

Nom	Version	Langage	Editeur	Catégorie	Date
GEonto_souslot1.2	V1	Fr		Recherche	15 juillet 2009

Rédacteurs

Eric Kergosien, Mauro Gaio

Le Poste : Configuration Requise

Type de machine : Intel Pentium 3, 4 (ou processeur AMD équivalent)

Système d'exploitation et logiciel(s) :

- Les différentes distributions de Linux et notamment : Debian Lenny - Ubuntu 8.10
- Windows 2000 Advanced Server (Service Pack 4 ou une version ultérieure), Windows Server 2003, Windows XP Pro ou Home (SP1 ou SP2);
- Une machine virtuelle java (JVM) 1.5 ou plus récente doit être installée sur le poste pour procéder ensuite à l'installation de Linguastream. Nous avons testé l'outil sur la version stable en date JDK 1.6 Update 3.

A noter :

- **Machine virtuelle Java 1.5 minimum requise.**
- **L'espace disque nécessaire pour Linguastream est de 100 Mo pour l'installation de l'édition Standard auquel doit être ajouté 30 Mo pour intégrer le Tree-tagger et Prolog (SWI). Prévoir de l'espace disque pour chaque chaîne de traitement qui seront définies ou intégrées à Linguastream. A titre d'exemple, la chaîne de traitement GEonto_souslot1.2 nécessite 50Mo d'espace disque en l'état actuel.**

Licences

Licences : Pour une utilisation académique veuillez vous reporter au condition d'utilisation sur les sites suivants:

- Linguastream : <http://http://www.linguastream.org/>
- TreeTagger : <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>
- SWI-Prolog : <http://www.swi-prolog.org/>
- le thésaurus RAMEAU : <http://rameau.bnf.fr/>

Commentaires

L'installation se déroule en 6 étapes :

Etape 1 : Installation du SDK java ;

Etape 2 : Installation du moteur prolog;

Etape 3 : Installation du Tree Tagger;

Etape 4 : Installation de Linguastream;

Etape 5 : Lancement et configuration de l'application Linguastream;

Etape 6 : Intégration de la chaîne Geonto_souslot1.2 V1

Quelques Précisions sur l'installation

Instructions :

- Posséder les droits d'administrateur du poste.
- Avant toute installation, fermer les applications ouvertes sur le poste.

Etapes :

Etape 1 : Installation de JDK 6 Update 3

- **Sous Linux (distributions debian et ubuntu)**

Rechercher le paquet sun-java6-bin via le gestionnaire de paquets synaptic et suivez les instructions d'installations en prenant tous les paquets dépendants recommandés par le gestionnaire. Les droits administrateurs sont nécessaires pour ajouter des paquets synaptic.

- **Sous Windows**

La version est disponible à l'adresse suivante : <http://java.sun.com/javase/downloads/?intcmp=1281>

Double clic sur l'exécutable. Cliquez sur next.

