# Multilingual Online Translation

## Annual report 2010 - 2011

MOLTO

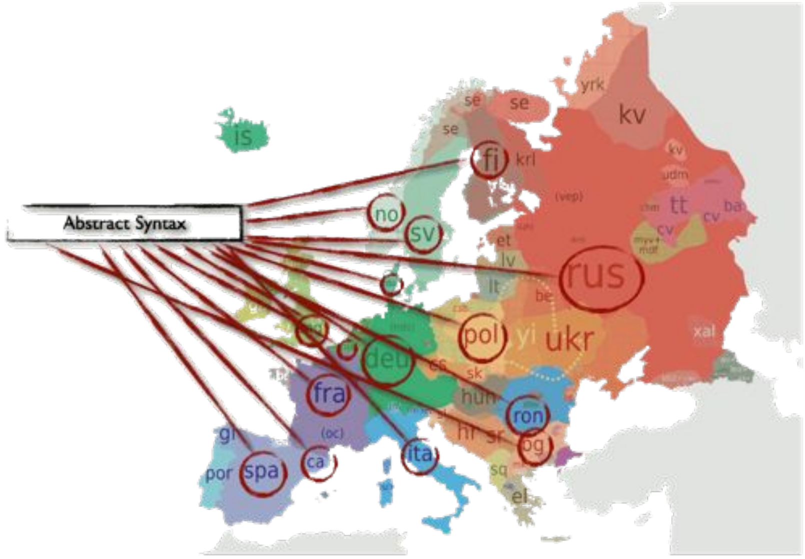
## Non multa, sed multum

## Contents

MOLTO's goal is to develop a set 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 (Grammatical Framework), a grammar formalism where multiple languages are related by a common abstract syntax. GF has been applied in several small-to-medium size domains, typically targeting up to ten languages but MOLTO 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 the technology accessible for domain experts without GF expertise and to reduce the effort needed for building a translator 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 (OWL) and GF grammars, and the extension of rule-based translation by statistical methods. The OWL-GF interoperability will enable multilingual natural-language-based interaction with machine-readable knowledge while the statistical methods will add robustness to the system when desired.

MOLTO technology will be released as open-source libraries for standard translation tools and web pages and thereby fit into standard workflows.



MOLTO's mission is 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.

# Project Objective

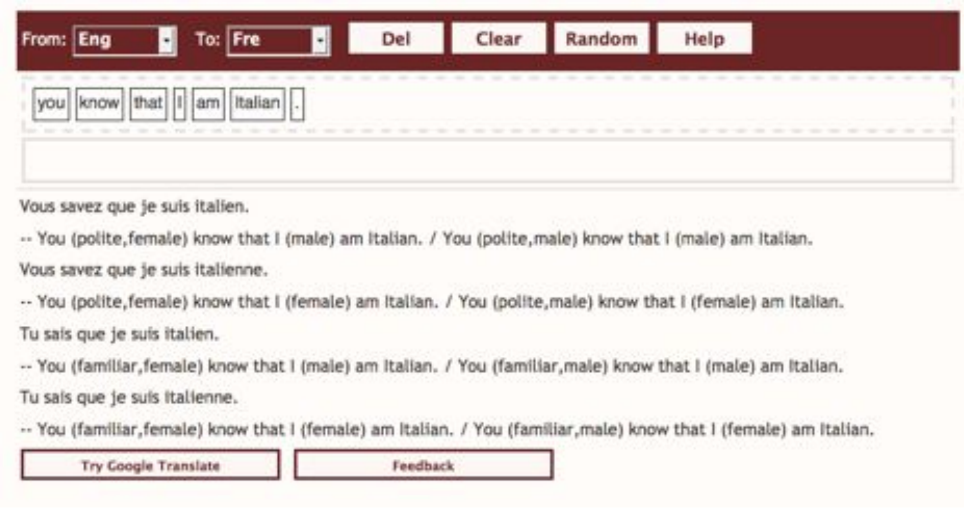
The project MOLTO - Multilingual Online Translation, started on March 1, 2010 and will run for 36 months. It promises to develop a set of tools for translating texts between multiple languages in real time with high quality. MOLTO will use multilingual grammars based on semantic interlinguas and statistical 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.

