MOLTO

D1.2 Progress Report

http://www.molto-project.eu

Contract No.:	FP7-ICT-247914
Project full title:	MOLTO - Multilingual Online Translation
Deliverable:	D1.2. Progress Report
Security (distribution level):	Confidential
Contractual date of delivery:	M7
Actual date of delivery:	18 Nov 2010
Type:	Report
Status & version:	Final
Author(s):	Olga Caprotti and Aarne Ranta et al.
Task responsible:	UGOT
Other contributors:	All

ABSTRACT

Progress report for the first semester of the MOLTO project lifetime, 1 Mar 2010 - 30 Sep 2010. The source URL for this document is http://www.molto-project.eu/wiki/d1.2.

Table of Contents

1 Publishable Summary

1.1 Project context and objectives

The project MOLTO - Multilingual Online Translation, started on March 1, 2010 and will run for 36 months. It promises to develop a set of tools for translating texts between multiple languages in real time with high quality. MOLTO will 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 based on 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 provided by GF [2], Grammatical Framework. GF technologies in MOLTO are complemented by the use of ontologies, such as used in the semantic web, and by methods of statistical machine translation (SMT) for improving robustness and extracting grammars from 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 on-going work on at least Arabic, Farsi, Hebrew, Hindi/Urdu, Icelandic, Japanese, Latvian, Maltese, Portuguese, Swahili, Tswana, and Turkish.

Tools like Systran (Babelfish) and Google Translate are designed for consumers of information, but MOLTO will mainly target the producers of information. Hence, the quality of the MOLTO translations must be good enough for, say, an e-commerce site to use in translating their web pages automatically without the fear that the message will change. Third-party translation tools, possibly integrated in the browsers, let potential customers discover, in their preferred language, whether, for instance, an e-commerce page written in French offers something of interest. Customers understand that these translations are approximate and will filter out imprecision. If, for instance, the system has translated a price of 100 Euros to 100 Swedish Crowns (which equals 10 Euros), they will not insist to buy the product for that price. But if a company had placed such a translation on its website, then it might be committed to it.

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 language. Three such domains will be considered during the MOLTO project: mathematical exercises, biomedical patents, and museum object descriptions. The MOLTO tools however will be applicable to other domains as well. Examples of such domains could be e-commerce sites, Wikipedia articles, contracts, business letters, user manuals, and software localization.

1.2 Main results achieved so far

A few results have been already achieved during the first semester of the project's lifetime. Two applications of the MOLTO translation web services are online on the project web pages:

1. The travel phrasebook [4] translates sentences to 14 different languages and shows some of the major end-user features available to MOLTO users: predictive typing and JavaScript-based GUI. Predictive typing prompts the user with the next available choices

mandated by the underlying grammar and offers quasi-incremental translations of intermediate results from words or complete sentences. JavaScript-based GUI using off-the-shelf functions can be readily deployed on any device where a browser is available.

2. The MOLTO KRI [5], Knowledge Reasoning Infrastructure, demonstrates the possibility of adding a natural language query language to retrieve answers from an OWL database. In this way, a query like Give me information about all organizations located in Europe is interpreted as the machine understandable SPARQL statement:

```
SELECT DISTINCT ?organization ?organization_label
       WHERE    {    ?organization    .    ?organization
              ?organizationloc. ?organizationloc
              "Europe"
             . ?organization ?organization_label
             . }
```

On the more technical level, MOLTO released:

- First version of the Python plugin for GF (based on the planned C plugin). The plugin makes GF primitives available from the Natural Language Tool [6], an open source collection of Python modules for research and development in natural language processing and text analytics, with distributions for Windows, Mac OSX and Linux.

Work has continued towards the forthcoming release of:

- Version 3.2 of GF, which features updates of the pgf format, complete type checker for dependent types, exhaustive generation of ASTs via lambda prolog, support for probabilities in the abstract syntax, random generation and parse results guided by probability, and example based grammar generation.
- Urdu resource grammar library and Turkish morphology.

1.3 Expected final results and their potential impact and use

The expected final product of MOLTO is a software toolkit made available via the MOLTO website. It will consist in a family of open-source software products:

- a grammar development tool, available as an IDE and an API, to enable the use as a plug-in to web browsers, translation tools, etc, for easy construction and improvement of translation systems and the integration of ontologies with grammars
- a translator's tool, available as an API and some interfaces in web browsers and translation tools
- a grammar library for linguistic resources
- a grammar library for the domains of mathematics, patents, and cultural heritage

These tools will be portable to different platforms as well as generally portable to new domains and languages. By the end of the project, MOLTO expects to have grammar resource libraries for 18 languages, whereas MOLTO use cases will target between 3 and 15 languages.

