Published on *Multilingual Online Translation* (http://www.molto-project.eu)

DX.2 Annual Public Report

Contract No.:	FP7-ICT-247914
Project full title:	MOLTO - Multilingual Online Translation
Deliverable:	DX.2 Annual public report
Security (distribution level):	Public
Contractual date of delivery:	M24
Actual date of delivery:	15 March 2010
Type:	Report
Status & version:	Final
Author(s):	O. Caprotti et al.
Task responsible:	UGOT [1]
Other contributors:	

ABSTRACT

Annual report on activities carried out in the framework of the MOLTO EU project. This report is designed for Web publishing, for a broad public outside the consortium. It documents the main results obtained by the MOLTO project during the first two years of activity and promotes the objectives of the project.

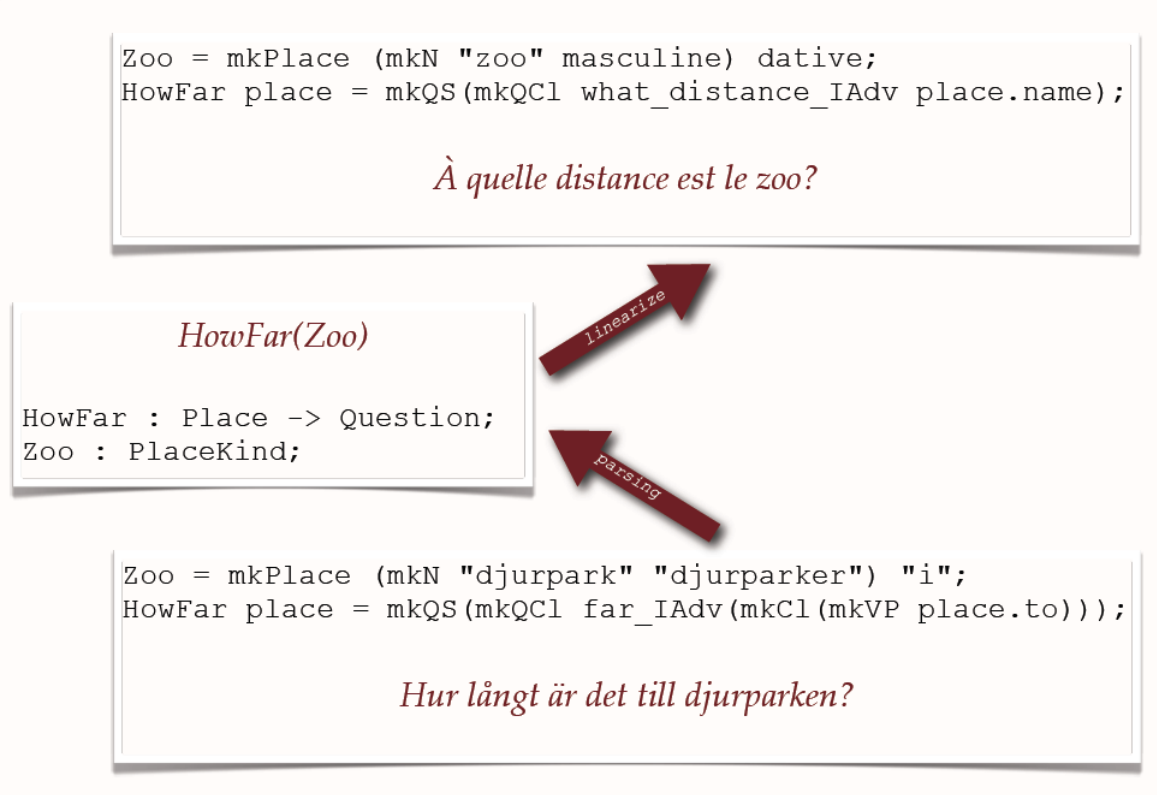
MOLTO's goal is to develop a suite of tools for translating texts between multiple languages in real time with high quality. MOLTO uses domain specific semantic grammars and ontology-based interlinguas implemented in GF [2] (Grammatical Framework), a grammar formalism where multiple languages are related by a common abstract syntax. Until now GF [2] has been applied in several small-

to-medium size domains, typically targeting up to ten languages, but during MOLTO we will scale this up in terms of productivity and applicability by increasing the size of domains and the number of languages.