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.

# Results



Vous savez que je suis italien.

-- You (polite,female) know that I (male) am Italian. / You (polite,male) know that I (male) am Italian.

Vous savez que je suis italienne.

-- You (polite,female) know that I (female) am Italian. / You (polite,male) know that I (female) am Italian.

Tu sais que je suis italien.

-- You (familiar,female) know that I (male) am Italian. / You (familiar,male) know that I (male) am Italian.

Tu sais que je suis italienne.

-- You (familiar,female) know that I (female) am Italian. / You (familiar,male) know that I (female) am Italian.

MOLTO Phrasedroid

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

1. The travel phrasebook[1] 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[2], 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
        . }
```

Moreover it is now equipped with an interface in Swedish which demonstrates how the same knowledge base of facts stored in English can be queried in a different language.

A pre-release of a web-based grammar editor[3] for creating and compiling GF application grammars in the cloud can be tested online. It is designed to assist novel authors of GF grammars for instance by prompting the writer with prefilled templates for each new concrete language.  The resulting application grammars can then be compiled directly online on

The expected final product of MOLTO is an open-source software toolkit consisting in:

- grammar development tool, as an IDE and an API, to allow 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
- translator’s tool, as an API and some interfaces in web browsers and translation tools
- grammar libraries for linguistic resources, and for the domains of patents, mathematics, and cultural heritage

[1] http://www.grammaticalframework.org/demos/phrasebook/

[2] http://molto.ontotext.com

[3] http://grammaticalframework.org/demos/gfse

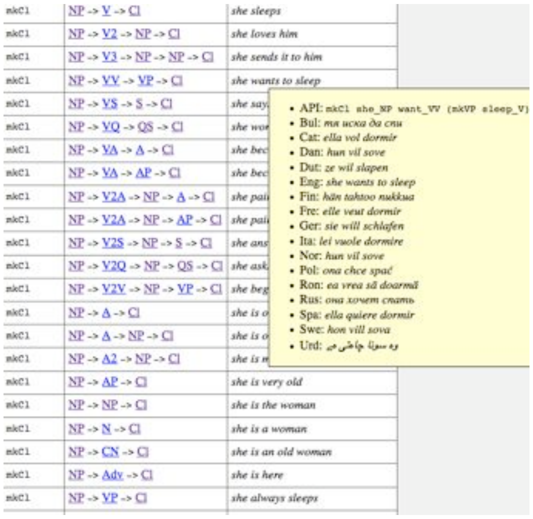
the server to web applications in javascript. Since all the workflow happens in the cloud, the authors do not have to install any software on their machines, with the added advantage of accessing the latest version of the libraries maintained by the developers.

On the more technical level, MOLTO released GF version 3.2 with simpler installation, runtime type checker and parser for dependant types, improved type errors reporting, probabilities in the abstract syntax, and example based grammar generation. The grammar API is now multilingual. New languages in the resource grammar library include Urdu, Amharic and complete morphology for Turkish and Punjabi.

The project also released the first version of the Python plugin for GF that makes GF primitives available from the Natural Language Toolkit, 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. Additionally, a Java runtine interprer for PGF grammars is operational for parsing and linearization and used on the andoid application of the phrasebook, the Phrasedroid. Finally, the PGF native C library, named "libpgf", is currently still under development. Basic linearization is already working, but both the interface and implementation require some further refinement. The ultimate goal of the native C library is to provide a full-fledged industrial-strength library for operating with PGF grammars, so that there are no longer any technical limitations to prevent the adoption of GF technology. In practice, "industrial-strength" means that the library should be embeddable, portable, lightweight, efficient, and robust.