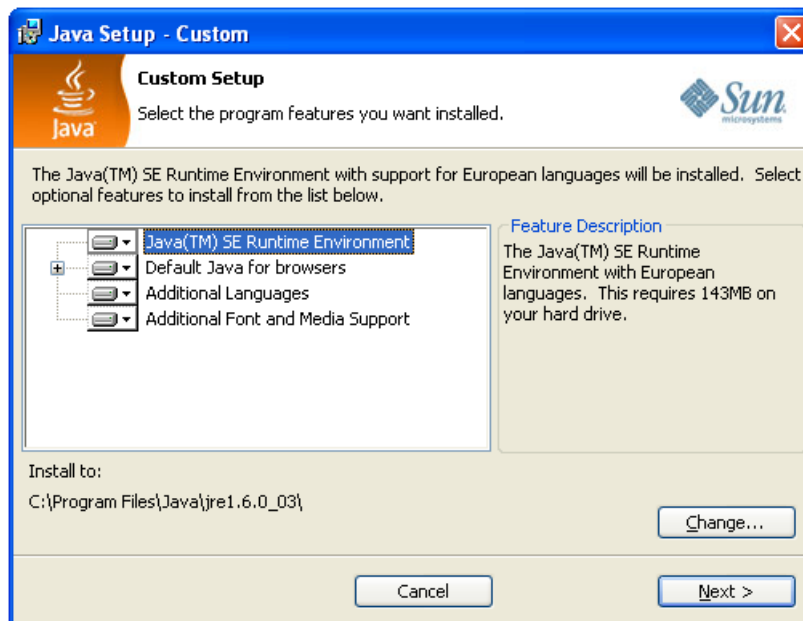


Figure 1. Installation Java 6 SDK Update 3 : choix des plugins et des répertoire d'installation

Choisissez installation complète et installer dans C:\jdk1.6.0.3

La JRE est installée dans C:\Program Files\Java\jre1.6.0_03.

Configuration des variables d'environnements

Allez dans démarrer ▪ panneau de configuration ▪ système : cliquez sur l'onglet avancé puis sur le bouton variables d'environnements :

- Ajouter la variable système suivante JAVA_HOME prenant pour valeur C:\jdk1.6.0.3 ;
- Modifier ensuite la variable suivante Path en lui ajoutantC:\jdk1.6.0.3\bin.

Etape 2 : Installation du moteur prolog

• Sous Linux

Rechercher le paquet swi-prolog (version 5.6.55-1 ou plus) via le gestionnaire de paquets synaptic et suivez les instructions d'installations en prenant tous les paquets dépendants recommandés par le gestionnaire. Les droits administrateurs sont nécessaires pour ajouter des paquets synaptic.

• Sous windows

Double clic sur l'exécutable *w32pl5644.exe*. Suivez la procédure d'installation telle qu'indiquée en gardant le même répertoire d'installation : C:\Program Files\pl.

Configuration des variables d'environnements

- Ajouter la variable SWI_HOME_DIR=<Path to SWI-Prolog>;
- Ajouter LIB de la façon suivante SWI_HOME_DIR%\lib ;
- Modifier PATH comme suit : %PATH%;%SWI_HOME_DIR%\bin;
- Ajouter la variable INCLUDE=%INCLUDE%; %SWI_HOME_DIR%\include.

Etape 3 : Installation du Tree Tagger

Le Tree Tagger est un système d'étiquetage automatique des catégories grammaticales des mots avec lemmatisation et possible.

- **Sous Linux**

Rechercher sur le site précité les éléments nécessaires à l'installation. La partie Download contient l'ensemble des éléments nécessaires (archive .tar.gz, le script tagging.scripts et le script d'installation) pour lancer l'installation.

- **Copier ensuite le fichier french.par dans le répertoire TreeTagger/lib pour la prise en compte du lexique français.**

- **Sous Windows**

Double clic sur l'archive *tree-tagger-windows-3.1.zip* et copier le répertoire dans C:\

Le tree tagger nécessite l'installation d'un interpréteur Perl que l'on peut obtenir à l'adresse <http://www.perl.com/pub/language/info/software.html>. Nous utilisons ici la version suivante : ActivePerl-5.8.8.822-MSWin32-x86-280952

Configuration des variables d'environnements

- Modifier PATH comme suit : %PATH, C:\TreeTagger\bin

Copier ensuite le fichier french.par dans le répertoire C:\TreeTagger\lib pour la prise en compte du lexique français.

Etape 4 : Installation de Linguastream

Une fois les différents composants installés, nous allons maintenant installer Linguastream.

- Aller récupérer la dernière version de Linguastream sur le site de l'éditeur <http://www.linguastream.org/> et suivre les étapes d'installation.

N.B.: Sous Windows il est préférable que le répertoire d'installation soit :

C:\Program Files\LinguaStream

- Une fois Linguastream installé, le lancer en cliquant sur l'icône placée sur le bureau ou en exécutant sous Windows le script : `C:\Program Files\LinguaStream\bin\linguastream.bat` et sélectionnez les plugins que vous désirez utiliser.