MOLTO aims to make its technology accessible to domain experts who lack GF [2] expertise so that building a multilingual application will amount to just extending a lexicon and writing a set of example sentences. The most research-intensive parts of MOLTO are the two-way interoperability between ontology standards (such as OWL and RDF) and GF [2] grammars and the extension of rule-based translation by statistical methods. The OWL-GF [2] interoperability enables multilingual natural language based interaction with machine-readable knowledge while the statistical methods add robustness to the system when desired. MOLTO technology is released as open-source libraries for third-party translation tools and web pages and thereby fits into standard workflows.

1. Project Objective

The EU project MOLTO - Multilingual Online Translation, started on March 1, 2010 and will run until June 2013.The Consortium, comprising the universities of Gothenburg, Helsinki and Polytechnical Barcelona together with the industrial Bulgarian partner OntoText, has been enlarged by the addition of University of Zurich and of the Dutch Be Informed.



[3] MOLTO's multilingual translation tools use multilingual grammars based on semantic interlinguas and statistical machine translation to simplify the production of multilingual documents without sacrificing the quality. The interlinguas are designed to model domain semantics and are equipped with reversible generation functions: namely translation is obtained as a composition of parsing the source language and generating the target language.

An implementation of this technology is alread available in the Grammatical Framework, GF [2]. As a result of the MOLTO project work, GF [2] technologies are complemented by the use of ontologies, viewed as formalisms employed by the semantic web for capturing structural relations, and by methods of statistical machine translation (SMT) for improving robustness and extracting grammars from linguistic data.

MOLTO is committed to dealing with 15 languages, which includes 12 official

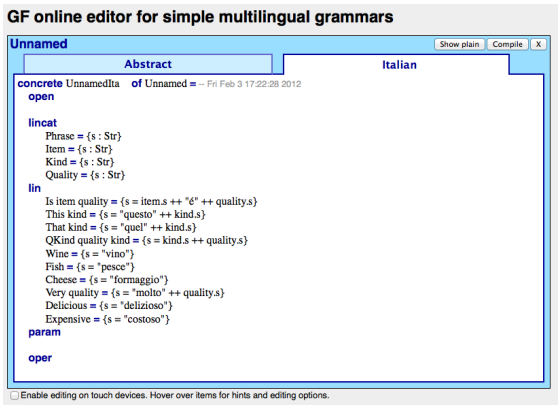
languages of the European
Union - Bulgarian, Danish, Dutch, English, Finnish, French, German, Italian, Polish, Romanian, Spanish, and Swedish - and 3 other languages - Catalan, Norwegian, and Russian. In addition, there is constant on-going work on creating new resource grammars, in particular Arabic, Farsi, Hebrew, Hindi/Urdu, Icelandic, Japanese, Latvian, Maltese, Portuguese, and Swahili. The coverage and accuracy of the GF [2] grammar library resource varies among the different languages and is documented on the web site of GF [2].

When comparing MOLTO to popular translation tools like Systran (Babelfish) and Google Translate, the main difference is the intended user of the tools: these tools target end-users of information whereas MOLTO targets producers of information.
By producers of information, MOLTO is able to handle well scenarios in which the language is constrained, as examples one may consider e-commerce sites, where products are often described with repeated linguistic expressions (e.g. Wikipedia articles, contracts, business letters, user manuals, and software localization), but even social networks often display usage of common phrases ("Happy birthday!" "I like it" "The hotel is located ...." "Your reservation is confirmed"). Ideally, MOLTO tools will enable publishers of websites to add multilinguality with little effort but most importantly with the certification that the meaning of the message conveyed stays unaltered across languages. MOLTO is also working on a multilingual semantic wiki .....

