Equipe Informatique Linguistique du LIGM

Journée des doctorants - ED MSTIC 2010

Myriam Rakho et Anthony Sigogne

10 *juin* 2010

Table des matières

Informatique Linguistique et applications

- Traitement automatique des langues (TAL)
- Créer des processus automatique permettant de traiter les langues naturelles
- On utilise tous les logiciels du TAL !!
  - Traduction automatique (google traduction, systran,...)
  - Correcteur orthographique (intégré à Word,...)
  - Résumé automatique de texte
  - Classification de documents (Yahoo, Google,...)
  - Moteurs de recherche (Yahoo, Google,...)
  - ...

Equipe Informatique Linguistique du LIGM

- Université Paris-Est Marne-la-Vallée
- Équipe dirigée par Eric Laporte :
  - 6 chercheurs permanents
  - 2 ingénieurs de recherche
  - 12 doctorants
  - 18 membres associés
- Équipe composée de linguistes et d'informaticiens
- http://infolingu.univ-mlv.fr/

Deux axes de recherche majeurs dans l'équipe

- Développement de processus automatiques traitant les langues naturelles
- Développement manuel de ressources lexicales et syntaxiques
  - Basées sur la langue naturelle, créées par des linguistes
  - Intégration dans les processus automatiques

Thèmes particuliers

- **Analyse syntaxique**
- **Annotation Sémantique**
- Traduction automatique
- Développement de ressources lexicales et syntaxiques
- Enrichissement de corpus annotés (Europarl, ...)
- Création de logiciels de traitements linguistiques (Unitex, Outilex, ...)

Présentation succinte des thèses

- Annotation sémantique
- Intégration d'un lexique syntaxique dans un analyseur syntaxique probabiliste

Annotation sémantique de documents

Étapes :

1. La problématique : l'accès aux masses de données
2. Mon sujet de thèse : l'annotation sémantique de documents

Analyse des grandes masses de données

Exploitation optimale des masses de données textuelles

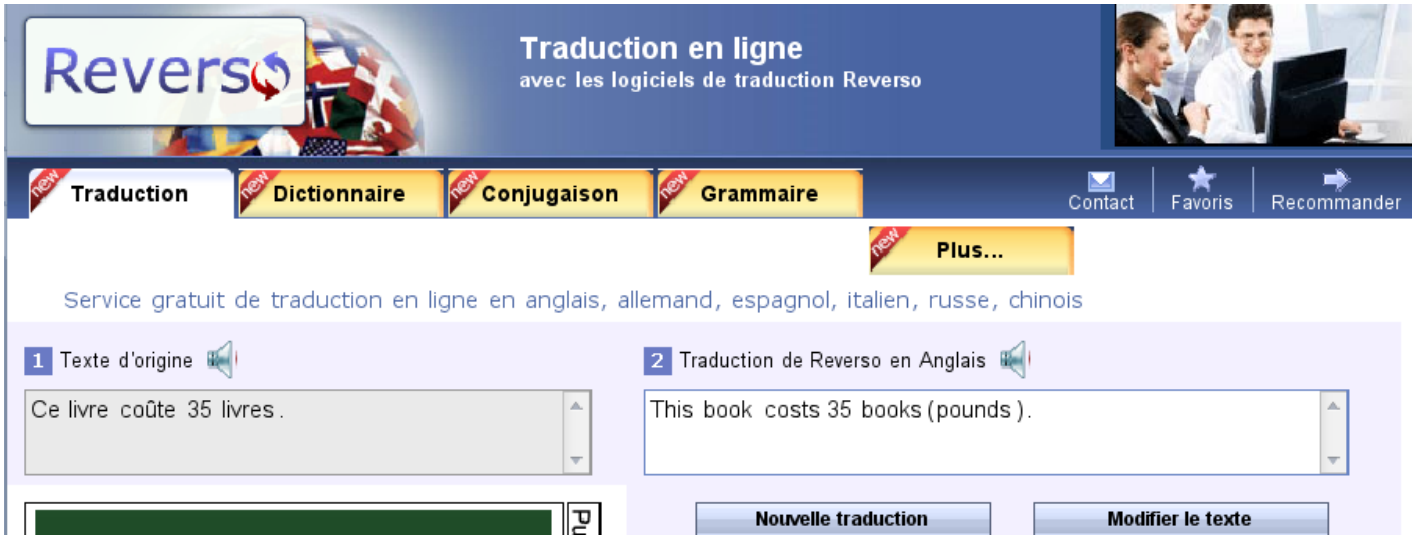
Multilinguisme => traduction automatique