Etape 5 : Lancement et configuration de l'application Linguastream

Une fois l'outil lancé, vous arrivez à l'interface suivante :

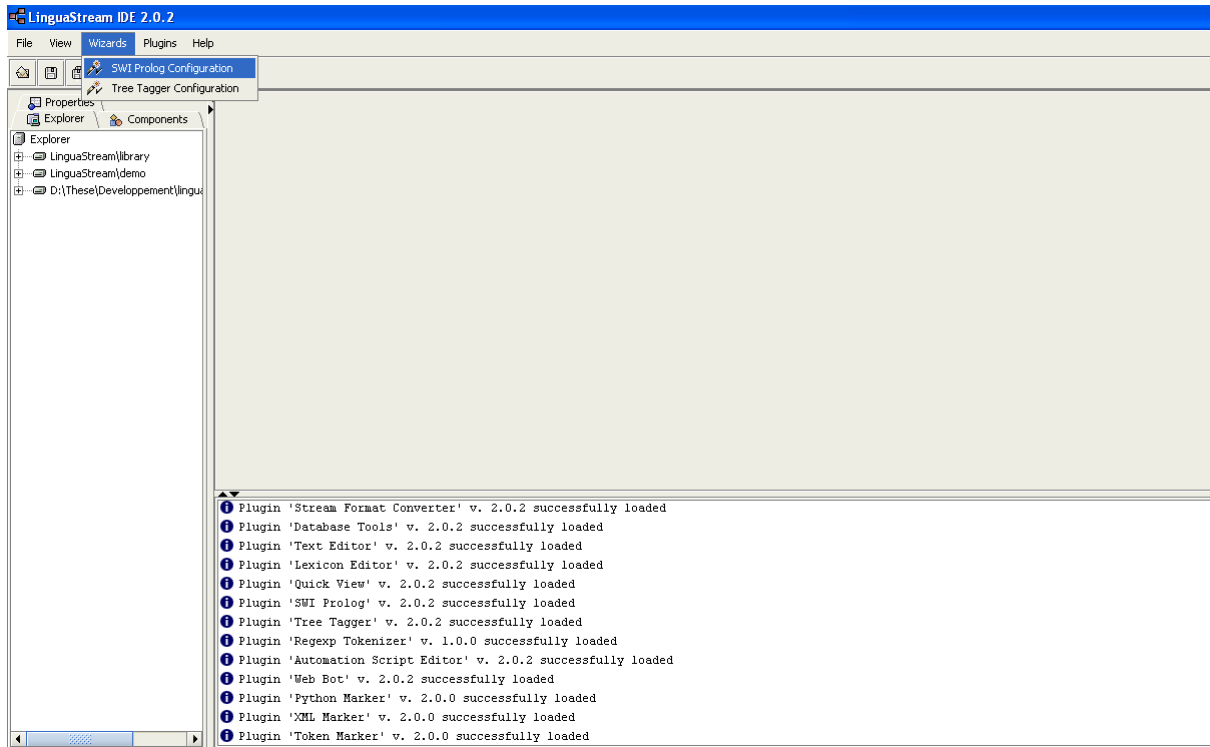


Figure 2. Interface de Linguastream

Nous allons maintenant configurer les variables d'environnement de l'outil pour pouvoir utiliser les modules DCG Marker (prolog) et TreeTagger.

- Sélectionner Wizards ▫ SWI-Prolog configuration. Une boîte de dialogue apparaît alors pour demander le chemin d'accès à l'exécutable Swi-prolog, vous n'avez plus qu'à donner le bon chemin d'accès.