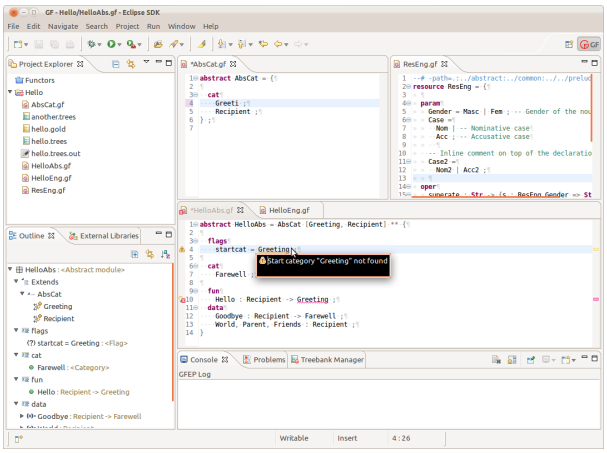
There is a well-known trade-off in machine translation: one cannot at the same time reach full coverage and full precision. In this trade-off, Systran and Google have opted for coverage whereas MOLTO opts for precision in domains with a well-understood codified language, either because it is of technical nature or because of common everyday usage.
The domains considered during the MOLTO project show a range of features of constrained natural languages: mathematical exercises and biomedical patents employ a technical and sophisticated jargon, whereas museum object descriptions use a language accessible to anybody.

2. Results

The expected final product of MOLTO is an open-source software suite of tools comprising a grammar development environment, an application programming interface and environment to assist the translators' workflow, and sample application grammar libraries for the domains of mathematical word problems, biomedical patents, and cultural artefacts.



Translation systems in MOLTO rely on multilingual grammars written in the GF [2] programming language. Until now, the development environments available to GF [2] grammarians consisted of a generic text editor, such as Emacs, used in combination with the GF [2] interactive command shell, and the online GF [2] documentation. This is a simple and effective environment for the experienced grammar developer. To better support less experienced grammar developers, one of the goals of the MOLTO project is to create an Integrated Development Environment for grammar development. The GF [2] Simple Editor (by Thomas Hallgren), an initial prototype of a web-based grammar development environment that offers the same core functionality as the traditional environment is now available at http://www.grammaticalframework.org/demos/gfse [4]. Its main features include grammar editing, grammar compilation, error detection, testing and visualization. Moreover, it enables the creation of web-based translation systems without installation of any software, as it is using web services to carry out compilation and interpretation tasks, and thus gives quick access to GF [2] to novice and occasional users. Intended scenario for this editor is in supporting fast testing and prototyping of example grammars in tutorial settings, for instance during teaching and demonstrating GF [2].



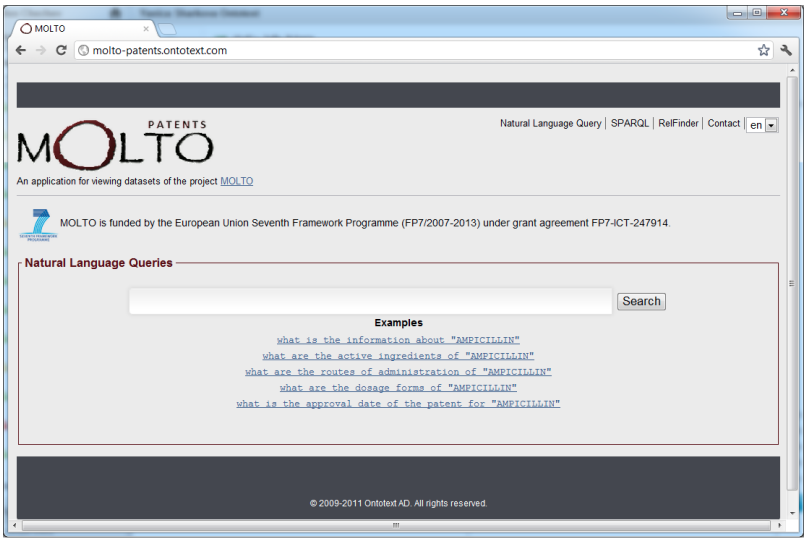
A different, more sophisticated high-level integrated development environment is based on the Eclipse platform and specifically tailors GF [2] grammar-writing. The GF [2] Eclipse plugin (by John Camilleri) currently features real-time syntax checking, automatic code formatting, import-aware auto-complete suggestions, cross-reference resolution, inline contextual documentation, "New Module" wizards, external library browsing, launch shortcuts to the GF [2] shell, and a visual tool for running treebank test suites. These new, powerful, time-saving development tools are aimed at both new users and GF [2] veterans alike. It is available online at http://www.grammaticalframework.org/eclipse/ [5] and at http://www.molto-project.eu/wiki/gf-eclipse-plugin [6].