Ambiguïté du langage => identification du sens des mots à traduire

Exemple :

Livre. *nom féminin ou masculin*

1. ouvrage (en. *book, writing, ...*)
2. monnaie (en. *pound sterling*)
3. unité de mesure de poids financier (en. *pound*)



Annotation sémantique de documents

Définition : (*Cross-lingual Word Sense Disambiguation*)

Procédure visant à identifier la signification d'une unité lexicale ambigüe :

-> étant donnée une liste prédéfinie des différentes significations d'un mot

Article. *nom masculin*
1. article de presse *en.* press report
2. partie de texte de lois *en.* rule, article

-> identifier son sens dans un contexte donné

*L'**article** sur les choux du futur ...*
*L'**article** 144 du Code Civil ...*

*L'enfant **joue avec** un ballon ... => - avec qqc.*
*On **a joué au** Scrabble. => - à qqc.*
*Max **joue du** piano. => - de qqc. (concret)*
*Marie **joue de** son charme pour passer. => - de qqc. (abstrait)*
*Les musiciens **jouent** une sonate de Mozart. => - qqc.*
*Luc **a joué** 15 euros sur cette course. => - qqc. sur qqc.*

Annotation sémantique de documents

Statistical Language Modeling (SLM) :

Modèles de représentation linguistique et statistique des données

Étape 1 :

Enrichir les données d'informations linguistiques pour en affiner le contenu

Exemple : représentation syntaxique du verbe *to end*

She ended their relationship after just two months. Elle a mis fin à leur relation après deux mois.
The concert ended with a Mozart violin concerto. Le concert s'est terminé par un concerto ...
The rain ended your plans to play tennis. La pluie a ruiné vos plans de jouer au tennis.
This contract is ended. Ce contrat a expiré.
The enterprise ended her contract. L'entreprise a annulé son contrat.

someone end sth.	=	mettre fin à qqc.
someone sth. end sth.	=	mettre fin à qqc. ; ruiner
someone end sth.with sth.	=	terminer qqc. par
sth. end with/in sth.	=	se terminer par
contract/agreement end	=	expirer
end a contract/agreement	=	annuler, résilier

Annotation sémantique de documents

Étape 2 :

Description statistique des données

Exemple : représentation statistique des relations de cooccurrence entre :

- les cooccurrents de *to end* (lignes)
- et les traductions de *to end* (lignes)

	contract.Subj	contract.Obj	relation.Obj	plan.Obj	concert.Suj	concert.Obj
expirer	45	0	0	0	0	0
annuler	0	6	0	0	0	0
mettre fin à	0	0	67	0	0	0
ruiner	0	0	15	54	0	0
se terminer par	0	0	21	0	33	0
finir	...	...	...	...	...	...

