Heloise — An Ariane-G5 compatible environment for developing expert MT systems online

Vincent Berment[1] *Christian Boitet*[2]
(1) INaLCO, Place du Maréchal de Lattre de Tassigny - 75775 Paris cedex 16
(2) GETALP, LIG-campus, 385, avenue de la Bibliothèque - 38041 Grenoble cedex 9

Vincent.Berment@imag.fr, Christian.Boitet@imag.fr

ABSTRACT

Heloise is a reengineering of the specialised languages for linguistic programming (SLLPs) of Ariane-G5 running both Linux and Windows. Heloise makes the core of Ariane-G5 available to anyone willing to develop "expert" (i.e. relying on linguistic expertise) operational machine translation (MT) systems in that framework, used with success since the 80's to build many prototypes and a few systems of the "multilevel transfer" and "interlingua" architecture. This initiative is part of the movement to reduce the digital divide by providing easily understandable tools that allow the development of lingware for poorly-resourced languages (π-languages). This demo article presents Heloise and provides some information about ongoing development using it.

KEYWORDS: machine translation, specialised languages for linguistic programming, SLLP, MT lingware, online lingware building, collaborative lingware building, Ariane-G5, Heloise, under-resourced languages

TITRE ET RÉSUMÉ EN FRANÇAIS

Héloïse — Un environnement compatible Ariane-G5 pour développer des systèmes de TA experte

Héloïse est une réingénierie des langages spécialisés (LSPL) d'Ariane-G5 tournant sous Linux et Windows. Héloïse rend le cœur d'Ariane-G5 accessible à toute personne désirant réaliser par elle-même des systèmes de traduction automatique (TA) experts (s'appuyant sur une expertise linguistique) opérationnels dans cet environnement, qui a été utilisé avec succès depuis les années 80 pour construire de nombreux prototypes et quelques systèmes adoptant une architecture de "transfert multiniveau" et d'"interlingua". Cette démarche s'inscrit dans le mouvement visant réduire la fracture numérique par la mise à disposition d'outils facilement appropriables, et permettant de développer des linguiciels pour des langues peu dotées (langues-π). Cet article démo présente Héloïse et fournit quelques informations sur les développements actuels réalisés avec Héloïse.

MOTS-CLÉS EN FRANÇAIS : traduction automatique, langages spécialisés pour la programmation linguistique, LSPL, linguiciels de TA, construction en ligne de linguiciels, Ariane-G5, Héloïse, langues peu dotées.

Proceedings of COLING 2012: Demonstration Papers, pages 9–16,
COLING 2012, Mumbai, December 2012.

1 Introduction

Ariane-G5 is a generator of machine translation systems developed and improved by the GETA group[1] during the years 1970 and 1980. This framework, despite the numerous publications and cooperative projects that made it widely known, remains of difficult access because of the "mainframe" environment under which it runs (zVM/CMS on z390). Ariane-G5 can be accessed either natively through a 3270 terminal emulator or using CASH, a portable "meta-environment" (written in Revolution) which contains the source files (lingware, corpus), and which communicates with Ariane-G5 that performs all the treatments (compilations and executions of "translation chains").

Heloise is a reengineering of compilers and "engines" of Ariane-G5's Specialized Languages for Linguistic Programming (SLLPs), running both Linux and Windows. The aim of its author when he developed this new version of Ariane-G5 SLLPs, was to make this system available to anyone wishing to design his own operational expert MT system (i.e. an MT system relying on linguistic expertise, as opposed to systems based on statistical properties of languages). This approach is part of the movement aiming at reducing the digital divide through the provision of tools, usable by non-specialists, and enabling them to develop their own language services.

This demo article aims at presenting Heloise and provides some information about ongoing development using it.