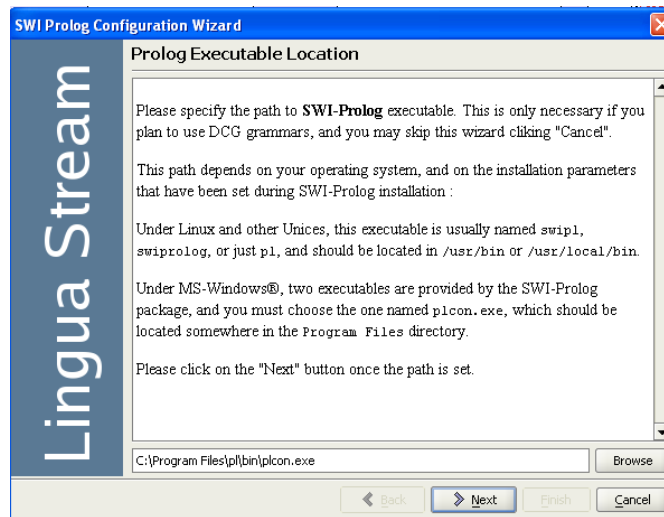


Figure 3. Configuration du prolog sous Linguastream

Une fois le chemin déterminé, cliquez sur suivant et valider la configuration.

- Sélectionner ensuite Wizards ▪ TreeTagger déterminer le chemin d'accès à l'exécutable Tree-tagger et au lexique french.par.

Une fois ces opérations réalisées, vous pouvez maintenant utiliser Linguastream et mettre en place vos propres chaînes de traitement.

Etape 6 : Intégration de la chaîne GEonto_souslot1.2

L'installation de la chaîne de traitement Geonto_souslot1.2 dans sa première version n'a été testée que sous Ubuntu 8.10 (figure 4).

- **Intégration de la chaîne à Linguastream :**

Pour intégrer la chaîne à Linguastream :

- décompresser l'archive ;
- sous Linguastream, cliquer bouton-droit sur l'explorer de fichier dans la fenêtre de gauche (figure 4), puis sélectionner « mount folder » et rechercher sur le disque le répertoire contenant la chaîne GEonto_souslot1.

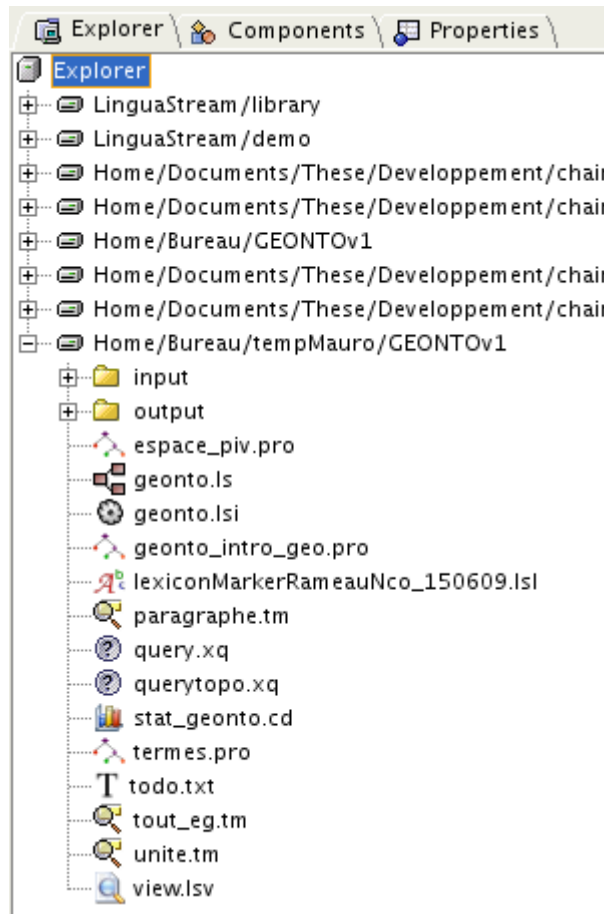


Figure 4 : Menu Explorer sous Linguastream intégrant la chaîne GEonto_souslot1.2

- Une fois la chaîne intégrée, double cliquer sur le fichier geonto.ls pour visionner l'ensemble de la chaîne dans la fenêtre principale

