

Projet Master 1 IC (2026-27)

Titre du sujet : Finalisation de TTEDIT en java

(rév. 28/9/23)

Nom des encadrants : Christian Boitet & Jean-Claude Durand
Email des encadrants : Christian.Boitet@imag.fr, Jean-Claude.Durand@imag.fr
Téléphone(s) : +33 (04574) 21459, 0660051969

1. Présentation du sujet

Il s'agit de contribuer à une partie du système Ariane-Y, en cours de développement par les encadrants. C'est un environnement de construction et d'exploitation (sur le web) de systèmes « experts » de TALN (traitement automatique des langues naturelles) et en particulier de TAO (traduction automatique ou automatisée). Il contient des langages spécialisés pour la programmation linguistique (LSPL) : ATEF pour l'analyse morphologique, EXPANS/TRANSF pour l'expansion et la traduction lexicales, ROBRA pour les systèmes transformationnels d'arbres « décorés » (utilisés en analyse, en transfert structural et en génération syntaxique), SYGMOR pour la génération morphologique étendue (mise en page, ponctuations, transcriptions...).

Ces LSPL permettent de construire des modules appelés *phases*, chacune comportant des *composants linguiciels* : déclarations d'attributs, formats (modèles), dictionnaires et grammaires-automates. Pour chaque type de composant, Ariane-Y contient un compilateur séparé, produit par ANTLR-4 (voir <https://wwwantlr.org>), par exemple ATEFgrammaire.g4, et qui se présente comme une méthode java, ici par exemple AYm::analyserGR. Le serveur Ariane-Y contient aussi la description des composants, sous forme d'un fichier « **arbre XML** », ne contenant que des balises avec leurs attributs. Un extrait est donné en annexe.

Le but du projet est de poursuivre l'implémentation en java de l'éditeur d'arbres TTEdit à 3 types d'arbres, à savoir, outre les arbres décorés et les arbres XML, les « **arbres simples** », simplement étiquetés. On veut aussi compléter la fonctionnalité de « réécriture d'arbres » en implémentant des **règles** de forme : **schéma**→**image**. Dans la première partie du projet (jusqu'à mars), il s'agira d'assimiler les techniques et outils utilisés dans le projet ARIANE-Y : non seulement (1) ANTLR, mais aussi (2) Doxygen pour la documentation des programmes et sa transformation en un ensemble de pages web ainsi qu'un fichier pdf, et (3) git (sur la forge du LIG) pour le développement partagé. On définira aussi une structure interne unique pour les 3 types d'arbres, qui sera la structure de travail de l'éditeur (lui-même graphique et bâti autour de JavaSwing), et on adaptera leurs compilateurs en conséquence.

Dans la seconde partie du projet (d'avril à mi-juin), on passera à la réalisation : développement de la partie « règles de réécriture » (en intégrant la documentation en Doxygen aux programmes), et tests unitaires sur des fichiers à fabriquer pour tester différents cas.

S'il reste du temps, on pourra étudier l'ajout de fonctions de vérification (par exemple, vérifier que toutes les chaînes font partie d'un vocabulaire (dictionnaire) donné en paramètre) et d'extraction (par exemple, valeurs des attributs).

2. Références

Ariane-Y : Spécifications du langage arbreXML (sous-langage de XML), à partir de la partie purement grammaticale de la grammaire ANTLR arbreXML.g4.

ANTLR : on se référera à <https://wwwantlr.org> et aux compilateurs déjà développés pour Ariane-Y.

Langages : Java supposé connu.

Outils : utilisation de git et de la forge du LIG.

Documentation avec Doxygen : on se référera à <https://www.doxygen.nl>.

3. Positionnement du sujet

- Indiquez le niveau d'innovation du sujet proposé

Très innovant ☐ ☐ ☐ ☒ ☐ Classique

- Indiquez la disponibilité de la documentation relative aux technologies à mettre en œuvre

Beaucoup ☐ ☒ ☐ ☐ ☐ Aucune

- Indiquez le niveau d'abstraction du sujet

Théorique ☐ ☐ ☒ ☐ ☐ Pratique

- Indiquez la quantité de développement à réaliser

Beaucoup ☐ ☐ ☒ ☐ ☐ Peu

- Indiquez le niveau de difficulté des algorithmes à mettre en œuvre

Difficile ☐ ☒ ☐ ☐ ☐ Facile

- Indiquez le niveau d'interaction avec d'autres composants logiciels

Ecosystème complexe ☐ ☐ ☒ ☐ ☐ Application seule

- Indiquez le nombre d'étudiant(e)s souhaité(e)s pour le projet : 2

- Indiquez les langages et technologies à utiliser :

Java, ANTLR, Doxygen

4. Encadrement

- Combien de temps pouvez-vous consacrer à l'encadrement de projets chaque semaine ?