2 Ariane-G5

2.1 General principles

Ariane-G5 is a generator of machine translation systems. It uses an expert approach (including a description of the languages handled) and the generated systems are generally based on a multilevel transfer linguistic architecture, and developed using a heuristic programming approach. It has also been used for "abstract pivot" approaches (IF semantico-pragmatic formulas for speech MT in the CSTAR and Nespole! projects in 1995-2003, and UNL linguistic-semantic graphs since 1997).

Ariane-G5 relies on five Specialized Languages for Linguistic Programming (SLLPs) operating on decorated trees. The specificity of an SLLP is that it offers high-level data structures (decorated trees or graphs, grammars, dictionaries) and high-level control structures (1-ary or N-ary non-determinism, pattern-matching in trees, guarded iteration).

A minimal translation system produced by Ariane-G5 includes 6 phases (MA, SA, LT, ST, SG, MG), grouped two by two into 3 steps:

- Morphological Analysis and Structural Analysis (analysis step),
- Lexical Transfer and Structural Transfer (transfer step),
- Structural Generation and Morphological Generation (generation step).

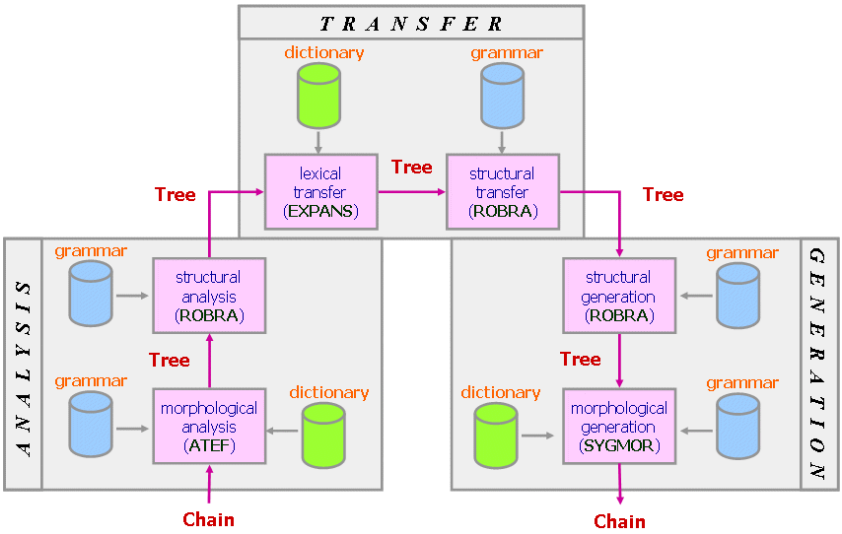

FIGURE 1 – Ariane: analysis, transfer et generation.

2.2 Lingware available to L-developers

The following lingware source is available under BSD license[2] since September 2010. This code greatly facilitates the implementation of new systems[3]. These lingware modules include:

- Large-scale prototype systems
  - Russian-French: RU5-FR5
  - French-English: BV-aéro/FE
  - English-French: B'VITAL
- Mockup systems
  - English-French teaching mockup: BEX-FEX
  - French-English teaching mockup: FEX-BEX
  - French-English (DGT, Telecommunications)
  - Portuguese-English
  - French-Russian (LIDIA)
  - French-German (LIDIA)
  - French-English (LIDIA)
  - UNL-Chinese: WNL-HN3
  - UNL-French: UNL-FR5

[2] See http://en.wikipedia.org/wiki/BSD_licenses.
[3] These lingware modules will soon be available for download.

- French-UNL : FR6-UNL
- English-Malay: ANG-MAL
- English-Thai: IN4-TH4
- English-(Chinese, Japanese, Arabic)
- Chinese-(English, French, German, Russian, Japanese)
- German-French

- Steps or groups of isolated phases
  - Analysis of Portuguese: AMPOR+ASPOR
  - Analyses of German: AMALX...
  - Analysis of Japanese (Annick Laurent's PhD thesis)