The main societal impact of MOLTO will be on contributing to a new perception for the possibilities of machine translation, moving away from the idea that domain-specific high-quality translation is expensive and cumbersome. MOLTO tools will change this view by radically lowering the effort needed to provide high-quality scoped translation for applications where the content has enough semantic structure.

1.4 Public website

The MOLTO website at http://www.molto-project.eu publishes the results, the news and all information related to the project. In addition, a Twitter feed is also available at http://twitter.com/moltoproject.

2 Core of the report

2.1 Project objectives for the period

The project objectives for the first semester focus on establishing the grounds for cooperation among the partners, hence three deliverables contribute to refine the goals of the project:

1. Setup of the website and definition of the work plan,
2. Definition of the dissemination strategy,
3. MOLTO test criteria, methods and schedule.

The first version of the MOLTO web services, due at Month 3 is the major concrete target for the period and demonstrates the technologies underlying the ideas of the project.

2.2 Work progress and achievements during the period

2.2.1 WP2 Grammar Developer's Tools - Month 6

The Grammarian' Tools include tools for using the GF grammar compiler and the Resource Grammar Library. In the first 6 months of MOLTO, we have worked on consolidating the compiler and the Library API, and also experimenting with the example-based grammar writing technique.

Clearly significant results include:

- Milestone 1, 15 languages in the library. Due September 2010; reached December 2009.
- Workflow for example-based grammar writing and estimated engineering effort: reported as a part of D10.2 [9]
- GF plugin to Python NLTK [10]
- GF syntax highlight plugin to XCode programming environment.
- Integrating probabilities with GF [2] grammars.
- Release of GF 3.1.6 in April 2010; GF 3.2 forthcoming before end of 2010.

No deviations from Annex I and the use of resources was as planned.

2.2.2 WP3 Translator's Tools - M6

Not yet started.

2.2.3 WP4 Knowledge Engineering - Month 6

During the first period we managed to clarify the needs for knowledge representation infrastructure of the case studies and software tools in MOLTO. We have also circulated a questionnaire describing the structured data sets which are expected to be of benefit for the project. Based on this information, we proceeded with deploying the knowledge representation infrastructure, which is now in place and accessible to the partners. It will be further described in D4.1 Knowledge Representation Infrastructure.
The second major direction during this period was the undoubtedly challenging grammar to ontology interoperability. For this we have chosen a quasi-exhaustive knowledge base of important named entities in the world and some relations between them. It is encoded according to PROTON – a basic-upper level ontology with about 300 classes of named entities. The first goal set for this interoperability was a transformation of questions expressed in natural language towards a formal query language – SPARQL. For this purpose, and on the basis of the ontology and the entities in the knowledge base, we have manually created a corpus of 500 sentences. This corpus is being used for development of the GF grammars handling the natural language questions and also for evaluation of the coverage of the grammars over this language space. After an initial grammar handling questions to the knowledge base has been developed for a subset of the English language, we have created a transformation function, rendering GF sentence trees to SPARQL queries. In order to show these initial results, we have developed a natural language based search interface over the knowledge base, with automatic suggestion of possible continuation of the questions, which is featured on the MOLTO website. The results of these questions are one or two-dimensional tables of entities, where each row is an individual “answer”.

2.2.4 WP5 Statistical and Robust Translation - M6

WP5 is planned to span from Month 7 to Month 30, but it is being conditioned by the delay on the Patents data. So, there is already some ongoing work we detail in the following.

Towards Milestone MS7 (M24): First prototypes of hybrid combination models.

Most of the objectives of the package depend on the compilation of the Patents corpus. Even the languages of study depend on the data that the new partner provides. In order to compensate the delay due to this both in WP5 and mainly in WP7 we started working here on hybrid approaches. The methodology now is to develop hybrid methods in a way independent of the domain and data sets used, so that they can be later adapted to patents.
At the moment, we are able to obtain phrases and alignments from a GF-generated synthetic corpus. This is a first step for the hard integration of both paradigms, and also for the soft integration methods led by SMT. We are currently going deeper into the latter, as it is a domain independent study.

Towards Deliverable D51 (M18): Description of the final collection of corpora.

Bilingual corpora are needed to create the necessary resources for training/adapting statistical MT systems and to extend the grammar-based paradigm with statistical information. We will compile and annotate general-purpose large bilingual and monolingual corpora for training basic SMT

systems. At the moment, we have compiled and annotated the European Parliament corpus for English and Spanish. Languages will probably finally be English, German, and Spanish or French, so as soon as this is confirmed the final general-purpose corpus can be easily compiled. The depth of the annotation will depend on the concrete languages and the available linguistic processors.
On the other hand, domain specific corpora will be needed to adapt the general purpose SMT system to the concrete domain of application in this project (Patents case study, WP7. We cannot build the final corpus, but some of the MOLTO members have joined the IRF so that a collection of patents data is available for individual research purposes. This has allowed to compile a preliminary parallel corpus on which we can start shortly to build a domain GF grammar and to develop a first pure SMT domain-adapted translator.

