ATOLL - INRIA Rocquencourt

http://atoll.inria.fr

Quel bilan pour l'ARC RLT?

Éric de la Clergerie

Eric.De_La_Clergerie@inria.fr

Réunion RLT

Jussieu – Jeudi 12 Décembre 2002

Rappel du thème de RLT

Examiner les problèmes d'acquisition et de représentation de ressources dans le contexte des TAG.

Rappel du programme proposé

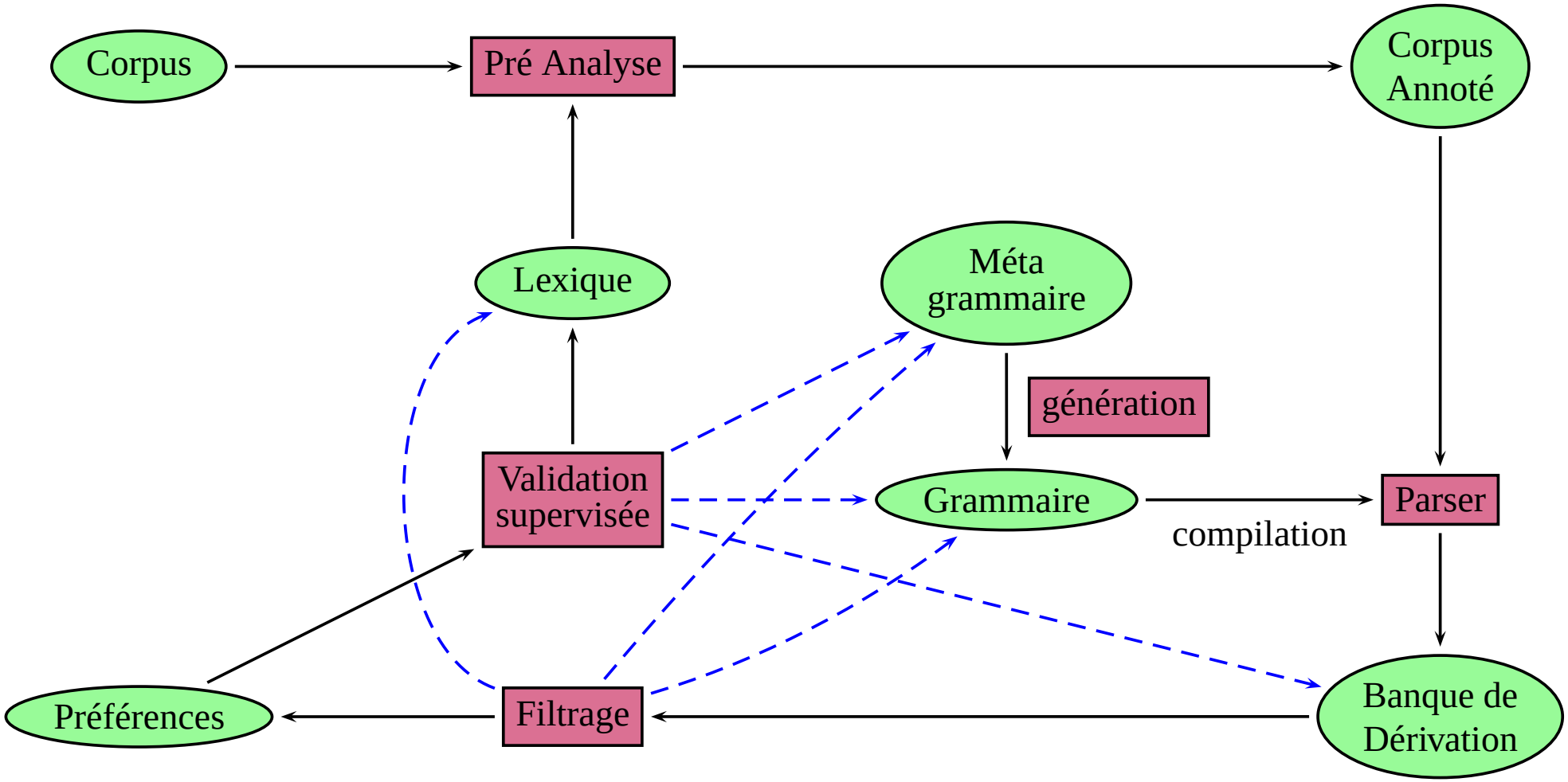
- Constitution ou sélection d'un corpus de textes
- Recensement les outils nécessaires et ceux disponibles
- Etude de rôle de la meta-grammaire
- Définition des représentations
- Définition du processus de constitution des corpus annotés
- Définition du processus d'émergence des ressources candidates
- Définition de l'interface de validation des ressources

Rappel des objectifs attendus

- des représentations normalisées pour les différents types de ressources manipulées (meta-grammaire, grammaires, corpus, lexiques)
- L'amélioration des techniques d'analyse syntaxique pour les TAG
- Une meilleure compréhension des meta-grammaires pour les TAG
- une méthodologie pour l'acquisition supervisée de ressources linguistiques, si possible fondée sur l'emploi d'une meta-grammaire
- un prototype de chaîne d'acquisition
- l'acquisition d'un premier jeu de ressources pour valider

Note : le but de l'action est avant tout de dégager une méthologie et des prototypes pouvant être adaptés (autres langues et formalismes).