3 Heloise, screenshots and comments

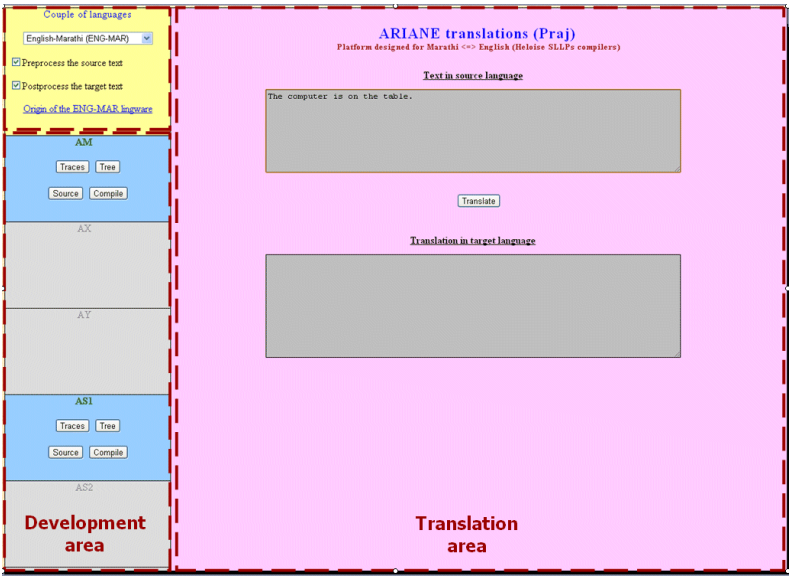

FIGURE 2 – Lingware development environment of Heloise

Heloise is an on-line tool available from any browser. The lingware developer (L-developer) is offered three areas as shown in figure 2:

- A "Selection" area in which he can select the current pair of languages,
- A "Development" area from which he controls the compiling and testing process,
- A "Translation" area in which he can make his translation trials.

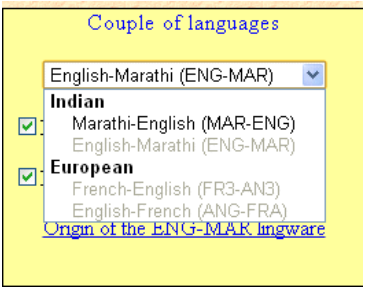

FIGURE 3 – Selection area

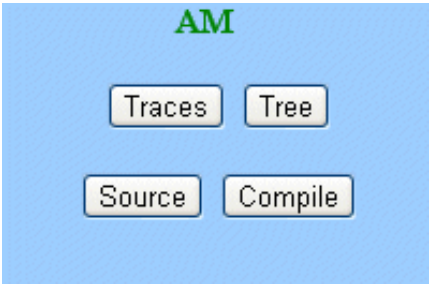

FIGURE 4 – Development area

The "Development" area provides four commands to the L-developer for each phase:

- The "Source" button, to get access and to manage the lingware files,
- The "Compile" button, to compile the lingware of the phase,
- The "Traces" button, to see the logs of a translation trial for the phase,
- The "Tree" button, to display the output tree of a translation for the phase.

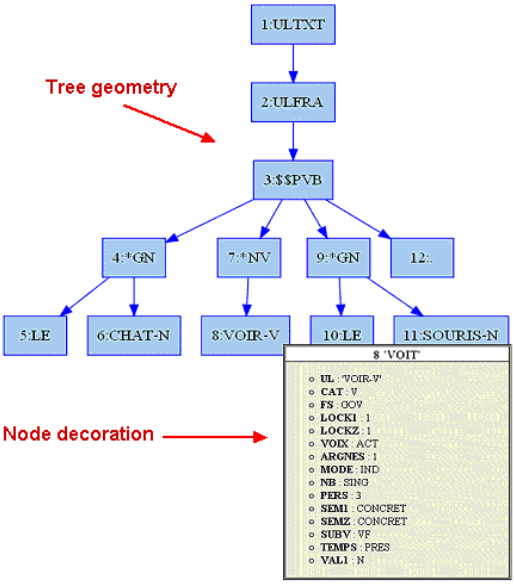

FIGURE 5 –Geometry and node decorations of an output tree

Source files of the Praj directory		
File name	Heloise => Local	Local => Heloise
FILASENG.FMATA.txt	Download	Choisissez un fichier Aucun fichier choisi Upload
FILASENG.GRAM1.txt	Download	Choisissez un fichier Aucun fichier choisi Upload
FILASENG.VARBG.txt	Download	Choisissez un fichier Aucun fichier choisi Upload

