Héloïse, une plate-forme pour développer des systèmes de TA compatibles Ariane en réseau

Vincent Berment[1,] Christian Boitet[2] Guillaume de Malézieux[1]
(1) INaLCO, 65 rue des Grands Moulins, 75013 Paris, France
(2) GETALP, 700 avenue Centrale, 38401 St Martin d'Hères, France
Vincent.Berment@inalco.fr, Christian.Boitet@imag.fr,
guillaume2l2m@gmail.com

RÉSUMÉ

Dans cette démo, nous montrons comment utiliser Héloïse pour développer des systèmes de TA.

ABSTRACT

Heloise, a platform for collaborative development of Ariane-compatible MT systems
In this demo, we present how to use Heloise for developing new MT systems.

MOTS-CLÉS : Traduction Automatique, langues peu dotées, développement collaboratif.
KEYWORDS: Machine Translation, under-resourced languages, collaborative development.

L'objectif de la démonstration est de montrer l'utilisation d'Héloïse 2.0, un environnement de travail permettant de développer des systèmes de Traduction Automatique (TA) à partir d'un navigateur internet.

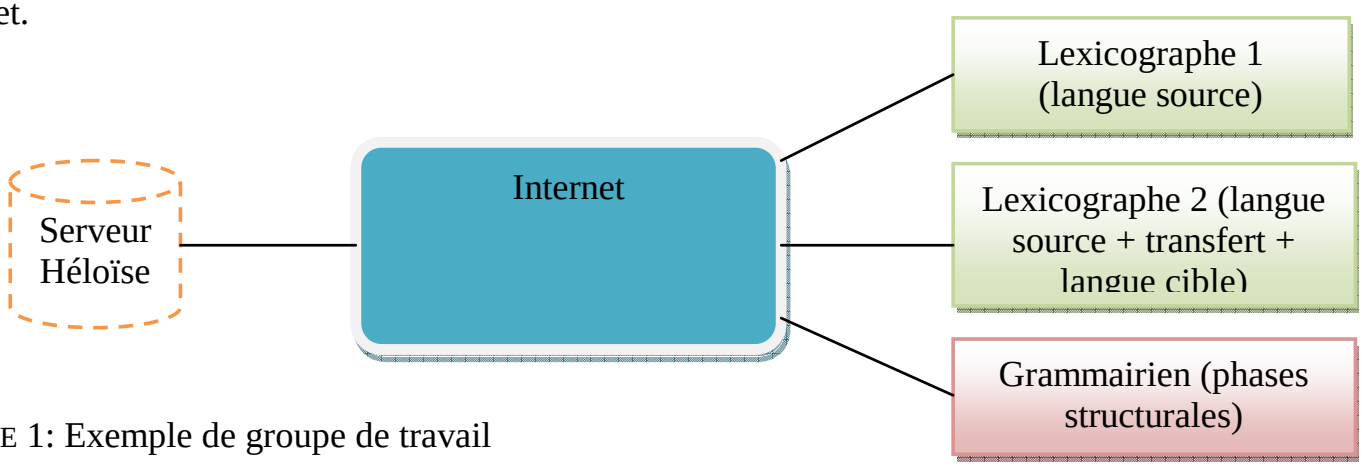
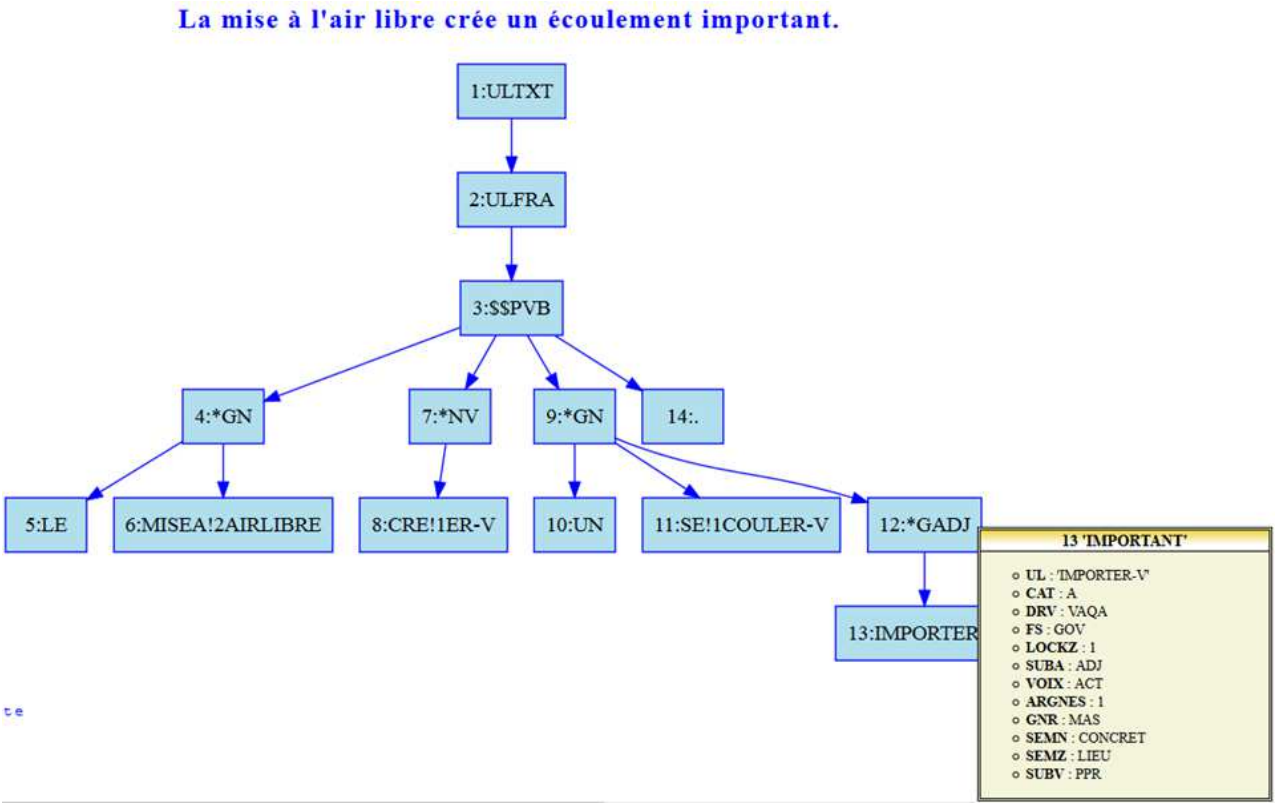


