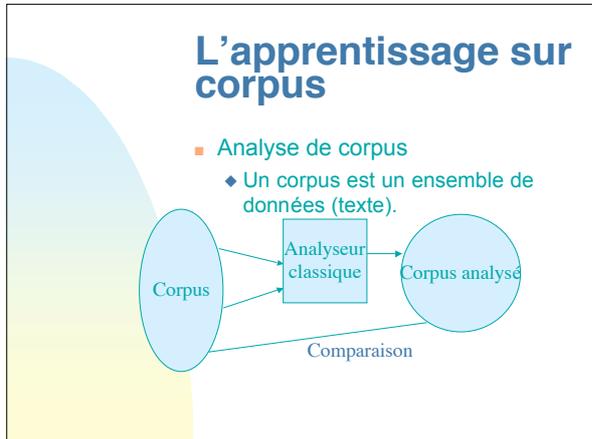
Utilisation des Grammaires

- ✦ (0) proposition -> groupe sujet
groupe verbal
- ✦ (1) groupe sujet -> groupe nominal
- ✦ (2) groupe nominal -> pronom
- ✦ (3) groupe nominal -> article nom
- ✦ (4) groupe verbal -> verbe
- ✦ (5) groupe verbal -> verbe groupe nominal



Les problèmes

- Le langage naturel n'est pas indépendant du contexte sur le plan grammatical
- Les grammaires de la langue ne sont pas complètes
- Les textes peuvent être a-grammaticaux



Problèmes et limites

- Problèmes
 - ◆ Compatibilité des ajouts ?
 - ◆ Non redondance ?
 - ◆ Mécanismes d'abstraction non directement prévus
 - ◆ Données incomplètes en lemmatisation
- Limites
 - ◆ Le bruit n'est pas géré.

Éléments de solution

- Problèmes
 - Vérifications manuelles (PennTree), réduction de l'absurdité
 - redondance par génération ou identité : suppression
 - Mécanismes d'abstraction: « raisonnement »
 - Etiquetage plutôt que lemmatisation.

A voir en option

- Analyseur lexical de Pitrat
 - ◆ Un thésaurus et des règles de conjugaison
- Etiqueteurs
 - ◆ dictionnaires des formes fléchies simplifiés
 - A apprentissage sur corpus d'entraînement : Brill, PennTree
 - Grammaticaux simples (markoviens, ATN, automates, etc.)
- Analyseurs morphosyntaxiques