2.2.5 WP6 Case Study: Mathematics - Month 6

Working towards deliverable D6.1:

1. Refactor prior code (WebALT grammars) into a separate module for each OpenMath Content Dictionary (CD).
2. Adapt said code to work with current GF resource libraries (3.1)
3. Test compilation of OpenMath layer for: English, Catalan, French, Italian, Spanish, German, Swedish.

Clearly significant results include:

4. OpenMath layer of D6.1 compiles correctly for said languages (English, Catalan, French, Italian, Spanish, German, Swedish)

WP6 was moved ahead to start on Month 5 (instead of 7) to buy time for WP5, which will be delayed due to lack of data.

2.2.6 WP7 Case Study: Patents - M6

WP7 was scheduled to start in Month 4. But the WP leader site, Matrixware, left the MOLTO Consortium during Month 3. We have had negotiations with replacing partners, and expect them to be concluded before November 2010 (Month 9 of MOLTO). Then we expect to start WP7 no later than January 2011 (Month 11).
While the delay is with several months, it need not imply great changes in the actual work. The original reason to start in Month 4 was to give the Matrixware site something to work on, since they we not highly involved in the other WP's. The new partner is expected to get started immediately, and the WP will also profit from the fact that some other MOLTO tools have become available (grammarian's tools from WP2 and grammar-statistics combination from WP5).
The actual work plan for WP7 may change in accordance with the preferences of the new partner. This will happen within the limits of the budget originally allocated to this WP.

2.2.7 WP8 Case Study: Cultural Heritage - M6

WP8 will start in Month 12, so no work can be reported yet.

2.2.8 WP9 User Requirements and Evaluation - M6

The objectives are to

(i) collect user requirements for the use cases, grammar development IDE and translation tools;
(ii) define criteria for evaluating the translation and the tools;
(iii) define diagnostic and evaluation corpora;
(iv) perform continuous quality control and monitor progress through iterative evaluation.

Deliverable 9.1 was the main target for this semester. Work has concentrated on setting up a local environment where the evaluation can take place and be organized. This included installed a project wiki, open to all partners, and to test technologies such as GF, OWLIM, SESAME and the Stanford Parser. The local evaluation platform is populated by generic and project specific translation quality evaluation tools: both statistical (BLEU) and language technology specific. A MOLTO Evaluation Cookbook[1] tracks on the wiki the current evaluation strategy in MOLTO.
The unavailability of the patent corpus has hindered the planning of evaluation for WP7 and is being postponed until the data becomes ready.
Lauri Alanko has worked on implementing a C language runtime for PGF. A first announcement was submitted at http://www.molto-project.eu/node/968. Currently, the implementation work continues on linearization. A first version of the linearization code is ready and under testing.
Inari Listenmaa has produced a Finnish version of the Mathematics vocabulary. She has started a study of ontology based vocabulary extraction from OntoText FactForge data as her MA topic.
Seppo Nyrkkö has been working on a method and tools for ontology based vocabulary extraction from free text. The method uses statistical wide-coverage parsers (e.g. the Stanford Parser) and matching parse graphs with existing seed ontologies. Work on a MOLTO vocabulary editor based on the TermFactory ontology based terminology tools is under way. The vocabulary editor is planned to appear as a tab plugin to the MOLTO translation editor by Krasimir Angelow.

2.2.9 WP10 Dissemination and Exploitation - M6

The stated objectives of this workpackage are to:

1. create a MOLTO community of researchers and commercial partners;
2. make the technology popular and easy to understand through light-weight online demos;
3. apply the results commercially and ensure their sustainability over time through synergetic partnerships with the industry.

The first task has been to setup the website for MOLTO, with information about MOLTO's technology and potential (D10.2, UGOT and Ontotext) targeted to research, industry and users. Bibliographic information on GF, on SMT and on knowledge retrieval is kept up-to-date and includes tutorial presentations delivered during the MOLTO workshops. The web site includes a News section with frequent informal posts on internal progress and plans and encouraging community contributions in the form of comments. More light newsflash items are published using the MOLTO Twitter feed. A specific section is devoted to Frequently Asked Questions and can be collaboratively maintained by the MOLTO partners.
This workpackage was responsible for two deliverables during the first semester:

[1] https://kitwiki.csc.fi/twiki/bin/view/MOLTO/EvaluationCookbook

- Dissemination plan with monitoring and assessment,
- MOLTO web service, first version.