Proposition de chaîne d'acquisition



Bilan : points faibles

- Flou entre TAGML1 et TAGML2 $\Rightarrow$ retard dans la mise à jour des outils et ressources
- Retard dans la conversion FTAG vers TAGML (stabilisation FTAG et outils)
- Pb dans la compilation de parseurs pour des grammaires de la taille de FTAG
- Pb technologiques sur Bases de données XML (banque de dérivations)
- Module de calcul d'information distributionnelle incomplet
- Interface de saisie/validation non terminée
- Liens lexique-métagrammaire peu étudiés
- Liens peu explicites entre MG et FTAG
- Pb de synchro entre les différentes étapes de la chaîne
- Pas de réelle validation de l'ensemble

Réalisations (1)

- Normalisation de ressources linguistiques
  - Extension TAGML2
  - Diffusion technologie XML, dont techniques de conversion (scripts,XSL)
  - format XML pour divers ressources (TAG, Forêts, morpho-syntaxe, méta-grammaires, ...)

  $\Rightarrow$ positionnement pour action Normalangue

- Meta-grammaires
  - Développement d'outils pour les MG (éditeur, compilateur)
    utilisation de ces outils
  - Développement de nouvelles MG et grammaires générées (français, espagnol, allemand)
  - Extensions MG de FTAG pour Noms et Adjectifs

  $\Rightarrow$ proposition d'ARC MGL

- Evolution technologie des parseurs
  - Parseur RCG pour les TAG (Pierre Boullier),
    efficace et valide pour de grosses grammaires (2-5Karbres)
  - Parseur DyALog pour les TAG, testé sans problème jusqu'à 500 arbres
  - Parseur LORIA et test SOAP
  - Test chaîne MG -> Grammaire -> Parseur pour DyALog, LFG et RCG

Réalisations (2)

Infrastructure de la chaîne de traitement :

- développement d'outils graphiques de visualisation d'arbres et graphes
- Mise au point de la chaîne morpho-syntaxe (Lionel)
  - fondée sur pipeline XML
  - tests avec parseur DyALog (petite grammaire du français) dans serveur de parseurs
- Tests de constitution de banque de dérivations à partir de corpus (Faycal Chami)
  petite grammaire du français hors chaîne
- Module de calcul d'info distributionnelle à partir de banques de dérivation (Faycal Chami)
  - fréquence de construction (famille) par lemme
  - goulet d'étranglement : importance d'une famille dans un exemple
  - ?fréquence de collocation (co-ancrage)
  - ?pas de constructions disponibles (symptome de grammaire incomplète)

Réalisations (3)

Suite infrastructure

- Définition de l'interface de saisie (José Aguirre-Ruiz)
  première mouture non opérationnelle, mais bon cahier des charges
  - sélection lemmes/familles (1er écran)
  - écran linguistique pour qualifier une « famille » (?liens avec MG)
    possibilité d'édition
  - écran statistique (affichage/synthèse info distributionnelle)
  - écran exemples tirés du corpus (banque de dérivations)

Volet « Social »

- Renforcement des liens entre nos équipes
- Travail collaboratif, échange d'information

Outils et ressources (1)

- Editeur MG et variantes (Bertrand, Benoît, Kim, Lionel)
- Compilateur MG et variantes (Bertrand, Benoît, Kim, Lionel)
- Chaine morpho-syntaxe (Lionel)
  tokenizer, lexed, Tree Tagger réentrainé, wrapper/unwrapper XML pour pipeline
- Parseur LORIA Patrice (Azim,Amalia), adapté TAGML2
- Parseurs ATOLL (DyALog, RCG ?)
- outils de débogage d'une grammaire LTAG (Azim)
- serveur de parseurs (Atoll)
- serveurs de forêts ou banque de dérivation (Atoll)
  + calcul d'info distributionnelle
- interface de saisie (ATOLL)

Outils et ressources (2)

- Corpus (faire le point)
- Grammaire FTAG (Anne,Barrier) + Documentation FTAG (Barrier)
- Petites MG et grammaires générées
- DTD TAGML2 (version lite) + APIs d'entrée/ sortie et visualisation (Azim)
- DTD Forest, TAGML et TAGML2 (Atoll)
- Outils de conversion TAGML (Atoll, LORIA)
- afficheur d'arbres TAGVIEWER (Benoît Crabbé)
- afficheur de graphes (LORIA)

Rappels des stages

- Stage postdoc de Lionel Clément
- Stage DEA 2001 de Stéphanie Werli – conversion corpus annoté
- Stage Cristophe Coquet (2001) – Représentation et édition de document XML (pour des graphes)
- Stage DEA 2002 de Faycal Chami – Banque de dérivation et module de filtrage
- Stage DESS 2002 de José-Manuel Aguirre-Ruiz – Interface de saisie

+ Stages Vartika Bhandari (serveur de forêts) et Abdelaziz Khajour (serveur de grammaire)

Evolution

Avec un peu de travail, la chaîne peut être complétée. Comment faire ?

- dégager des priorités dans les développements
- interesser d'autres partenaires : publicité et valorisation sur cette expérience
  publications, ouvertures, réponse à des appels d'offre
  mise en place d'un lieu d'accès aux outils et ressources
- trouver une place dans la nouvelle ARC MGL
- modifier les objectifs (+MG, -TAG)