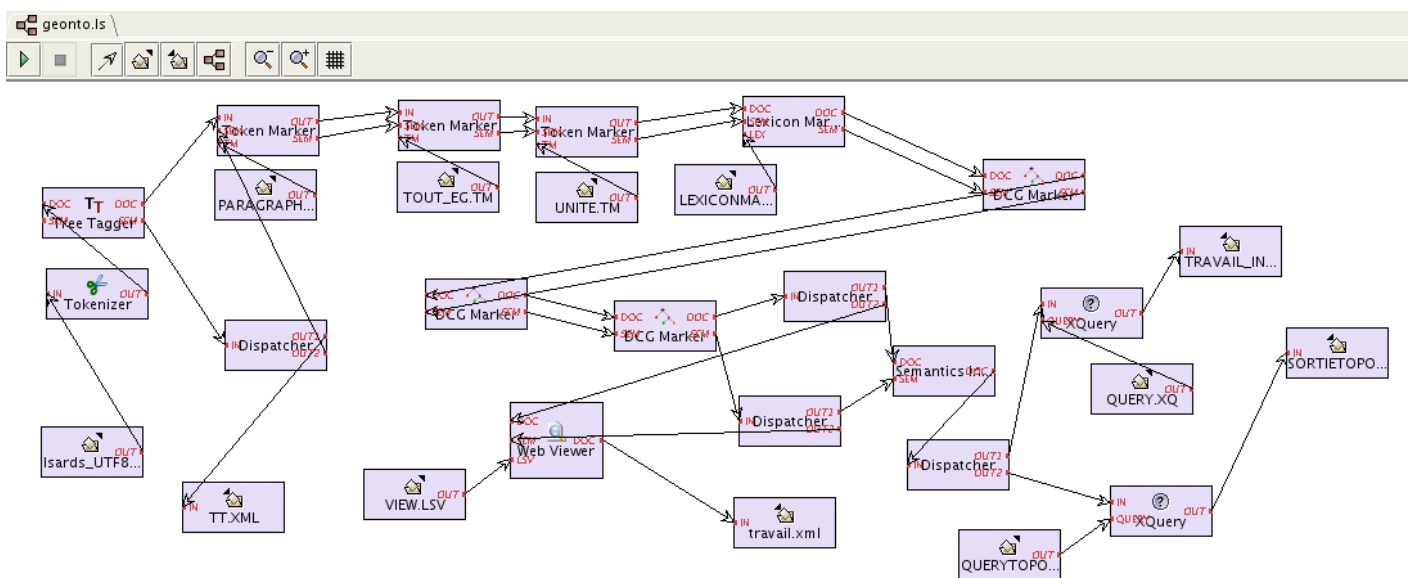


Figure 5 : Chaîne de traitement GEonto_souslot1.2

- **Composition de l'archive :**

Comme l'indique la figure 4, la chaîne est composée de :

- Répertoire `input` : contenant 10 livres type récits de voyage gracieusement fournis par la Médiathèque de Pau. Les documents au format TXT sont prêts à être traités dans la chaîne ;
- Répertoire `Output` : répertoire destiné à accueillir les fichiers traités par la chaîne ;
- Fichier `lexiconMarkerRameauNco_150609` : Thésaurus RAMEAU transformé au format `Isl` pour traitement sous Linguastream. Ce lexique regroupe l'ensemble des noms communs et le module `lexiconMarker` permet de marquer l'ensemble des mots et groupes de mots du lexique identifiés dans le texte traité ;
- Fichiers `.pro` : règles prolog permettant d'identifier les relations entre toponymes candidats identifiés et les noms communs qualifiants ces toponymes (qu'ils proviennent du lexique RAMEAU ou pas);
- Fichiers `.tm` : détection de motifs à partir d'expressions régulières. Ici pour nous, ce sont les entités nommées spatiales candidates ;

Lors de son exécution (bouton lecture), la chaîne de traitement prend sur une poste client tel que préconisé environ 1 minutes et 30 secondes pour traiter le livre 1 des œuvres type récits de voyages mis à disposition dans le répertoire `input`.

