ATOLL - INRIA Rocquencourt

http://atoll.inria.fr

Normalisation de données linguistiques

liens avec SYNTAX?

Éric de la Clergerie

Eric.De_La_Clergerie@inria.fr

Réunion SYNTAX

INRIA Rocquencourt – Mercredi 11 Décembre 2002

ATOLL

Atelier d'Outils Logiciels pour le Language naturel

ATOLL

Atelier d'Outils Logiciels pour le Language naturel

- surtout concerné par le traitement syntaxique
  multiformalismes, multi-niveaux, gestion de ambiguïté

ATOLL

Atelier d'Outils Logiciels pour le Language naturel

- surtout concerné par le traitement syntaxique
  multiformalismes, multi-niveaux, gestion de ambiguïté

- mais implication dans les questions d'infrastructure logicielle
  pour le traitement linguistique

ATOLL

Atelier d'Outils Logiciels pour le Language naturel

- surtout concerné par le traitement syntaxique
  multiformalismes, multi-niveaux, gestion de ambiguïté

- mais implication dans les questions d'infrastructure logicielle
  pour le traitement linguistique

- et validation de nos outils dans
  - acquisition de connaissances (linguistique)
  - fouille de textes

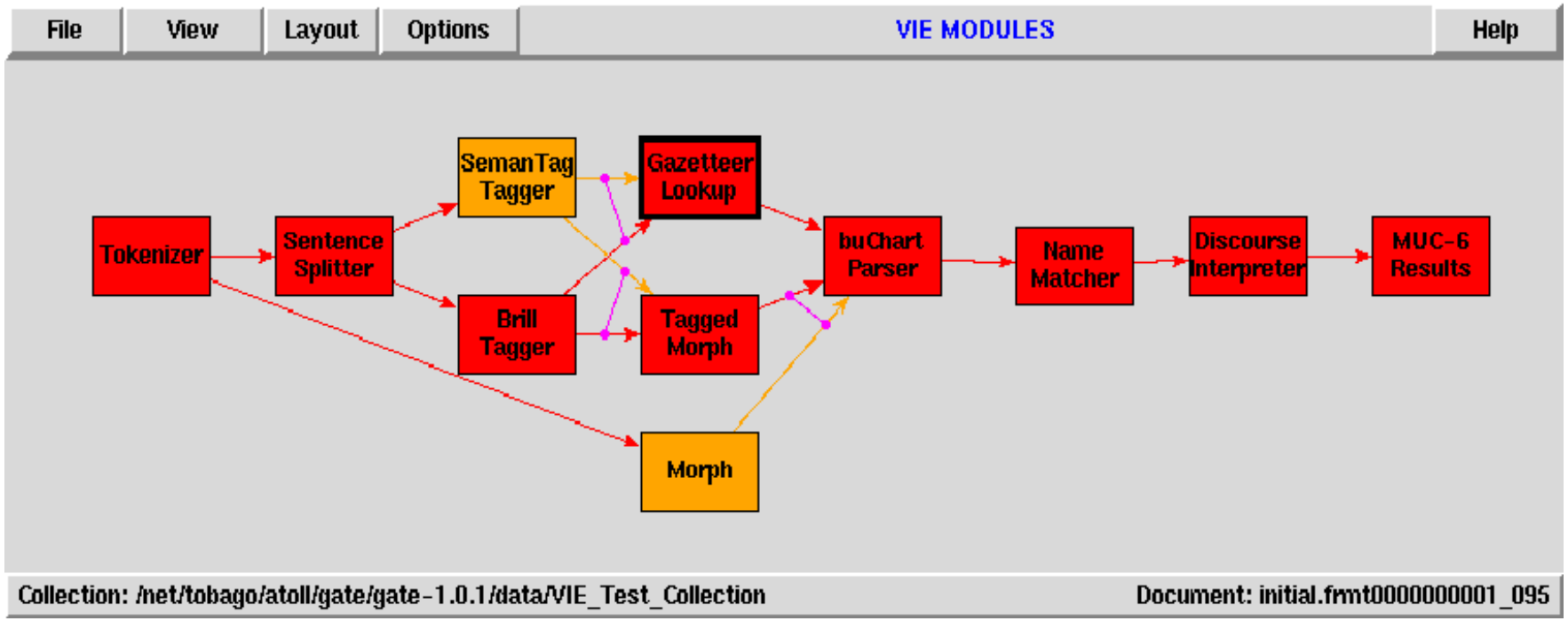
  $\Rightarrow$ traitement de documents (corpora)

