SygAnnot, un annotateur automatique lexical.
Alain Ibrahim, Fabien Priotto, Tarek Khelouf, Matthieu Laclau
Groupe Cydia Master 1 Informatique UM2 2011 http://informe.free.fr/sygannot

Travail d'étude et de Recherche encadré par la Pr. Violaine Prince avec la précieuse collaboration du Pr. Jacques Chauché

Résumé du travail en cours

SYGFRAN de Jacques Chauché est un analyseur d'énoncé français morpho-syntaxique écrit en SYGMART qui permet de transformer une phrase (texte brut) en un arbre syntaxique (élément structuré) enrichi d'informations sur les constituants. Devant ce potentiel, l'idée est née d'élaborer une application pour l'annotation automatique des textes offrant un double intérêt : un procédé pratique et une approche expérimentale pour l'annotation.

Il s'agit donc de faire réaliser, par un ordinateur, le travail d'un re-lecteur dont le but n'est pas de corriger le texte initial mais d'y insérer des précisions.

Notre application, codée en langage JAVA se compose de 3 modules :

- Le gestionnaire **AppliLexi** pour la création et la mise à jour des lexiques. Il permet d'extraire une liste d'étiquettes d'un corpus constitué de textes à forte cohérence thématique. Dans une première version, nous nous concentrons sur les groupes nominaux. Les syntagmes les plus fréquents issus sont retenus pour constituer les items lexicaux.

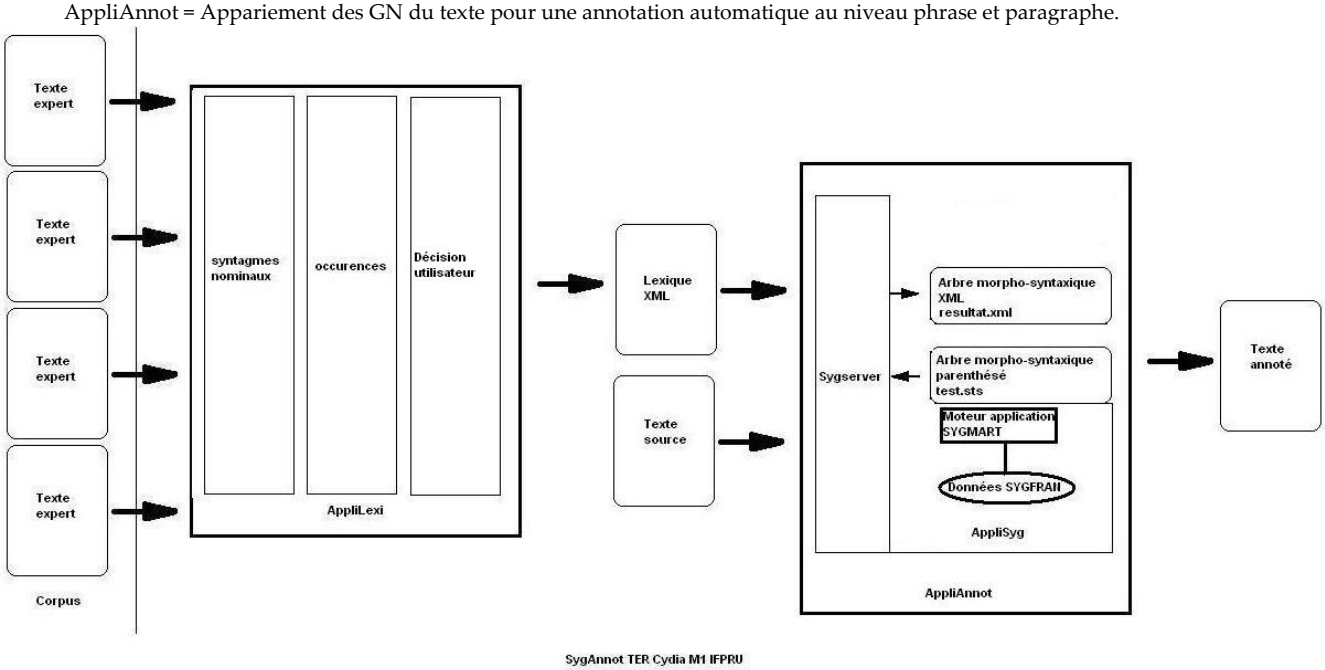
- Le moteur d'application SYGMART **AppliSyg** de Jacques Chauché qui génère l'arbre morpho-syntaxique d'un texte. Sygserveur, un utilitaire fournit par Alexandre Labadié transforme l'arbre parenthèsé généré par AppliSyg en fichier XML plus simple à parcourir.

- L'annotateur **AppliAnnot** qui exploite l'arbre SYGMART pour insérer des « notes » à partir du lexique thématique fabriqué par AppliLexi.

On distingue deux niveaux d'annotation : annotation au niveau de la phrase et au niveau du paragraphe. Nous recherchons la cohérence thématique. On convient d'une structure basée sur l'existence d'une super classe qui est le gouverneur du groupe local.

Principe de SygAnnot

AppliLexi = Extraction lexicale des GN depuis un corpus fournit par SYGFRAN
AppliAnnot = Appariement des GN du texte pour une annotation automatique au niveau phrase et paragraphe.



SygAnnot TER Cydia M1 IFPRU

Travail effectué :

Le module AppliLexi est terminé. Nous générons automatiquement un lexique XML que l'utilisateur peut facilement épurer pour ne garder que les étiquettes qui ont un intérêt dans le thème choisi.

Travail qui reste à faire :

Il nous reste à concevoir et programmer la procédure d'appariement dans le cœur du module AppliAnnot.

Difficultés rencontrées :

Une des difficultés rencontrées est liée au fait que nous ne somme pas maître de l'utilitaire SygServeur qui permet la conversion de l'arbre SYGMART vers XML. Il nous a fallu un certain temps pour l'obtenir du fait d'une incompatibilité avec le moteur SYGMART fournit initialement. La procédure d'identification des groupes nominaux qui tient compte des règles de la grammaire française n'a pas été facile à programmer.

Précisons que depuis le départ nous souffrons d'un handicap à cause de la formalisation tardive de nos objectifs (mardi 8 mars) faisant suite aux 3 rejets successifs de nos propositions pour un choix de TER.