Controlled natural languages are controlled subsets of natural languages, which are normally used in technical domains. The purpose of these languages is to reduce the complexity involved in natural languages, and to eliminate the ambiguity. The users of these languages are experts within their domain, and are trained to use these languages.

The MOLTO Phrasebook (by Aarne Ranta [7] et al.) is one such controlled natural language, whose domain is that of touristic phrases. It covers greetings and travel phrases such as "this fish is delicious", "how far is the airport from the hotel" in 17 languages. The translations show the kind of quality that can be hoped for when using a GF [2] grammar that can handle disambiguation in conveying gender and politeness, for instance from English to Italian. It is available both on the web from http://www.grammaticalframework.org/demos/phrasebook/ [8] and as a stand-alone, offiline Android application, the PhraseDroid, from http://tinyurl.com/7tyzvfd [9]. Screenshots of the mobile application are shown in the image on the side.

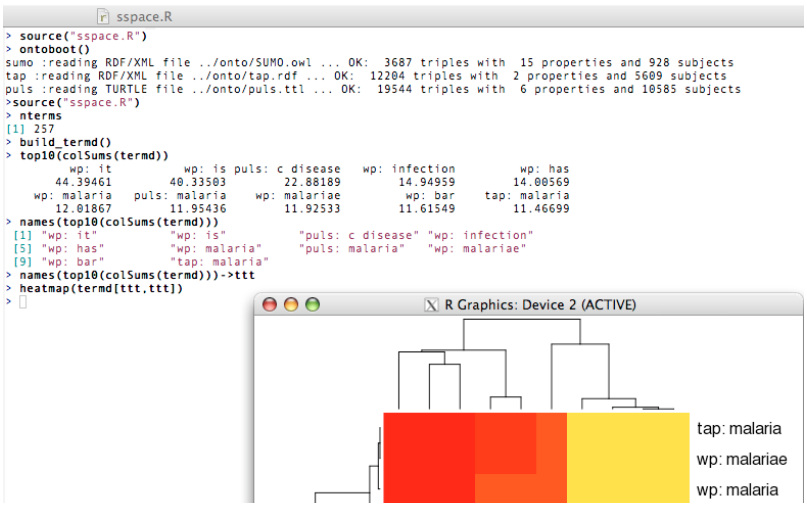
A different kind of controlled natural language is one that is used to command an interactive software system, for instance a computational engine such as Sage. The GFSage software application (by Jordi Saludes [10]) shows a command-line tool able to take commands in natural language, have them executed by Sage, and have the answers rendered in natural language too. The image on the side shows the web interface of Sage augmented by the MOLTO natural language command module. Note that this application demonstrates how a MOLTO library can add multimodality to a system originally developed with keyboard input/output as user interface. In fact, by piping the results to a speech engine, one can have the results aurally thus increasing accessibility of the computational systems to the visually impaired. The natural language interface relies on the Mathematical Grammar Library that can be tested at http://www.grammaticalframework.org/demos/minibar/mathbar.html [11] and

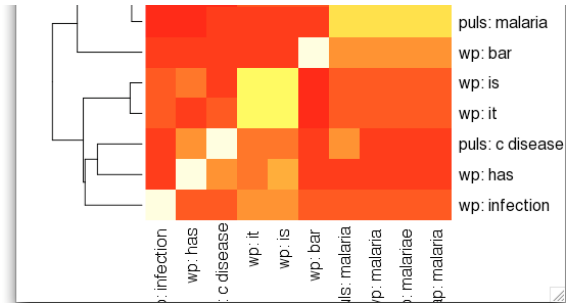
documentation on the GFSage module is available as deliverable http://tinyurl.com/78bh4ap [12] from the MOLTO wiki http://www.molto-project.eu/wiki/d62-prototype-comanding-cas [13].