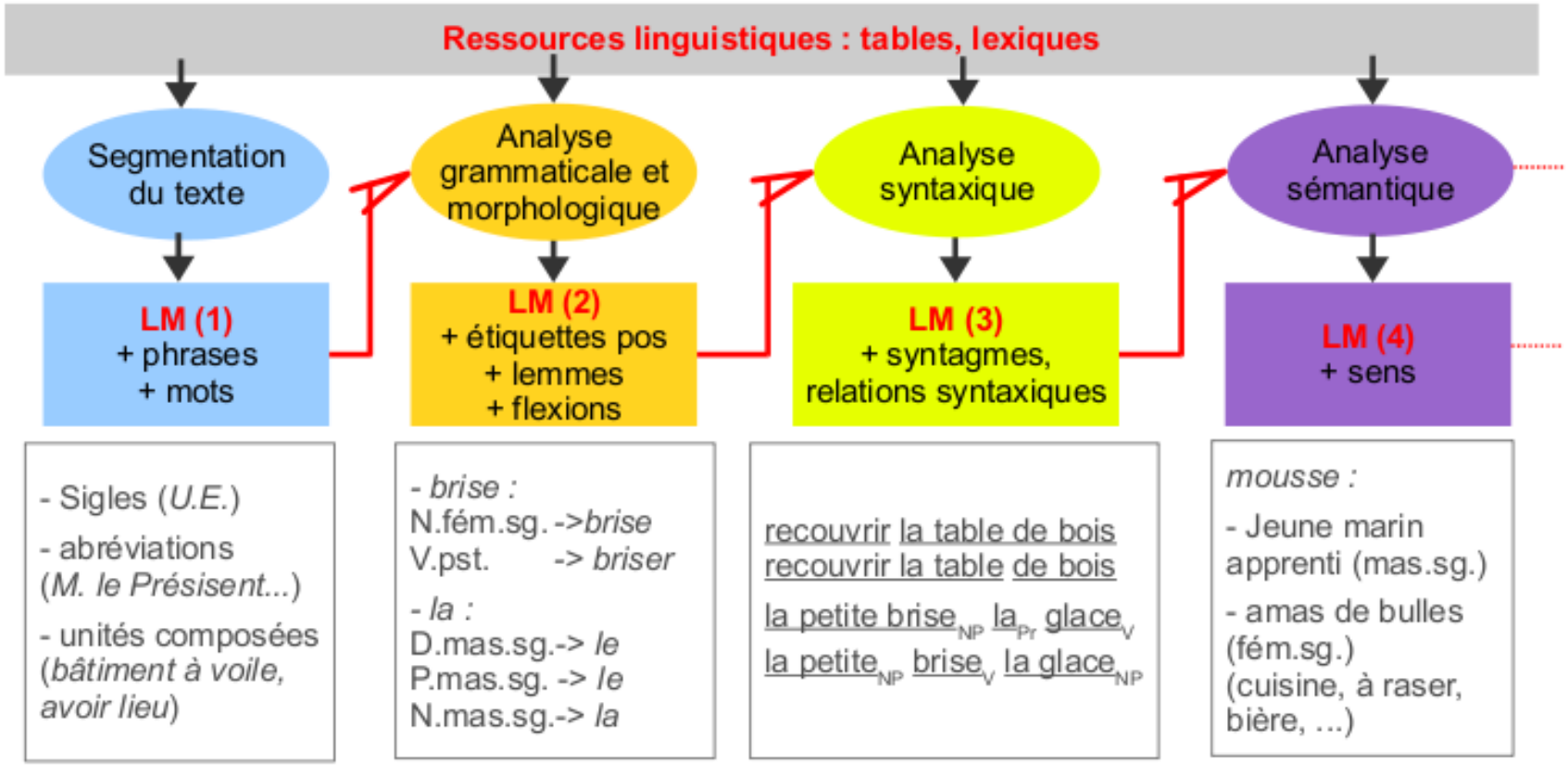
Explication : Lorsque *end* occurre avec les mots en entrée des colonnes,
il est traduit par les mots en entrée des lignes.

Cooccurrents = termes qui apparaissent ensembles dans un même contexte.

Annotation sémantique de documents

Problème : l'ambiguïté inhérente au langage (qui fait sa richesse !)

LM (*i*) = modélisation linguistique et statistique enrichie à chaque étape
Le modèle construit à l'étape ***i+1*** utilise des informations issues du modèle de l'étape ***i***.