The first experiments with statistical engines for translation have been presented at the MOLTO events as a first step towards the hybridization with GF. From the GF part, some preliminary results discuss the usage of GF to produce synthetic phrase alignments; also the methodology to extract high quality alignments from the domain corpora is being developed. The GF parser has been also adapted to deal robustly with this general domain corpora. Finally, the MOLTO workshop "GF meets SMT" (Gothenburg, November 2010) served the UPC and UGOT teams to brainstorm and discuss on the main hybridization strategies that will be carried out in the near future.

The MOLTO project plans to test its approach in three case studies: mathematical exercises, museum artifacts descriptions and biomedical patents. The mathematical case study is the most advanced test bed and covers mathematical expressions, following OpenMath, in 10 different languages.



| | | |
|---|---|---|
| mkCl | NP -> V -> Cl | she sleeps |
| mkCl | NP -> V2 -> NP -> Cl | she loves him |
| mkCl | NP -> V3 -> NP -> NP -> Cl | she sends it to him |
| mkCl | NP -> VV -> VP -> Cl | she wants to sleep |
| mkCl | NP -> VS -> S -> Cl | she say |
| mkCl | NP -> VQ -> QS -> Cl | she wor |
| mkCl | NP -> VA -> A -> Cl | she bec |
| mkCl | NP -> VA -> AP -> Cl | she bec |
| mkCl | NP -> V2A -> NP -> A -> Cl | she pai |
| mkCl | NP -> V2A -> NP -> AP -> Cl | she pai |
| mkCl | NP -> V2S -> NP -> S -> Cl | she ans |
| mkCl | NP -> V2Q -> NP -> QS -> Cl | she ask |
| mkCl | NP -> V2V -> NP -> VP -> Cl | she beg |
| mkCl | NP -> A -> Cl | she is o |
| mkCl | NP -> A -> NP -> Cl | she is o |
| mkCl | NP -> A2 -> NP -> Cl | she is m |
| mkCl | NP -> AP -> Cl | she is very old |
| mkCl | NP -> NP -> Cl | she is the woman |
| mkCl | NP -> N -> Cl | she is a woman |
| mkCl | NP -> CN -> Cl | she is an old woman |
| mkCl | NP -> Adv -> Cl | she is here |
| mkCl | NP -> VP -> Cl | she always sleeps |

- API: mkCl she_NP want_VV (mkVP sleep_V)
- Bul: *тя иска да спи*
- Cat: *ella vol dormir*
- Dan: *hun vil sove*
- Dut: *ze wil slapen*
- Eng: *she wants to sleep*
- Fin: *hän tahtoo nukkua*
- Fre: *elle veut dormir*
- Ger: *sie will schlafen*
- Ita: *lei vuole dormire*
- Nor: *hun vil sove*
- Pol: *ona chce spać*
- Ron: *ea vrea să doarmă*
- Rus: *она хочет спать*
- Spa: *ella quiere dormir*
- Swe: *hon vill sova*
- Urd: وہ سونا چاھتی ہے

MOLTO generalizes Controlled Languages to Multilingual Controlled Language Systems supporting also ambiguities

# Dissemination

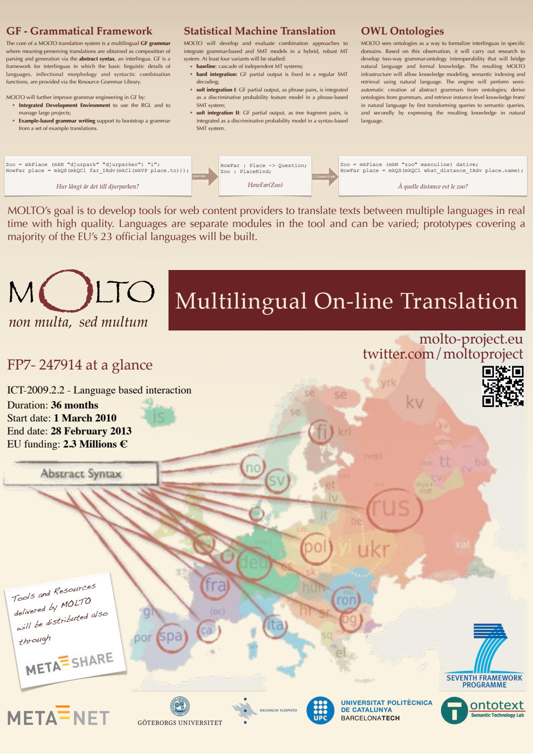
The MOLTO website was setup to popularize the MOLTO technologies and to help create a community of researchers and commercial partners. It makes available all the project's results and advertises the meetings and events organized by the partners. Every presentation delivered at international meeting as well as at internal workshops is archived on the website. The MOLTO news updates are posted as RSS feed suitable for aggregation by interested portals. Informal newsflash items are published using the MOLTO Twitter feed.