- **Sélection du fichier d'entrée :**

Pour lancer la commande, cliquer sur la boîte `file input` en bas à droite de la chaîne puis sur l'onglet `Propriétés` du menu général, fenêtre indiquant les propriétés de chaque éléments de la chaîne.

Modifier la variable `fileName` en allant rechercher via le navigateur de fichiers le fichier au format TXT qui doit être traité.

- **Visualisation de la sortie du traitement :**

Vérifier le fichier de sortie dans les propriétés du module `fileOutput` positionné à droite dans la chaîne sous Linguastream (le fichier de sortie et son emplacement sont modifiables avant de lancer le traitement) ;

Aller après traitement dans le répertoire `Output`, hors de Linguastream, et ouvrir le fichier résultat avec un éditeur XML adéquat.

Exemple de sortie :

```
<es-negation="false".type_es="esa".par_id="1".es_id="7">
  <esa-type_en="ascension".nom="Aneto">
    <texte>Aneto</texte>
  </esa>
</es>
<es-negation="false".type_es="esa".par_id="1".es_id="8">
  <esa-type_en="sommets".nom="Pyrénées">
    <texte>Pyrénées</texte>
  </esa>
</es>
```

Figure 6 : Extrait de sortie du traitement en version 1

Ce document contient un ensemble d'entités nommées spatiales candidates identifiées dans le texte traité avec pour chaque entité son qualifiant. Par exemple, l'entité spatiale « Pyrénées » a dans la figure 6 le qualifiant « sommets ».

Description de la chaîne

La chaîne de traitement GEonto_souslot1.2 dans sa première version intègre les éléments suivants :

- **Module Tokenizer** : ce service prend en charge la segmentation du texte. Il prend en entrée le texte et produit un autre flux dans lequel chaque mot est identifié et isolé dans une balise.
- **Module Tree-tagger** : ce service se charge de l'analyse morpho-syntaxique du texte.
- **Module Token Marker** : détection de motifs à partir d'expressions régulières. Il prend en entrée le flux XML et un fichier de ressources contenant les expressions régulières (au format tm) qui définissent les patrons à détecter. Le flux de sortie est augmenté de balises pour les textes correspondant à ces patrons.
- **Module LexiconMarker** : permet d'identifier dans le texte l'ensemble des mots et groupes de mots intégrés au lexique. Il donne la possibilité de typer une occurrence d'un mot par un autre mot renvoyant au même sens. Par exemple, en utilisant le lexique provenant du thésaurus RAMEAU, « Gouffre » identifié dans le texte sera marqué par le représentant « Grottes ».
- **Module DCG Marker** : ce service a pour but de détecter les relations qui existent entre les entités spatiales candidates. Il s'appuie sur des grammaires DCG (implémenté à l'aide du langage Prolog). Ces grammaires permettent de s'appuyer sur les mécanismes d'inférence et d'unification de Prolog à l'aide de règles simples. Il nous permet ici d'identifier les expressions composées d'une entité spatiales candidates et d'un groupe de mots qualifiants contenant un nom commun (qu'il provienne du lexique RAMEAU ou non).
- **Modules XQuery** : lors du traitement sémantique, chaque token est identifié par un numéro de paragraphe et un identifiant (dans le fichier doc). Ce module vient alors compléter le flux XML des identifiants des entités spatiales candidates. Il permet d'obtenir un flux XML valide par rapport au schéma défini (grâce aux fichiers.xslt donnés en entrées). Un extrait de la sortie finale après transformation via en ensemble de requête XQuery est proposé figure 6.