Différents niveaux d'analyse, différents niveaux d'ambiguïté

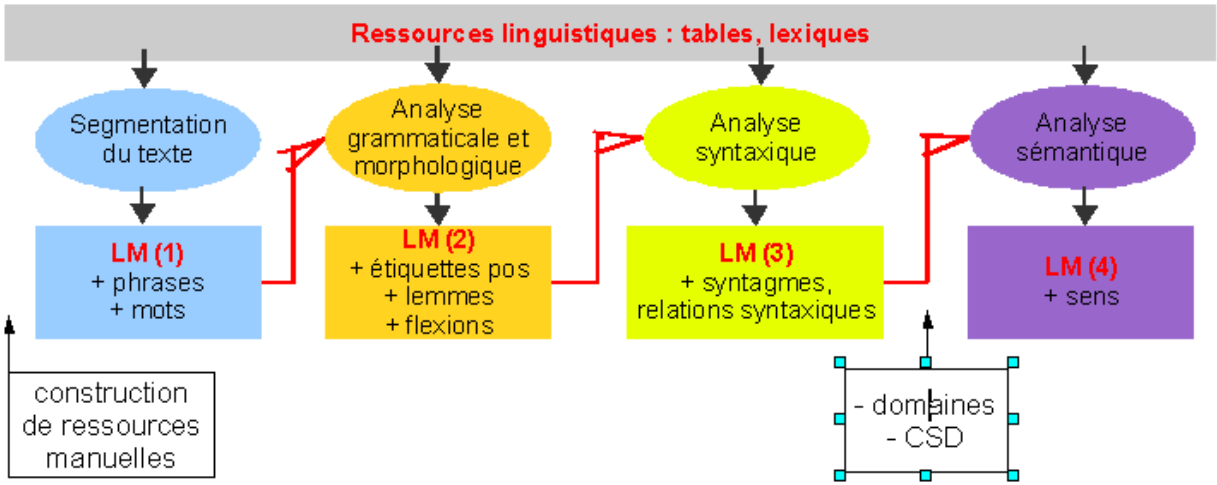
Annotation sémantique de documents (3/4)

Mon travail : Paramétrage de la modélisation des données :

*** du point de vue linguistique :**

enrichir le modèle avec des informations supplémentaires

- expressions multi-mots
- sur le domaine : Wordnet (chimique, médical, juridique, ...)
- sur les cadres de sous-catégorisatino : les tables du LADL
- etc.



*** du point de vue statistique :**

- tester des algorithmes nouveaux
- combiner différents algorithmes en fonctions des catégories de mots

et donc des modèles linguistiques

Annotation sémantique de documents

Objectif :

* Trouver le modèle optimal de représentation des contextes d'un mot

  = la combinaison d'informations linguistiques-statistiques la plus pertinente
  = le modèle qui permettra au mieux de distinguer ses différents sens

* en fonction de sa catégorie grammaticale, son type de polysémie, ...

Annotation sémantique de documents

Exemples d'informations linguistiques utiles pour la WSD :

type de polysémie	exemple	informations linguistiques utiles pour distinguer les différentes acceptions
homonymie de genre	le mousse vs. la mousse	grammaticales : catégorie Nom (vs. verbe *mousser*) + morphologiques : genre masc. ou fém.
polysémie	bière mousse à raser végétal	+ syntaxe (arguments) + champs thématiques + domaine (boissons, cosmétiques, botanique)
polysémie	*jouer*	+ cadre de sous-catégorisation - avec qqn./qqc. (jouer avec un ballon) - à qqc. (Monopoly, Scrabble, ballon, poupée) - de qqc. - concret (piano, violon, flûte) - de qqc. - abstrait (charme, influence) - qqc. (air, sonate, morceau) - qqc. sur qqc. (jouer 15 euros sur une course)
polysémie	faible	point - ; homme - ; - distance ; sexe - ; caractère - se sentir ; jambes ; ...

Intégration d'un lexique syntaxique dans un analyseur syntaxique probabiliste

Deux étapes :

- Créer un analyseur syntaxique probabiliste pour le français
  - Déterminer automatiquement le ou les sens de la phrase
- Intégrer des ressources syntaxiques dans le processus automatique
  - ressources créées manuellement, plus pertinentes que les statistiques

Analyse syntaxique ?

- Comment les mots se combinent en phrases qui ont un sens ?
- Dans quel ordre ?
- Quelles sont les possibilités de combinaison ?

