Chaîne de traitements linguistiques pour l'acquisition automatique de lexique

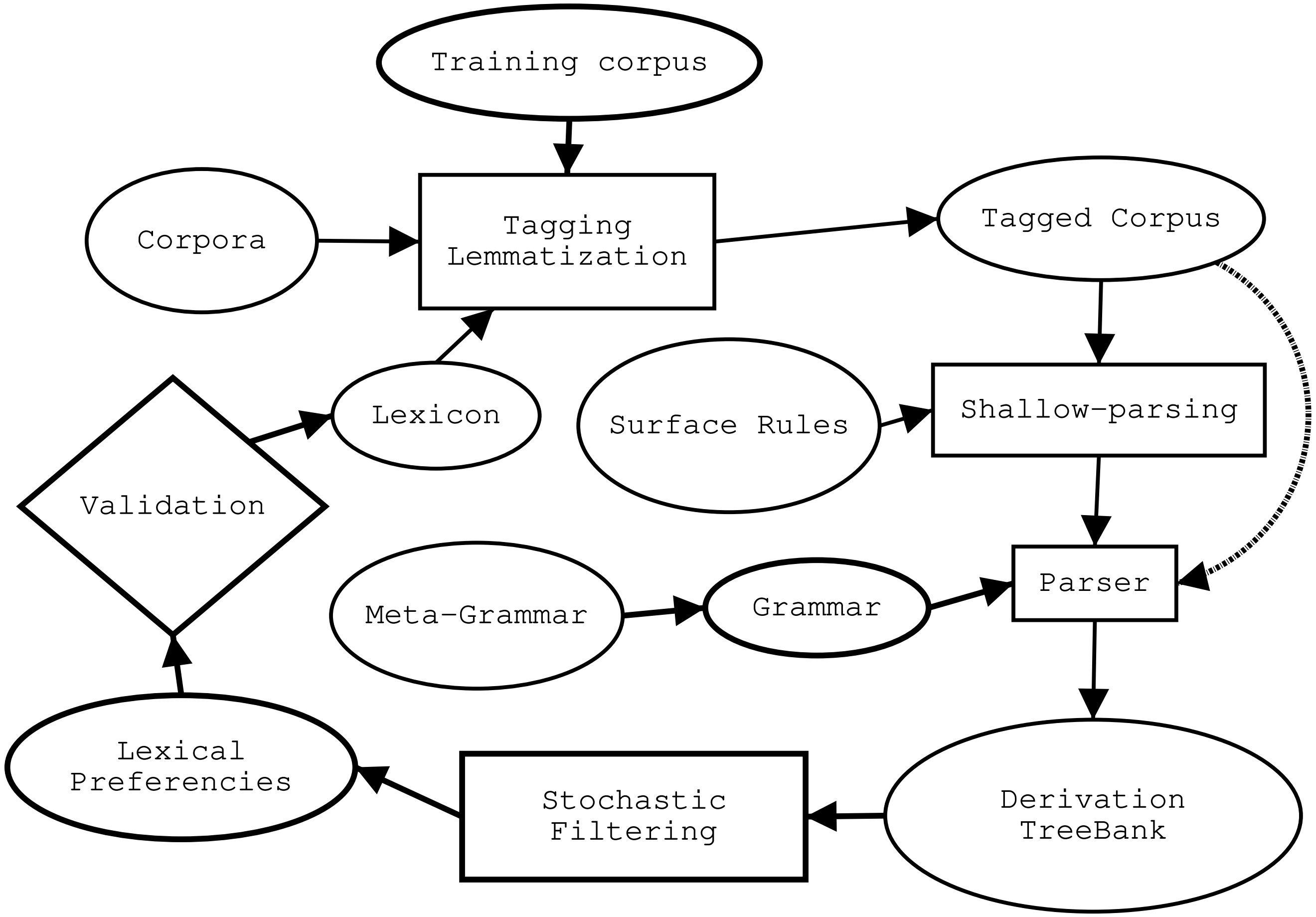
Lionel Clément

Projet Atoll

INRIA Rocquencourt

Lionel.Clement@inria.fr

Journée RLT – le 12 décembre 2001


Training corpus
Corpora
Tagging
Lemmatization
Tagged Corpus
Lexicon
Surface Rules
Shallow-parsing
Validation
Meta-Grammar
Grammar
Parser
Lexical
Preferencies
Stochastic
Filtering
Derivation
TreeBank