The project has been presented with the travel phrasebook demo and accompanying poster at a number of international conferences such as EAMT2010, ACL2010, SLTC2010, and FreeRMBT2010. MOLTO was also presented at regional meetings such at the Jornada sobre la Indústria de la Traducció entre Llengües Romàniques, held in September 2010 in València. Ranta delivered a tutorial on GF at LREC 2010 and one on a specific GF application to Multilingual Controlled Languages at CNL 2010.

Press coverage for MOLTO included, among articles in papers, also an interview at the Bulgarian National Radio a few days after the start of the project's lifetime.

The MOLTO project has also become part of the META-Share alliance of META-NET with the intent to share the linguistic resources produced during its lifetime.

Three project meetings have been held so far, in Barcelona in March 2010, in Varna in September 2010 and the yearly meeting in Gothenburg in March 2011. The meetings always include an Open Day with presentations aimed at the general audience.



**GF - Grammatical Framework**

The core of a MOLTO translation system is a multilingual **GF grammar** where meaning-preserving translations are obtained as composition of parsing and generation via the **abstract syntax**, an interlingua. GF is a framework for interlinguas in which the basic linguistic details of languages, inflectional morphology and syntactic combination functions, are provided via the Resource Grammar Library.

MOLTO will further improve grammar engineering in GF by:

- **Integrated Development Environment** to use the RGL and to manage large projects;
- **Example-based grammar writing** support to bootstrap a grammar from a set of example translations.

**Statistical Machine Translation**

MOLTO will develop and evaluate combination approaches to integrate grammar-based and SMT models in a hybrid, robust MT system. At least four variants will be studied:

- **baseline**: cascade of independent MT systems;
- **hard integration:** GF partial output is fixed in a regular SMT decoding;
- **soft integration I**: GF partial output, as phrase pairs, is integrated as a discriminative probability feature model in a phrase-based SMT system;
- **soft integration II**: GF partial output, as tree fragment pairs, is integrated as a discriminative probability model in a syntax-based SMT system.

**OWL Ontologies**

MOLTO sees ontologies as a way to formalize interlinguas in specific domains. Based on this observation, it will carry out research to develop two-way grammar-ontology interoperability that will bridge natural language and formal knowledge. The resulting MOLTO infrastructure will allow knowledge modeling, semantic indexing and retrieval using natural language. The engine will perform semi-automatic creation of abstract grammars from ontologies; derive ontologies from grammars, and retrieve instance level knowledge from/in natural language by first transforming queries to semantic queries, and secondly by expressing the resulting knowledge in natural language.