Chaîne de traitement linguistique

De nombreux composants et ressources à mettre en jeu, dépendant de la tâche, du domaine et de la langue !

Chaîne de traitement linguistique

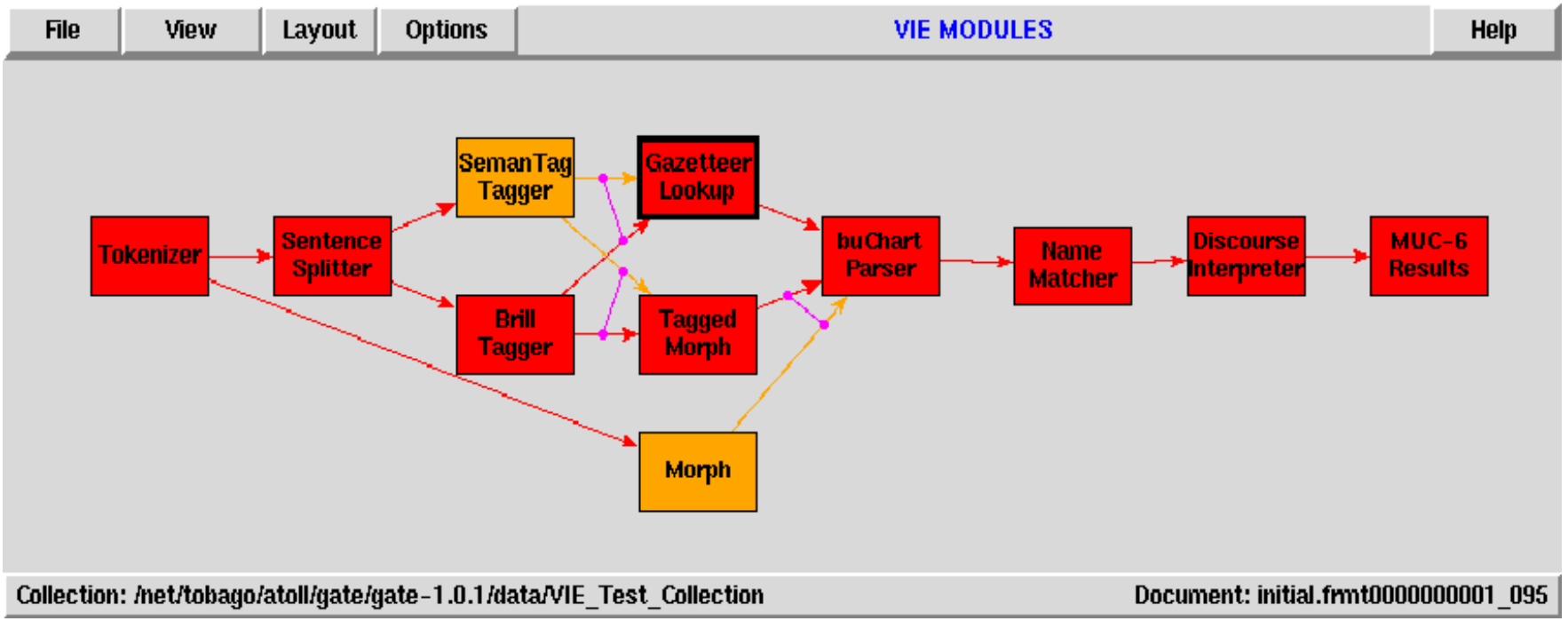
De nombreux composants et ressources à mettre en jeu, dépendant de la tâche, du domaine et de la langue !