Corpus

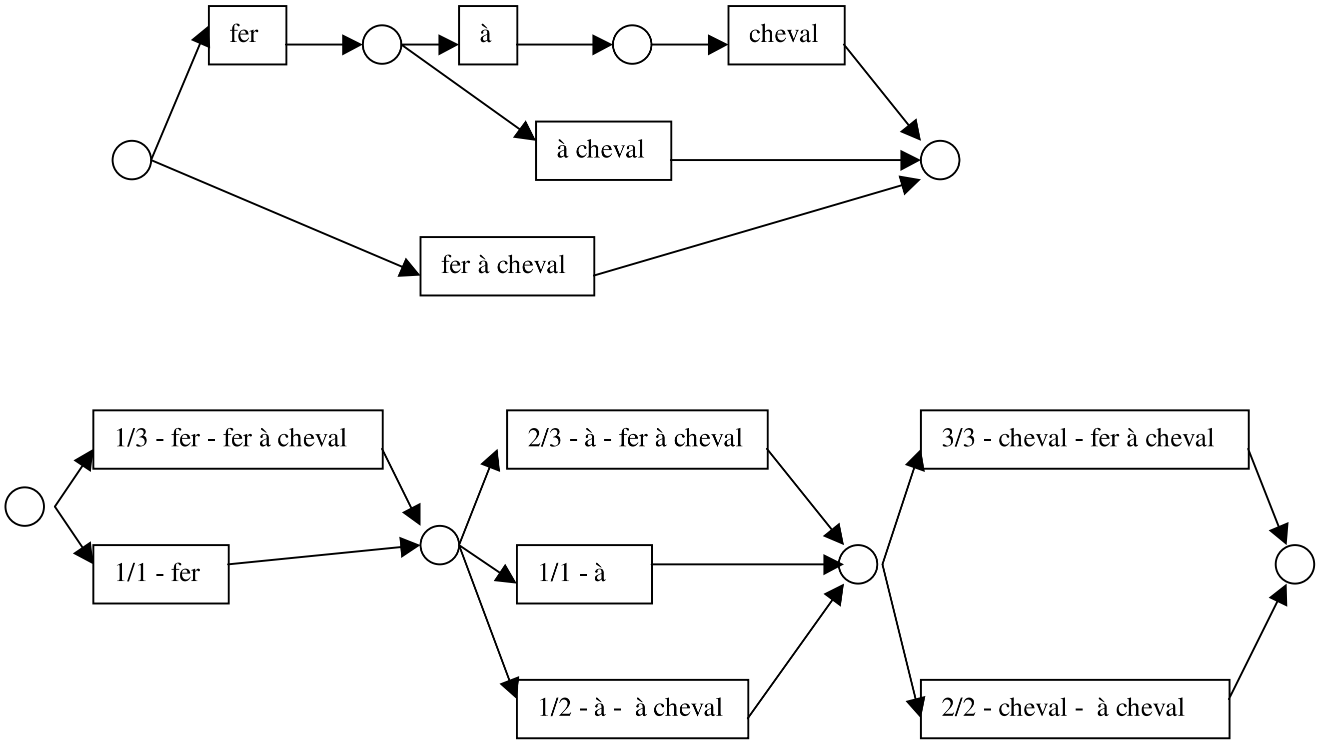
- Le Monde
- Bibliothèque Nationale de France
- Corpus de presse ?
- Test Suites (TSNLP – Eurotra)
- Corpus Botanique

Segmentation en mots

- LT XML toolkit (W3C)
  - Segmentation en chaînes **sgmltoken** (*aujourd / ' / c / ' / - / est / - / à / - / dire*)
  - Segmentation en tokens **sgmlseg** (*aujourd' / c' / - / est / - / à / - / dire*)
  - **à faire**→ Segmentation en mots (dont mots composés) grâce à des DAG (DTD)
    - amalgames (mots *porte-manteaux*) *desquels, aux*, etc.
    - discontinuités *afin de*
    - ambiguïté hors contexte mots simples/mots composés

Segmentation en mots

- Segmenteur (*tokenizer*) développé en **Flex**
  - Éclater les mots amalgamés
  - Segmenter les *tokens*
  - **à faire**→ interface avec **Lexed** et lexique mots composés : regrouper les mots composés


fer
à
cheval
à cheval
fer à cheval
1/3 - fer - fer à cheval
1/1 - fer
2/3 - à - fer à cheval
1/1 - à
1/2 - à - à cheval
3/3 - cheval - fer à cheval
2/2 - cheval - à cheval

Segmentation – DTD

- Proposition consortium Genelex
- Utilisation d'une DTD XML pour représenter des DAG
- Impératifs :
  1. Le seul contenu des balises PCDATA est le texte source
  2. Possibilité d'exprimer des ambiguïtés (mots simples/mots composés, *tagging*)

Étiquetage

- Pourquoi réentraîner ?
  - Nouveau corpus de référence le Monde
  - Conventions différentes
  - Étiquettes riches
  - Corpus d'entraînement : 749 860 mots – nbre d'étiquettes : 208
  - Corpus d'évaluation : 118 618 mots
- Problèmes
  - Les outils sont souvent conçus pour un jeu donné (**Spak**, **Flemm**, etc.)