```
Zoo = mkPlace (mkN "djurpark" "djurparker") "i";
HowFar place = mkQS(mkQCl far_IAdv(mkCl(mkVP place.to)));
```

*Hur långt är det till djurparken?*

parse

```
HowFar : Place -> Question;
Zoo : PlaceKind;
```

*HowFar(Zoo)*

linearize

```
Zoo = mkPlace (mkN "zoo" masculine) dative;
HowFar place = mkQS(mkQCl what_distance_IAdv place.name);
```

*À quelle distance est le zoo?*

MOLTO's goal is to develop tools for web content providers to translate texts between multiple languages in real time with high quality. Languages are separate modules in the tool and can be varied; prototypes covering a majority of the EU's 23 official languages will be built.

MOLTO
*non multa, sed multum*

Multilingual On-line Translation

molto-project.eu
twitter.com/moltoproject

FP7- 247914 at a glance

ICT-2009.2.2 - Language based interaction
Duration: **36 months**
Start date: **1 March 2010**
End date: **28 February 2013**
EU funding: **2.3 Millions €**

*Tools and Resources delivered by MOLTO will be distributed also through*

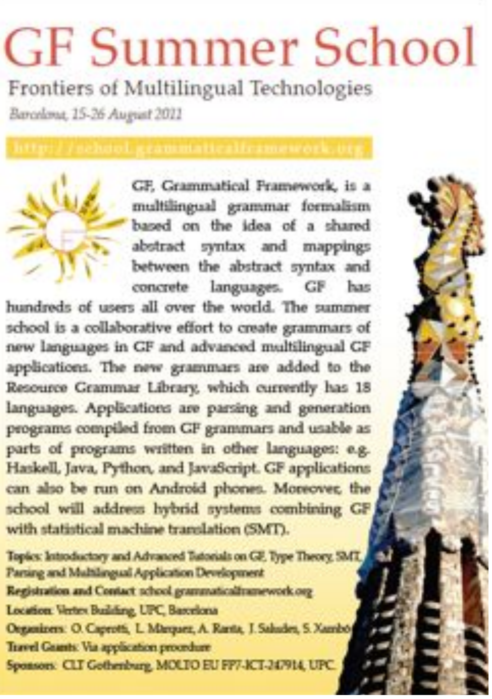
META-SHARE

SEVENTH FRAMEWORK PROGRAMME

META-NET
GÖTEBORGS UNIVERSITET
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
ontotext Semantic Technology Lab

Tools like Systran (Babelfish) and Google Translate are designed for consumers of information, but MOLTO will mainly target the producers of information

http://molto-project.eu



GF Summer School

Frontiers of Multilingual Technologies

*Barcelona, 15-26 August 2011*

http://school.grammaticalframework.org

GF, Grammatical Framework, is a multilingual grammar formalism based on the idea of a shared abstract syntax and mappings between the abstract syntax and concrete languages. GF has hundreds of users all over the world. The summer school is a collaborative effort to create grammars of new languages in GF and advanced multilingual GF applications. The new grammars are added to the Resource Grammar Library, which currently has 18 languages. Applications are parsing and generation programs compiled from GF grammars and usable as parts of programs written in other languages: e.g. Haskell, Java, Python, and JavaScript. GF applications can also be run on Android phones. Moreover, the school will address hybrid systems combining GF with statistical machine translation (SMT).

**Topics:** Introductory and Advanced Tutorials on GF, Type Theory, SMT, Parsing and Multilingual Application Development

**Registration and Contact:** school.grammaticalframework.org

**Location:** Vertex Building, UPC, Barcelona

**Organizers:** O. Caprotti, L. Màrquez, A. Ranta, J. Saludes, S. Xambó

**Travel Grants:** Via application procedure

**Sponsors:** CLT Gothenburg, MOLTO EU FP7-ICT-247914, UPC.

# Forthcoming

During Summer 2011, the project will release the Mathematical Grammar Library for simple drill problems in mathematics. The GF Grammar IDE is planned to be ready by Autumn 2011 as well as the Translation tools API and a prototype of Grammar-Ontology Interoperability. Another prototype planned for September 2011 is the first result on Patent Machine Translation with Information Retrieval case study.

In terms of events organized by MOLTO, the second "GF Summer School: Frontiers of Multilingual Technology" will be held in Barcelona in August 15-26, 2011. A tutorial on GF is planned for CADE 2011. The MOLTO partners have also agreed to organize in June 2012 the next conference on FreeRBMT in Gothenburg.

The third MOLTO project meeting will take place at the beginning of September 2011 in Helsinki.

Stay tuned by subscribing to the MOLTO RSS feed or follow us on Twitter.

## Contact Points

**Goeteborgs universitet**
Aarne Ranta,
<aarne@chelmers.se>

**Helsingin Yliopisto**
Lauri Carlson,
<lauri.carlson@helsinki.fi>

**Universitat Politècnica de Catalunya**
Jordi Saludes,
<jordi.saludes@upc.edu>

**Ontotext AD**
Borislav Popov,
<borislav.popov@ontotext.com>