To demonstrate the MOLTO Knowledge Reasoning Infrastructure, the Patent retrieval prototype (by Milen Chechev from Ontotext [14] in collaboration with the UPC [15] and the UGOT [1] teams), at http://molto-patents.ontotext.com [16], shows examples of queries in natural language to a set of patents in the pharmaceutical domain. Users can ask question in French and English like 'what are the active ingredients of "AMPICILLIN"', 'que sont les formes posologiques de "AMPICILLIN"'. The system is still under development: at present the online interface allows to browse the retrieved patents and returns the semantic annotations that explain why any particular patent has matched the user's criteria. Similar technology for knowledge retrieval is being applied also in the case of cultural heritage, namely with descriptions of artefacts from the museum of Gothenburg, in order to allow multilingual query and retrieval. For this task, an ad-hoc ontology has been created and its preliminary GF [2] application grammar can be tested by selecting "Painting.pgf" at http://www.grammaticalframework.org/demos/minibar/minibar.html [17].

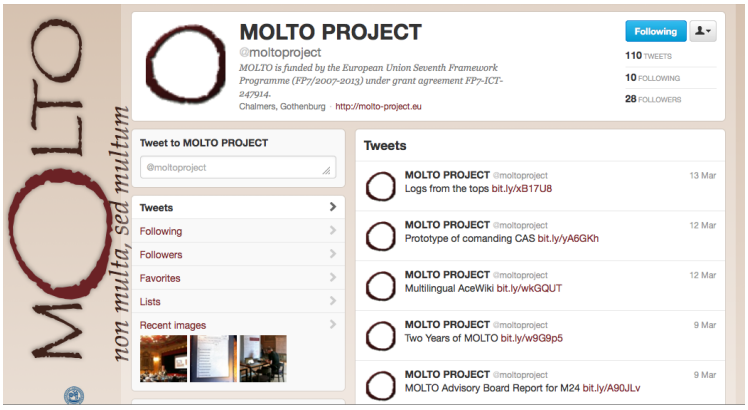
The MOLTO translation environment is being developed (by UHEL [18] with contributions of UGOT [1]) as a customization the GlobalSight translation system (www.globalsight.com [19]). The aim is to be



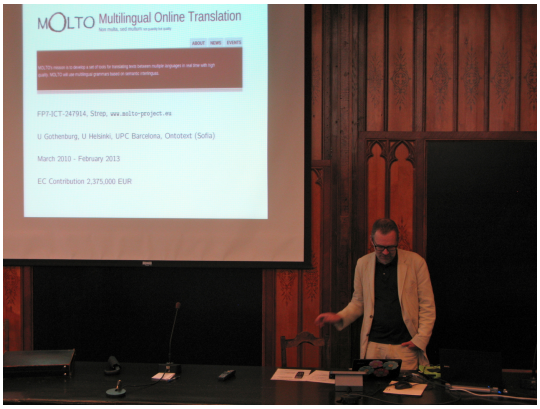


able to embed MOLTO translation tools to a third-party translation platform. MOLTO tools are designed with a focus only on translation. GlobalSight is an open source translation management platform, which provides the infrastructure needed in a professional translation workflow. More specifically, a MOLTO translation editor will be available on the side of conventional editors and be characterized by the possibility of fetching terms from the FactForge ontology via the TermFactory database, allowing to import and export terms in TermFactory. Terminology work is also supported by OntoR, an ontology extraction system (by Seppo Nyrkkö) implemented as s semi-supervised machine learning process, where new term dictionary candidates may be found in given text, by finding "closest matches" in previously known _ontologies_ (i.e. hierarchical vocabulary, term structure, usually industry or domain specific). A corpus-harvested new term can be _aligned_ with its closest matches in an prior existing term ontology. New term's functional and semantic environment is analyzed, and the feature variables extracted are compared to values of previously known terms. The user is given the supervision control to decide the best alignment match and thus refine the ontology incrementally. These tools are not yet ready for distribution but a preview can be seen during the project meetings' open days.

3. Dissemination



The main dissemination venues for the results of MOLTO are the MOLTO website and the project meetings. The website at www.molto-project.eu [20] makes available all the project's results and advertises news, deliverables, and events organized by the partners. It also archives all MOLTO publications, both delivered at international meeting as well as at internal workshops. The MOLTO news updates are posted as RSS feed suitable for aggregation by interested portals that is distributed by the MOLTO twitter feed [21] and via the MOLTO LinkedIn group [22].