Étiquetage

- Tagger TnT (Thosten Brants)
  - Problème Licence
  - (3-grams)
  - 6 219 mots mal taggés soit (5,23 %)
- TreeTagger
  - arbre de décision
  - 3 872 mots mal taggés soit (3,25 %)
- Brill Tagger
  - Pas essayé

Mots les plus souvent mal taggés

TreeTagger		TnT	
31	que C-S-	38	qui PRO-rel-3fp
34	en P–	39	le D-def-fp
35	l’ D-def-ms	40	qui PRO-rel-3ms
36	qui PRO-rel-3fp	44	que ADV–
37	se CL-refl-3fs	45	les D-def-fp
38	que ADV–	46	l’ D-def-ms
39	qui PRO-rel-3ms	46	se CL-refl-3fs
40	comme P–	72	qui PRO-rel-3fs
49	le D-def-mp	82	le D-def-mp
69	qui PRO-rel-3fs	185	% N-C-mp

Lemmatisation

- **TreeTagger**
  - Pas d'ambiguïté possible ?
  - Pas de traitement des mots inconnus
- **Flemm** (F. Namer)
  - Ne fonctionne qu'avec conventions **TreeTagger** ou **Brill Tagger**
  - Utile que pour mots inconnus
  - → portage du code pour nos conventions (encodage des données en dur)
- **Lexed** + Lexique dico morpho
  - Traitement des mots inconnus à faire
    - Interface avec **Flemm** ?
    - Application de règles régulières (suffixe + tag → lemme)

Shallow parsing

- **Spak** Modifications du code en concertation avec A. Kinyon :
  - Ajouts :
    - actions (insert(), closeWhenClose(X, Y), closeWhenOpen(X, Y), closeUntil(X))
    - conditions (lemme, joker)
    - entrée/sortie (unicode, chaîne)
  - Produit des balisages valides
  - À l'étude :
    - compléter tests sur constituance
    - autres tests sur état de l'analyseur ?
- problèmes non résolus :
  - encodage
  - optimisation

```
////////////////////////////////////////////////////////////
// R9 (436):
// If current is quant adv
//   and next is prep "de",
// Then R9 bis (440):
//   if not inside a PP
//     or previous is a relative pronoun,
//   then close current constituent.
//   start <NP> and insert current.
//
{!PP}  (:$quantAdv:) (:"de":$Prep) then
close(), open(NP), insert();

{}  P(::$relPro) (:$quantAdv:) (:"de":$Prep) then
close(), open(NP), insert();
```

Alternatives Chunking et/ou shallow-parsing

- Chunking : Reconnaîssance balises de constituants en regexp (pas essayé)
- Shallow-parsing : **Spak** : Reconnaîssance balises de constituants grâce à un **PDA**
- Shallow-parsing : RCG Robuste temps linéaire (P. Boullier)
  - TAG $\rightarrow$ RCG (P. Boullier 98 in TAG+4)
  - RCG $\rightarrow$ 1-RCG
  - RCG $\rightarrow$ RCG' augmentée tel que RCG' reconnaisse V* en plus de L-RCG

Analyse – Analyse partielle

- Parser RCG (mais traits et co-ancres)
  - À l'étude : Parser DRCG (Definite Range Concatenation Grammar)
    $\psi \to \psi_1 \ldots \psi_m$
    $\psi_i = A(\alpha_1, \ldots, \alpha_p)(\theta_1, \ldots, \theta_q)$
- RCG + parser DyALog
- Tests sur parser d'Anoop

Interface des différents composants

- Encodage XML
  - Segmentation des mots
  - Annotation morpho-syntaxique
  - Annotation des constituants
  - Grammaire
  - Forêts d'analyse
- XSL – XSLT
  - Interface avec outils non XML (**TreeTagger, Flemm, Spak**)
- **Cocoon servlet WWW + XSLT**
  - Serveur de lexique, de grammaire (A. Khajour), de forêts d'analyse (V. Bhandari) via une DB et un langage de requêtes, de parsers.

Filtrage

- Sources
  - Textes étiquetés et lemmatisés
  - Textes segmentés
  - Textes shallow-parsés
  - Textes parsés
- Calculs stastistiques
  - Représentativité d'une occurence (segment, lemme, forme, fonction, constituant).
  - Représentativité des collocations (de termes ou terme/construction)

Filtrage

- Cibles
  - Formes – Lemmes
    * termes, noms propres, idiotismes, mots *inconnus*
    * traits (restriction de sélection)
  - Schèmes
    * co-ancrage (figements, sous-cat)
    * traits (restriction de sélection)
  - Familles
    * couverture des instances d'une même famille