Exemple

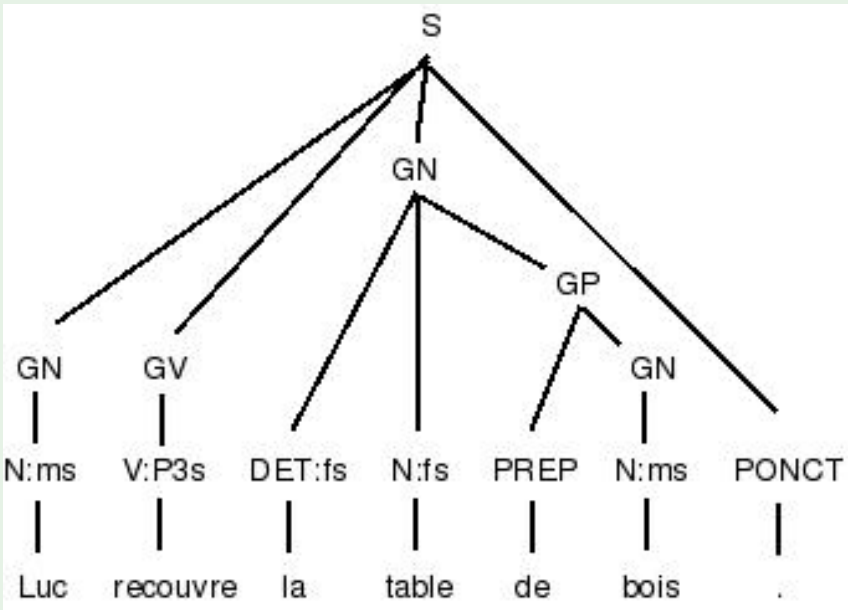
Luc recouvre la table de bois.
* recouvre la table Luc de bois

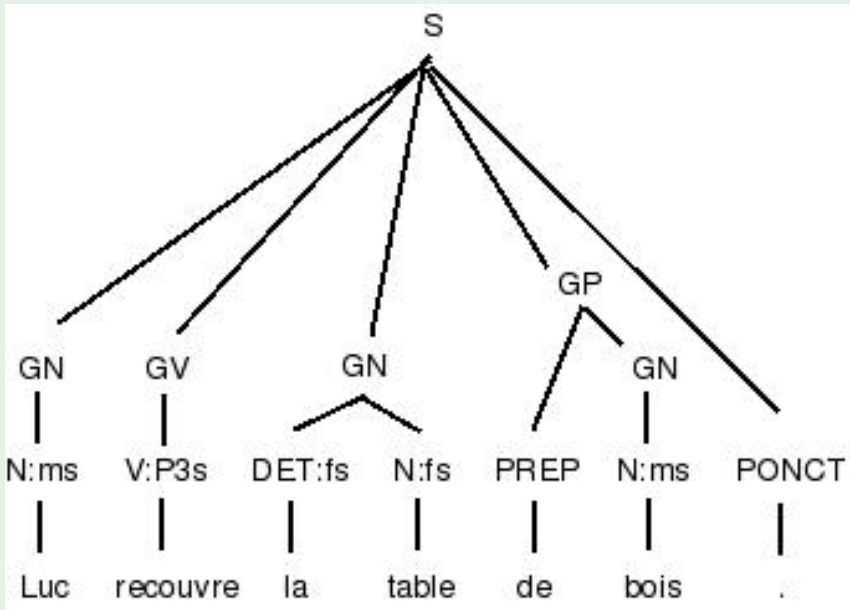
Analyse syntaxique d'une phrase

- Déterminer les constituants syntaxiques (Groupes nominaux, verbaux,...)
- Déterminer les relations entre ces constituants

Exemple

Luc recouvre la table de bois.





Analyse syntaxique probabiliste

- Processus automatique de l'analyse syntaxique
- Utilisation d'un modèle probabiliste (Modèles de Markov,...)
- Entraînement sur un corpus annoté

Exemple

	règles+fréquences		**Probabilités associées**
	GP -> PREP GN : 100		GP -> PREP GN : 0.01
corpus $\Longrightarrow$	GN -> N:ms : 700	$\Longrightarrow$	GN -> N:ms : 0.12
	GN -> DET:fs N:fs GP : 2501		GN -> DET:fs N:fs GP : 0.50

Problématiques et Innovations

1ère Problématique :

- La recherche sur l'analyse syntaxique probabiliste n'est pas récente, cependant :
  - Corpus anglais, Penn-Treebank : première version 1993
  - Corpus français, French-Treebank : première version 2001
- Peu d'analyseurs et résultats moyens dans le cadre du français.

Solutions :

- Méthodes d'analyses récentes.
- Modèles probabilistes complexes récents (CRF, Maximum entropy, SVM...).

