MOLTO

D1.3 Progress Report

http://www.molto-project.eu

Contract No.:	FP7-ICT-247914
Project full title:	MOLTO - Multilingual Online Translation
Deliverable:	D1.3. Progress Report T12
Security (distribution level):	Confidential
Contractual date of delivery:	M13
Actual date of delivery:	18 Apr 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 year of the MOLTO project lifetime, 1 March 2010 - 28 Feb 2011.

Declaration by the scientific representative of the project coordinator

I, as scientific representative of the coordinator of this project and in line with the obligations as stated in Article II.2.3 of the Grant Agreement declare that:

- The attached periodic report represents an accurate description of the work carried out in this project for this reporting period;
- The project (tick as appropriate) [1]:

  □ has fully achieved its objectives and technical goals for the period;

  □ has achieved most of its objectives and technical goals for the period with relatively minor deviations.

  □ has failed to achieve critical objectives and/or is not at all on schedule.

- The public website, if applicable

  □ is up to date

  □ is not up to date

- To my best knowledge, the financial statements which are being submitted as part of this report are in line with the actual work carried out and are consistent with the report on the resources used for the project (section 3.4) and if applicable with the certificate on financial statement.
- All beneficiaries, in particular non-profit public bodies, secondary and higher education establishments, research organisations and SMEs, have declared to have verified their legal status. Any changes have been reported under section 3.2.3 (Project Management) in accordance with Article II.3.f of the Grant Agreement.

Name of scientific representative of the Coordinator: ....

Date: ..../ ..../ ....

For most of the projects, the signature of this declaration could be done directly via the IT reporting tool through an adapted IT mechanism.

[1] If either of these boxes below is ticked, the report should reflect these and any remedial actions taken.

Table of Contents

Declaration by the scientific representative of the project coordinator......................................2
1 Publishable Summary..4
1.1 Project context and objectives...4
1.2 Main results achieved so far..4
1.3 Expected final results and their potential impact and use...5
1.4 Public website 5
2 Core of the report...5
2.1 Project objectives for the period 6
2.2 Work progress and achievements during the period 6
2.2.1 WP2 Grammar Developer's Tools – M12 6
2.2.2 WP3 Translator's Tools – M12 6
2.2.3 WP4 Knowledge Engineering – M12 7
2.2.4 WP5 Statistical and Robust Translation – M12 8
2.2.5 WP6 Case Study: Mathematics – M12..11
2.2.6 WP7 Case Study: Patents – M12...12
2.2.7 WP8 Case Study: Cultural Heritage – M12 13
2.2.8 WP9 User Requirements and Evaluation – M12...13
2.2.9 WP10 Dissemination and Exploitation – M12..14
2.3 Project management during the period...15
2.4 Deliverables and milestones tables...16
3 Use of the resources...18
4 Financial statements...19

Index of Tables

Table 2.1. Deliverables 16
Table 2.2. Milestones..17
Table 3.1. Use of resources for UGOT for the period 18
Table 3.2. Use of resources for UHEL for the period...18
Table 3.3. Use of resources for UPC for period 19
Table 3.4. Use of resources for Ontotext for period 19

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 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 [4], described in D1.2.
2. The MOLTO KRI [5], described in D1.2, has been extended to allow natural language queries in Swedish as well as in English.

A pre-release of a web-based grammar editor for creating and compiling GF application grammars in the cloud can be tested online[2]. 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 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

[2] http://www.grammaticalframework.org/demos/gfse/

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 runtime interpreter for PGF grammars is operational for parsing and linearization and used on the android 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 these 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.

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. A LinkedIn group (http://www.linkedin.com/groups?mostPopular=&gid=3703935) has also been created for people interested in MOLTO.

2 Core of the report

2.1 Project objectives for the period

The project objectives for the second semester focus on laying foundational infrastructure for the upcoming concrete results due later in the project. The major deliverables due for this period are:

- MOLTO test criteria, methods and schedule, defining evaluation and quality assessment,
- Knowledge Representation Infrastructure, to handle ontologies and tasks related to information retrieval
- GF Grammar Compiler API, to support development of hybrid SMT-GF models and multilingual applications

These deliverables fix the tools to be used, improved and integrated during the rest of the project.

2.2 Work progress and achievements during the period

2.2.1 WP2 Grammar Developer's Tools – M12

The main achievements in WP2 in the period were

(1) the release of GF 3.2 and
(2) the web-based grammar development environment.

They are both important steps toward making GF more accessible for application programmers.

In the GF 3.2 release, the ease of installation was radically improved: on all major platforms (Linux, Mac, Windows), one just downloads one file and gives one command to install the GF environment. The library documentation in GF 3.2 is made multilingual so that all constructs are exemplified in 16 languages. This uses an automatic generation technique that will stay up to date when the library changes.