The dissemination plan [21] can be accessed on the consortium-restricted pages and will be amended during the project's lifetime if needed. The project has been presented in a few meetings and international events, most notably at LREC2010, EAMT2010, and ACL2010.
The first version of the MOLTO web service consists of an online demonstration of a multilingual travel phrasebook, described online in Deliverable D10.2 [22].

2.3 Project management during the period

Management tasks carried out during the first semester of MOLTO finalized the administrative and organizational setup of the project. The website for the project is online at http://www.molto-project.eu. The Consortium Agreement had been signed before the Grant Agreement in December 2009. The work plan for MOLTO (Deliverable D1.1) is hosted on the wiki pages on the website.
The Steering Group of MOLTO, elected during the Kick-Off meeting, presently consists of voting members A. Ranta (UGOT, Chair), J. Saludes (UPC), B. Popov (Onto), and L. Carlson (UH). The Steering group held monthly calls to discuss the project's progress and recorded the minutes on the website. The MOLTO Advisory Board has been established, with members Prof. Stephen Pulman (Computing Laboratory Oxford) and Keith Hall (Google Research Zurich).
The project had to face a major challenge with the dissolution of the Consortium partner company Matrixware. Upon learning of this, the Coordinator informed the Commission and proceeded to formalize the dismissal of Matrixware, that left the Consortium at the end of Month 2, on April 23, 2010. In order to be able to carry out the tasks set forward in the MOLTO DoW, with minor disruption, MOLTO started negotiations with EPO, European Patent Office, to incorporate it as new member of the MOLTO Consortium. This process has taken a long time, about 3 months and we expect to learn their final decision at the end of October. In case of positive outcome, then EPO will step in and we expect little changes to the original work plan. In case of negative outcome, then MOLTO will discuss changing the work plan for Workpackage 7, the Patent Case Study, possibly to a different domain. MOLTO partners have been approached by several interested parties with use case study domains that could be suitable test beds for the tools developed during the project, these potential partners will be approached first.
The original work plan has been slightly modified to cope with changes in the Consortium, mainly by shifting the start of two workpackages. The loss of Matrixware affected the MOLTO activities scheduled for Workpackage 7: Case Study Patents (led by Mxw) from Month 4 to Month 30. The major task that has been put on hold is the preparation of a parallel patent corpus (Mxw) to fuel the training of statistical MT (UPC). The work on Workpackage 7 will start as soon as the Consortium situation clarifies. UPC, the most directly affected partner (whose tasks depended on the work of Mxw), has begun the work on Workpackage 6: Case Study Mathematics in Month 5 instead of Month 7.
Two project meetings have been organized, in Barcelona, 8-10 March 2010, and in Varna 10-12 September 2010. A bilateral meeting, between UH and UGOT, has been organized in Helsinki on 5-6 May 2010.

3 Deliverables and milestones tables

3.1 Deliverables for Period M1-M6

ID		Due date	Diss. level	Nature	Publication
D1.1	Workplan for MOLTO	1 Apr, 2010	Co	R	[28]
D10.1	Dissemination plan, with monitoring and assessment	1 Jun, 2010	Co	R	[30] [29]
D10.2	MOLTO web service, first version	1 Jun, 2010	P	P	[9][31]
D9.1	MOLTO test criteria, methods and schedule	1 Sept, 2010	P	R	[32]

3.2 Milestones for Period M1-M6

ID	Title	Due date
MS1	15 Languages in RGL [36]	1 Sept, 2010
MS2	Knowledge Representation Infrastructure [37]	1 Sept, 2010

4 Use of the resources

Not available for midterm reporting.

5 Financial statements

Not available for midterm reporting.

Links:
[2] http://www.grammaticalframework.org
[4] http://www.molto-project.eu/demo/phrasebook
[6] http://www.nltk.org/
[9] http://www.molto-project.eu/node/1044
[10] http://www.molto-project.eu/node/959
[12] http://www.molto-project.eu/node/28
[14] http://www.molto-project.eu/node/880
[17] http://www.molto-project.eu/node/874
[19] http://www.molto-project.eu/node/827
[21] http://www.molto-project.eu/wiki/d10.1
[22] http://www.molto-project.eu/wiki/d10.2
[28] http://www.molto-project.eu/d1.1

[29] http://www.molto-project.eu/d10.1
[30] http://www.molto-project.eu/node/1030
[31] http://www.molto-project.eu/d10.2
[32] http://www.molto-project.eu/node/914
[36] http://www.molto-project.eu/node/1061
[37] http://www.molto-project.eu/node/1062