Capture d'écran du système **GATE** développé à l'université de Sheffield (UK).

Chaîne de traitement linguistique

De nombreux composants et ressources à mettre en jeu, dépendant de la tâche, du domaine et de la langue !



Capture d'écran du système **GATE** développé à l'université de Sheffield (UK).

- Ressources linguistiques : accès et normalisation
- Outils : accès, agencement, communication, arbitrage

Normalisation / Standardisation

Expériences de standardisation de ressources :

- pour des grammaires TAG, [DTD TAGML]
- pour des sorties d'analyse syntaxique, forêts de dérivation, [DTD FOREST]
- pour des informations de morpho-syntaxe, Lionel Clément
- pour les structures de traits, basées sur TEI

Normalisation / Standardisation

Expériences de standardisation de ressources :

- pour des grammaires TAG, [DTD TAGML]
- pour des sorties d'analyse syntaxique, forêts de dérivation, [DTD FOREST]
- pour des informations de morpho-syntaxe, Lionel Clément
- pour les structures de traits, basées sur TEI

Utilisation de XML comme format d'échange et éventuellement de communication.
Conversions (XSLT, scripts Perl, BD, . . .) vers des formats de travail

Normalisation / Standardisation

Expériences de standardisation de ressources :

- pour des grammaires TAG, [DTD TAGML]
- pour des sorties d'analyse syntaxique, forêts de dérivation, [DTD FOREST]
- pour des informations de morpho-syntaxe, Lionel Clément
- pour les structures de traits, basées sur TEI

Utilisation de XML comme format d'échange et éventuellement de communication.
Conversions (XSLT, scripts Perl, BD, . . .) vers des formats de travail

Demarrage du projet **Normalangue**, dans le programme national Technolangue

Normalangue

- sur 3 ans, deux sous-projets : RNIL et Technolangue (technologies vocales)

Normalangue

- sur 3 ans, deux sous-projets : RNIL et Technolangue (technologies vocales)
- Terminologie pour les ressources linguistiques
- Schémas de représentation (méta-modèles) pour les ressources linguistiques, dont
  - annotations, morpho-syntaxe et syntaxe,contenus sémantiques multimodaux
  - données multilingues et mémoires de traduction
  - BD lexicales

Normalangue

- sur 3 ans, deux sous-projets : RNIL et Technolangue (technologies vocales)
- Terminologie pour les ressources linguistiques
- Schémas de représentation (méta-modèles) pour les ressources linguistiques, dont
  - annotations, morpho-syntaxe et syntaxe,contenus sémantiques multimodaux
  - données multilingues et mémoires de traduction
  - BD lexicales
- Mise au point d'API entrées/sorties pour les outils linguistiques (XML, Java, Open Source, UML)
- Jeux de tests

Normalangue

- sur 3 ans, deux sous-projets : RNIL et Technolangue (technologies vocales)

- Terminologie pour les ressources linguistiques
- Schémas de représentation (méta-modèles) pour les ressources linguistiques, dont
  - annotations, morpho-syntaxe et syntaxe,contenus sémantiques multimodaux
  - données multilingues et mémoires de traduction
  - BD lexicales

- Mise au point d'API entrées/sorties pour les outils linguistiques (XML, Java, Open Source, UML)
- Jeux de tests

- Vaste consortium [RNIL]
  académique : INRIA (L&D, ATOLL), TALaNa, ATILF, LLF, LIMSI, IRIN, CLIPS, RESO, CEA
  industrie : XRCE, EDF R&D, Systran, FT R&D, EADS, Softissimo, Sinequa, Lucid-IT, J-Way
  association : AFNOR
- En relation avec ISO TC37 SC4 et le groupe miroir français de l'AFNOR
- Liens avec initiatives internationales passées ou en cours : W3C, TEI, EAGLES, MATE, ISLE, ...
- Liens avec initiatives françaises : XMiner, Codex-terme, EVALDA/Easy, plateforme e-LinguiM