FIGURE 1: Exemple de groupe de travail

L'idée sous-jacente est de faciliter le travail en commun de personnes géographiquement dispersées et souhaitant collaborer au développement d'un (ou plusieurs) système(s) de TA. Ceci est intéressant en particulier pour les langues mal couvertes par les systèmes de TA existants (diasporas). Héloïse inclut une réécriture des compilateurs Ariane-G5[1], ce qui lui confère une totale compatibilité ascendante (Berment, Boitet, 2012). Les systèmes de TA développés sous Ariane par le GETA ont été mis sous licence BSD et sont ainsi disponibles à titre d'exemples dans Héloïse.

[1] Ariane-G5 est un environnement de développement conçu et réalisé par le GETA à Grenoble dans les années 1970-1990.

L'un des principaux avantages d'Ariane et par conséquent d'Héloïse est la (relative) facilité de réalisation des systèmes de TA. Ceci est dû à l'existence de langages de programmation linguistique qui permettent aux linguistes de définir leurs propres objets (ex. : paradigmes morphologiques, classes morphosyntaxiques, relations argumentaires et sémantiques…) ainsi qu'à l'existence d'une méthodologie linguistique qui guide le développeur. Cette méthodologie, qui s'appuie sur des théories linguistiques dont celle de Lucien Tesnière, explique comment obtenir une représentation abstraite aussi indépendante de la langue que possible (relations entre les prédicats et leurs arguments, relations sémantiques...), qui mixe un arbre de constituants, un arbre de dépendances et un graphe de relations prédicat-arguments et sémantiques.



Ces structures, appelées structures multiniveaux de Vauquois, permettent ensuite de générer la traduction de l'énoncé dans n'importe quelle langue. Il est à noter que la méthodologie permet plusieurs approches dont le passage par un transfert ou par un pivot sémantique comme UNL.

L'environnement Héloïse est constitué d'une zone d'information (wiki, blog, communauté…) et d'une zone dédiée au développement (environnement complet de développement « linguiciel »). Chaque utilisateur a des identifiants pour une protection maximale des données. Depuis la sortie de la version bêta en 2010, le développement de plusieurs systèmes a été entrepris dont :

- des analyseurs morphologiques : l'allemand (Guilbaud et al., 2013), du lituanien (Kapočiūtė-Dzikienė et al., 2016)), du quéchua (Maximiliano Duran) et du russe réalisé à partir des données lexicales de Vincent Benet, professeur à l'INaLCO,
- un projet multilingue impliquant des personnes distants géographiquement et visant à traduire le Petit Prince de Saint-Exupéry entre de nombreuses langues dont des langues d'Asie du Sud-Est (birman, cambodgien, lao, thaï…),
- un projet visant à dériver un système espagnol-anglais à partir du système existant portugais-anglais, ce qui a permis de montrer l'efficacité de la méthode puisqu'un premier système a pu voir le jour en six mois environ.

Références

BERMENT V., BOITET C. (2012). Heloise – An Ariane-G5 compatible environment for developing expert MT systems online. Actes de *COLING 2012 (Demonstration Papers)*, 9-16.

KAPOČIŪTĖ-DZIKIENĖ J., BERMENT V., RIMKUTĖ E. (2016). A Lithuanian Lemmatizer Designed for Open Online Collaborative Machine Translation. Article soumis à *Baltic HLT 2016*.

GUILBAUD J-P., BOITET C., BERMENT V. (2013). Un analyseur morphologique étendu de l'allemand traitant les formes verbales à particule séparée. Actes de *TALN 2013 (Volume 2 : papiers courts)*, 755-763.