MOLTO sponsored the GF [2] Summer School 2001, Frontiers of Multilingual Technologies during August 15-26, 2011 hosted by UPC [15] in Barcelona, Spain. The two weeks program included lectures from "Getting started with GF [2]", to "GF [2] application development" and "Resource grammar development" and was attended by around 20 participants from around the world. The use case studies of MOLTO were amply presented by members of the Consortium. On August,1 2011 Aarne Ranta [7] was invited to give a tutorial on GF [2] during CADE-the 23rd International Conference on Automated Deduction, in Wroclaw, Poland. The lecturing material "Grammatical Framework: A Hands-On Introduction" [23] is online. At the same meeting, Jordi Saludes [10] has presented the Mathematical Grammar Library during the affiliated workshop THedu'11, Computer Theorem Proving Components for Educational Software. "A Framework for Improved Access to Museum Databases in the Semantic Web" was presented during the meeting Recent Advances in Natural Language Processing (RANLP 2011), in September 2011, at Hissar, Bulgaria. Similar work, "Reason-able View of Linked Data for Cultural Heritage" was presented during The Third International Conference on SOFTWARE, SERVICES & SEMANTIC TECHNOLOGIES, also in September, 2011 in Bourgas, Bulgaria. The MOLTO project was presented at Tsukuba University and during the meeting "Digitization and E-Inclusion in Mathematics and Science 2012" (DEIMS2012) in February 2012 in Tokyo Japan by Olga Caprotti [24].
Two demonstrations of MOLTO prototypes on query and retrieval in the cultural heritage and in the patent domains have been accepted for presentation at the European Track of the World Wide Web 2012 conference. A paper on GF [2], "Smart Paradigms and the Predictability and Complexity of Inflectional Morphology", will also be presented at the conference of the European Association for Computational Linguistics in April 2012.

The list of conference papers funded by MOLTO can be retrieved under Publication [25] from the website.

Project meetings of MOLTO include always an open day with a program of presentations aimed at a general audience, the last MOLTO open days took place in Gothenburg on March 9, 2011 during the second project meeting, on September,2 2011 in Helsinki during the third project meeting, and on January,12 2012 in Gothenburg for the MOLTO-EEU kick off meeting.

4. Forthcoming

The project is looking forward to the final development phase especially with the addition of the new case studies, which will bring feedback to existing tools and ongoing work. In terms of events sponsored by MOLTO, the Third International Workshop on Free/Open-source Rule-based Machine Translation will take place in Gothenburg, Sweden, between 13-15 June 2012. Chair of the meeting is the MOLTO coordinator A. Ranta and local organization is managed by the MOLTO project manager. The fifth MOLTO project meeting will take place in September 2012 in The Netherlands in cooperation with the MONNET project. Stay tuned by subscribing to the MOLTO RSS feed or follow us on Twitter.

Source URL: http://www.molto-project.eu/wiki/living-deliverables/dx2-annual-public-report

Links:
[1] http://www.molto-project.eu/University of Gothenburg
[2] http://www.grammaticalframework.org
[3] http://www.molto-project.eu/sites/default/files/linearize-parse.png
[4] http://www.grammaticalframework.org/demos/gfse
[5] http://www.grammaticalframework.org/eclipse/
[6] http://www.molto-project.eu/wiki/gf-eclipse-plugin
[7] http://www.molto-project.eu/user/3
[8] http://www.grammaticalframework.org/demos/phrasebook/
[9] http://tinyurl.com/7tyzvfd
[10] http://www.molto-project.eu/user/6
[11] http://www.grammaticalframework.org/demos/minibar/mathbar.html
[12] http://tinyurl.com/78bh4ap
[13] http://www.molto-project.eu/wiki/d62-prototype-comanding-cas
[14] http://www.molto-project.eu/Ontotext AD
[15] http://www.molto-project.eu/Universitat Politècnica de Catalunya
[16] http://molto-patents.ontotext.com
[17] http://www.grammaticalframework.org/demos/minibar/minibar.html
[18] http://www.molto-project.eu/University of Helsinki
[19] http://www.globalsight.com
[20] http://www.molto-project.eu
[21] http://twitter.com/moltoproject
[22] http://www.linkedin.com/groups?gid=3703935
[23] http://www.grammaticalframework.org/gf-tutorial-cade-2011
[24] http://www.molto-project.eu/user/4

[25] http://www.molto-project.eu/biblio