Infrastructure dans ATOLL

Infrastructure dans ATOLL

- Expérience (ancienne) avec GATE,
  Document primaire + base d'annotations

Infrastructure dans ATOLL

- Expérience (ancienne) avec GATE,
  Document primaire + base d'annotations

- Expériences récente avec des « pipelines XML » (Lionel Clément),
  dans la lignée de LT XML (Language Technologie Group, Edinburgh)
  $\Rightarrow$ notion de Document enrichi XML

Infrastructure dans ATOLL

- Expérience (ancienne) avec GATE,
  Document primaire + base d'annotations

- Expériences récente avec des « pipelines XML » (Lionel Clément),
  dans la lignée de LT XML (Language Technologie Group, Edinburgh)
  $\Rightarrow$ notion de Document enrichi XML

- Expériences de stockage en BD de ressources (forêts de dérivations) sous format XML,
  avec language de requêtes et accès WEB (Cocoon)
  $\Rightarrow$ Bases de données XML et services WEB

Infrastructure dans ATOLL

- Expérience (ancienne) avec GATE,
  Document primaire + base d'annotations

- Expériences récente avec des « pipelines XML » (Lionel Clément),
  dans la lignée de LT XML (Language Technologie Group, Edinburgh)
  $\Rightarrow$ notion de Document enrichi XML

- Expériences de stockage en BD de ressources (forêts de dérivations) sous format XML,
  avec language de requêtes et accès WEB (Cocoon)
  $\Rightarrow$ Bases de données XML et services WEB

- Mise en oeuvre dans l'ARC RLT « Ressources linguistiques pour les TAG »,
  acquisition semi-automatique de lexique pour les TAG (sémantique lexicale)
  $\Rightarrow$ bootstrap [TAL $\longrightarrow$ ressources linguistiques $\longrightarrow$ TAL] $\longrightarrow$ fouille de texte

Traitement de documents (SIMBIO)

Une expérience en cours dans ATOLL de traitement de documents Botaniques,
corpus de descriptions d'espèces végétales (Polynésie et Cameroun, $\sim$ 40 volumes, origine IRD)

Traitement de documents (SIMBIO)

Une expérience en cours dans ATOLL de traitement de documents Botaniques,
corpus de descriptions d'espèces végétales (Polynésie et Cameroun, $\sim$ 40 volumes, origine IRD)

- Correction post-OCR (traitements statistiques et linguistiques)
  distance d'édition, transducteurs et étiqueteurs
- Structuration XML – retrouver la structure logique
  $\sim$ « chunking »
- Extraction terminologique
- Acquisition d'ontologies
- Fouille de textes, vers base de connaissances
- Système d'information (navigation, recherche, identification, « éducation »)

Traitement de documents (SIMBIO)

Une expérience en cours dans ATOLL de traitement de documents Botaniques,
corpus de descriptions d'espèces végétales (Polynésie et Cameroun, $\sim$ 40 volumes, origine IRD)

- Correction post-OCR (traitements statistiques et linguistiques)
  distance d'édition, transducteurs et étiqueteurs
- Structuration XML – retrouver la structure logique
  $\sim$ « chunking »
- Extraction terminologique
- Acquisition d'ontologies
- Fouille de textes, vers base de connaissances
- Système d'information (navigation, recherche, identification, « éducation »)

Déjà, deux corpus structurés XML avec interfaces WEB de navigation

ATOLL et SYNTAX

- Expérience en traitement linguistique (dont syntaxe), avec utilisation de XML.
- Coordination avec Normalangue et ISO TC37SC4 sur les questions de normalisation et d'API
  Communication dans les 2 sens
- Recherche d'une infrastructure facile à mettre en oeuvre pour les traitements linguistiques et autres
  Utilisation de UML ?
- Aide bienvenue dans nos expériences de traitement de documents