Les 2 encadrants devraient être disponibles :

1^{ère} partie : 2 ½ journées par semaine.

2^{ème} partie : 4 à 5 ½ journées par semaine.

- Indiquez vos contraintes quant à l'encadrement

Pas de contraintes a priori.

- Indiquez vos contraintes quant au sujet proposé

Ce sujet implique un bon équilibre entre théorie et pratique. Il faut non seulement aimer développer, avec rigueur (suivre les préceptes et méthodes des cours de GI !), mais aussi avoir de la curiosité, en particulier pour s'initier aux arcanes de l'analyse syntaxique LL(*) mise en œuvre dans ANTLR par Terence Parr. ANTLR4 est un outil extrêmement puissant. Il sera utile de lire le livre de Terence Parr (voir sur le Web).

Au terme de ce projet, les étudiants auront beaucoup progressé en GI, techniquement (ANTLR) et méthodologiquement (Doxygen, Git).

À retourner à :

Damien.Pellier@univ-grenoble-alpes.fr

Extrait d'un fichier ArbreXML (ici fichInitDLING.arbDS). Ici, l'élément racine est <DLING> avec comme liste d'attributs :

```
codeBalAY = "DLING" symbDLING = "DLING" nomBalAY = "DLING-données_linguistiques_chargées"
dateCreation = "20140221" dateDerniereModif = "20200404" statutDLING = "init"
```

```

38 <DLING codeBalAY = "DLING"
39 symbDLING = "DLING" nomBalAY = "DLING-données_linguistiques_chargées"
40 dateCreation = "20140221" dateDerniereModif = "20200404"
41 statutDLING = "init">
42 <MV codeBalAY = "MV" nomBalAY = "MV_AM_TUX" admin = "?XAN JPG JCD EB"
43 dateCreation = "20210909" dateDerniereModif = "20220401"
44 statut = "?test" taille = "0">
45 <PHASES codeBalAY = "PHASES" nomBalAY = "PHASES_de_MV_AM_TUX"
46 dateCreation = "20210909" dateDerniereModif = "20220405"
47 statutPHASES = "mise_au_point">
48 <!-- Phase ATEF -->
49 <Phase codeBalAY = "Phase" symbPhase = "AM"
50 nomBalAY = "AM/atef"
51 codeLangueSource = "RUS" nomLangueSource = "russe"
52 codeLangueCible = "?LC" nomLSPL = "ATEF"
53 syntaxeCompling = "AY" conventionChaines = "chaineAY"
54 encodageOriginal = "UTF-8"
55 statutChargement = "charg_0" statutVerification = "verif_0"
56 dateCreation = "20140221" dateDerniereModif = "20140221" taille = "0">
57 <!-- Mettre ici le contenu de la phase (composants) -->
58 <DVAR codeBalAY = "DVAR" nomBalAY = "DVAR" bDVAR = "?0"
59 statutChargement = "charg_0"
60 statutVerification = "verif_0" nbErreurs = "0"
61 dateCreation = "20140221" dateDerniereModif = "20210909"
62 adresseSI = "0" taille = "0" > </DVAR>
74
75 <FORMATS codeBalAY = "FORMATS" bFORMATS = "0"
76 nomBalAY = "FORMATS_des_différents_LSPL" > </FORMATS>
98
99 <GRAMMAIRES codeBalAY = "GRAMMAIRES"
100 nomBalAY = "GRAMMAIRES" bGRAMMAIRES = "0"
101 dateCreation = "20140507" dateDerniereModif = "20140507"
102 lGram = "1 2 3 4 5 6 7 8" > </GRAMMAIRES>
144
145 <DICTIONNAIRES codeBalAY = "DICTIONNAIRES"
146 nomBalAY = "DICTIONNAIRES" bDICTIONNAIRES = "0"
147 dateCreation = "20140507" dateDerniereModif = "20190508"
148 lDicBaz = "" lDicAff = ""
149 pDicBA = "9 8 7 6 5 4 3 2 1" lDicTourn = ""

```

Écran TTEDIT montrant un arbre décoré.