Problématiques et Innovations

2ème Problématique :

- La plupart des analyseurs probabilistes ne se basent que sur les statistiques.

Solution :

- On dispose de nombreuses ressources syntaxiques pour le français.
  - Lexique-Grammaire (Maurice Gross)
  - Pour chaque entrée du lexique, on dispose des structures syntaxiques possibles.

Problème :

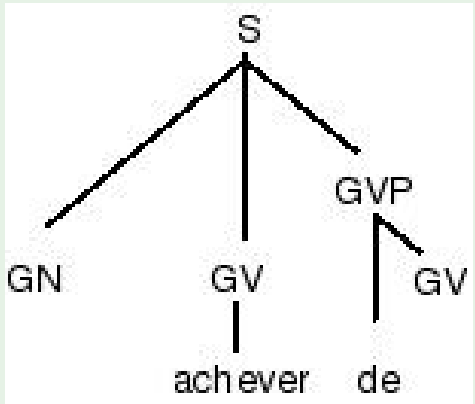
- Les données du lexique ne sont pas exploitables directement par l'ordinateur.

Exemple

Entrée : Verbe **achever**

Structure syntaxique : GN **achever** de GV

Transformation sous forme d'arbre syntaxique :



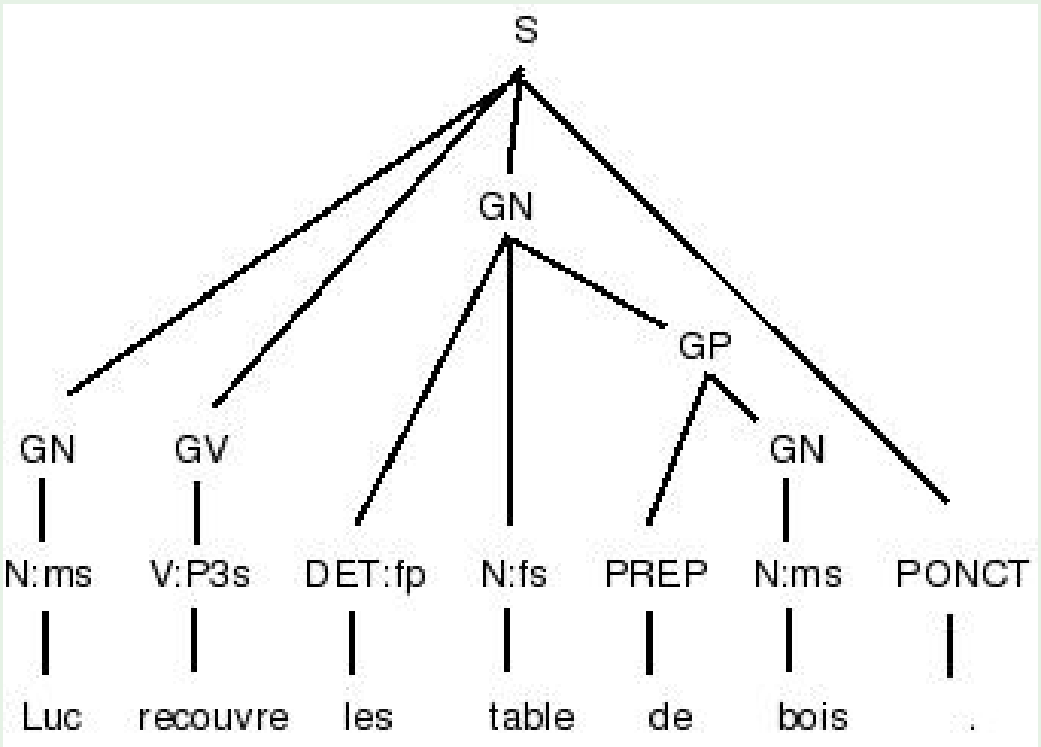
Applications à la correction grammaticale

Un correcteur grammatical fait de multiples vérification :

- Conformité des mots aux règles grammaticales (ordre des mots et accords).
- Présence des mots dans les dictionnaires lexicaux.

Exemple

* Luc recouvre les table de bois.



FIN

Merci pour votre attention!