The web-based grammar development environment makes it possible to get started with GF without any installation at all. The grammars are built "in the cloud", and compile into translation systems likewise in the cloud. The editor guides the user, requiring no previous knowledge of GF syntax. New languages can be added by minimal modifications of old ones.

Related Publications

K. Angelov and A. Ranta. *Implementing Controlled Languages in GF*. N. Fuchs (ed.), CNL-2009 Controlled Natural Languages, LNCS/LNAI 5972, 2010.

O. Caprotti, K. Angelov, R. Enache, Ramona, T. Hallgren, and A. Ranta: *The MOLTO Phrasebook*. Swedish Language Technology Conference SLTC 2010.

A. Ranta, K. Angelov, and T. Hallgren. *Tools for multilingual grammar-based translation on the web*. Proceedings of the ACL 2010 System Demonstrations, ACM Digital Library, 2010.

A. Ranta, Grammatical Framework: Programming with Multilingual Grammars, CSLI Publications, Stanford, 2011, ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

2.2.2 WP3 Translator's Tools – M12

The main objectives for this period were the Tools API package (due at M18). The API has proceeded with three major steps:

- the MOLTO translation editor (by Krasimir Angelov),
- the MOLTO vocabulary editor (by Junyou Shen), and
- the TermFactory back end (by Lauri Carlson).

Simultaneous to the API, tasks related to content production have proceeded with three aspects:

- Term harvesting from Web of Data (by Inari Listenmaa)
- Ontology / vocabulary learning (by Seppo Nyrkkö)
- WordNet en-fi (supporting work done at UHEL FinnWordNet project)

We have clearly significant progress with developing and demonstrating the C language GF runtime support (by Lauri Alanko).

Work towards Deliverable D3.1 MOLTO translation tools API (M18)

The tools API package is in steady progress to be completed. Work is primarily needed in integrating the vocabulary editor with MOLTO translation editor and TermFactory. Also, TermFactory is to be integrated with the MOLTO KRI. Additionally, the integration plans include transferring WordNet en-fi entries to the KRI, as well as populating TermFactory entries from the KRI. Related content production tasks are proceeding at UHEL, supported by parallel research projects, including FinnWordNet and FIN-CLARIN.

The current progress meets the goals described in the Annex I. Use of resources in WP 3 shows no major deviations from Annex I (Description of Work).

Related publications

I. Kudashev, I. Kudasheva, L. Carlson: TermFactory: A Platform for Collaborative Ontology based Terminology Work. In Proceedings of the XIV Euralex International Congress (Leeuwarden, 6–10 July 2010).

I. Kudashev, L. Carlson, I. Kudasheva: TermFactory: Collaborative Editing of Term Ontologies. In Proceedings of Terminology and Knowledge Engineering Conference 2010 : Presenting Terminology and Knowledge Engineering Recourses Online: Models and Challenges. Dublin, 2010.

A. Norta, R. Yangarber, L. Carlson: Utility Evaluation of Tools for Collaborative Development and Maintenance of Ontologies. Enterprise Distributed Object Computing Conference Workshops (EDOCW), 2010

2.2.3 WP4 Knowledge Engineering – M12

The objectives of the work package in this period were:

- knowledge representation infrastructure to be delivered, installed and made accessible to partners
- KRI to be loaded with data sets useful for the work of the consortium
- initial GF – ontology interoperability to be developed in the direction of NL query – GF – SPARQL

During the first year of the project we have collected the requirements of the case studies and the R&D partners with respect to the knowledge representation infrastructure needs and the data sets which will be needed. We have deployed KRI and loaded it with the PROTON basic upper-level ontology and the WKB knowledge base of entities and facts of general importance, a pre-linked data set following the same principles and serving as a basis for the transition to the more domain-specific data sets needed by the case studies. KRI provides a SPARQL end point, RDF DB API and a user interface for search and navigation over the structured data sets.

Major efforts in WP4 have been dedicated to the NL – ontology interoperability through the research and implementation of GF – ontology mapping. For this purpose we have developed a corpus of natural language queries that match the ontology schema in the default data set. Afterwards, grammars covering this corpus have been developed. We came up with a language for definitions of mappings between ontology constructs and GF sentence structures.

A new approach for the two-way interoperability between the GF grammars and OWL has been created after M6, which lead to a solution independent of the underlying GF grammar and ontology, compared to the pre-M6 version, which was based on concrete syntax mapping abstract syntax trees.

New ways for optimization are being investigated, for retrieving the abstract syntax trees of querying GF from Java, than the current approach of using piped command line scripts.

The Swedish grammar is applied and successfully tested. The support of the distinct Swedish language letters (åäö) is checked and proved. The interoperability with GF requires specification of the used language and encoding of the text of the queries.

The interoperability engine between the supported GF structures and SPARQL queries contains sets of GF expressions including constants, shortcuts, macros definitions, wildcard notations and logical conditions for better expressivity, generalization and restricting or enlarging more matches in the GF grammar to groups of queries in the ontologies.