FIGURE 6 – From this table, the L-developer downloads and uploads his lingware source files

Conclusion

L-development works are currently undertaken for several pairs of languages, including:

- Several π-languages (under-resourced languages) such as Khmer, Lao and Marathi,
- European languages, such as English, French and German.

The quality reached for the MT systems is the one obtained with Ariane-G5's methodology and tools, but without the size limitations of Ariane-G5 (~50 pages input text, 64Knodes trees…).

References

Bachut D., Le projet EUROLANG : *une nouvelle perspective pour les outils d'aide à la traduction*, Actes de TALN 1994, journées du PRC-CHM, Université de Marseille, 7-8 avril 1994.

Bachut D., Verastegui N., *Software tools for the environment of a computer aided translation system*, COLING-1984, Stanford University, pages 330 à 333, 2-6 juillet 1984.

Berment V., *Méthodes pour informatiser les langues et les groupes de langues peu dotés*, Thèse de doctorat, Grenoble, 18 mai 2004.

Boitet C., *Le point sur Ariane-78 début 1982 (DSE-1), vol. 1, partie 1, le logiciel*, rapport de la convention ADI n° 81/423, avril 1982.

Boitet C., Guillaume P., Quézel-Ambrunaz M., *A case study in software evolution: from Ariane-78.4 to Ariane-85*, Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Colgate University, Hamilton, New York, 14-16 août 1985.

Boitet C., *Current machine translation systems developed with GETA's methodology and software tools*, conférence Translating and the Computer 8, 13-14 novembre 1986.

Boitet C., *La TAO à Grenoble en 1990, 1980-90 : TAO du réviseur et TAO du traducteur*, partie des supports de l'école d'été de Lannion organisée en 1990 par le LATL et le CNET, 1990.

Boitet C., *A research perspective on how to democratize machine translation and translation aids aiming at high quality final output*, MT Summit VII, Kent Ridge Digital Labs, Singapour, pages 125 à 133, 13-17 septembre 1999.

Boitet C., *A roadmap for MT: four « keys » to handle more languages, for all kinds of tasks, while making it possible to improve quality (on demand)*, International Conference on Universal

Knowledge and Language (ICUKL 2002), Goa, 25-29 novembre 2002.

Boitet C., *Les architectures linguistiques et computationnelles en traduction automatique sont indépendantes*, TALN 2008, Avignon, 9-13 juin 2008.

Del Vigna C., Berment V., Boitet C., *La notion d'occurrence de formes de forêt (orientée et ordonnée) dans le langage ROBRA pour la traduction automatique, Approches algébrique, logique et algorithmique*, Journée thématique ATALA sur la traduction automatique, ENST Paris, 1er décembre 2007.

Guillaume P., Ariane-G5 : *Les langages spécialisés TRACOMPL et EXPANS*, document GÉTA, juin 1989.

Guilbaud J.-P., Ariane-G5 : *Environnement de développement et d'exécution de systèmes (linguiciels) de traduction automatique*, Journée du GDR I3 co-organisée avec l'ATALA, Paris, novembre 1999.

Nguyen H.-T., *Des systèmes de TA homogènes aux systèmes de TAO hétérogènes*, Thèse de doctorat, Grenoble, 18 décembre 2009.

Vauquois B., *Aspects of mechanical translation in 1979*, Conference for Japan IBM Scientific program, juillet 1979.

Vauquois B., *Computer aided translation and the Arabic language*, First Arab school on science and technology, Rabat, octobre 1983.