There were two simultaneous developments. The fastest and more stable approach has been selected. The developed component in the KRI has been introduced with the new grammars after intensive tests with Swedish and English examples.

The latest changes in the GF framework have been installed in relation with the newest abstract and concrete language grammars.

Related Publications

[Accepted, September 2010] Enache, R., Angelov, K.: Typeful Ontologies with Direct Multilingual Verbalization. In 2nd Workshop on Controlled Natural Language, Marettimo Island, Sicily (ITALY) September 13-15, 2010.

[Submitted, October 2010] Enache, R., Angelov, K.: Typeful Ontologies with Direct Multilingual Verbalization. In Special Issue of the Studia Logica Journal on Logic and Natural Language, 2011

2.2.4 WP5 Statistical and Robust Translation – M12

WP5 is active from Month 7 to Month 30. The original schedule for this work package was designed so that the corresponding case of study, WP7, having started three months before, made available data and some basic components already by the beginning of WP5. Due to the delay of WP7, WP5 has suffered some reordering of the tasks, although the tasks themselves remain the same.

Work towards Deliverable D5.1 (Month 18)

D5.1: 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 have compiled general-purpose large bilingual and monolingual corpora for training the basic SMT systems. We used a subset of the data sets included in the Fifth Workshop on Statistical Machine Translation

(WMT10[3]) including European Parliament Proceedings, Newspaper articles and United Nations Proceedings. The corpus consists of 2 million sentence pairs randomly selected from the various sources keeping their respective proportion. The corpus has been annotated at several linguistic levels such as lemma, part-of-speech and chunk.

Besides the out-of-domain corpus, domain specific corpora are needed to adapt the general purpose SMT system to the concrete domain of application in this project (Patents case study, WP7). For this, WP5 uses the patents parallel corpus selected in WP7.

The in-domain corpus has been divided in three sets: a training set with 279,282 aligned parallel fragments in English, French and German, a small development set (993 fragments) and a test set (1008 fragments). As a task of the work package the corpus will be annotated with syntactic information such as part-of-speech or chunk tags. The depth of the annotation will depend on the concrete language and the available linguistic processors.

Work towards Milestone MS7 (Month 24)

MS7: First prototypes of hybrid combination models.

As previously said, there have been variations in the order of the tasks to be taken during the first year of the work package. In a practical way, that means that some work on the baseline translation systems has been postponed and some work for the hybrid translation systems has been advanced. From the original first-year plan:

1. Compilation and annotation of corpora from the patents domain.
2. Training and adaptation of the base SMT systems.
3. Statistical extension of the patents GF grammar.
4. Evaluation and comparison of the baselines (GF, SMT and cascade systems) in real domain data.
5. Initial experiments with the combination approaches.

We have started to implement generic modules for the first experiments with the combination approaches (point number 5). On the other hand, the final corpus definition (1) and the GF grammar and baseline (3) will be finished later in time according to the evolution of WP7, but still within the one year schedule. In the following we show the most significant achievements.

SMT baseline

In order to quantify the quality of the hybrid SMT-GF system, this WP defines three different baselines to compare the results:

- SMT system
- GF translation system
- Naïve (cascade) combination SMT-GF as a first hybridization approach

Current work has been focused on training the in-domain phrase-based SMT system. This system is based on Moses and has been developed using standard SMT tools:

- Corpus: WP7 selected corpus
- Language model: 5-gram interpolated Kneser-Ney discounting, SRILM Toolkit
- Alignments: GIZA++ Toolkit
- Translation model: Moses package
- Weights optimization: MERT against BLEU

[3] http://www.statmt.org/wmt10/

- Decoder: Moses

Table 5.1 shows a first evaluation of this baseline using the BLEU metric and its comparison with two public SMT systems for general translation: Bing[4]and Google[5]. These systems can be considered the state-of-the-art of a SMT open domain translator. Our in-domain trained system performs significantly better than the two general purpose ones mainly because of two reasons. First, it has been trained on the specific domain and, second, the use of the tokenization tools developed in WP7 to deal with chemical compounds is a key element for an adequate translation.

	EN2DE	DE2EN	EN2FR	FR2EN	DE2FR	FR2DE
Bing	0.33	0.43	0.43	0.45	0.20	0.24
Google	0.45	0.58	0.53	0.62	0.43	0.39
Domain SMT	***0.58***	***0.65***	***0.62***	***0.70***	***0.56***	***0.53***

Table 5.1. BLEU scores for the in-domain SMT baseline. The table shows all the translation pairs given the languages: English (EN), French (FR) and German (DE).

A deeper evaluation analysis is also being carried out. Table 5.2 shows as an example of this analysis the results for several lexical metrics applied to the same three systems for the English-German language pair. In the future, the set of metrics will be extended to semantic and syntactic metrics as well. To do this, we will make use of the Asiya software as explained in WP9.

Besides the work on the evaluation of the SMT baseline, a fourth system trained with the out-of-domain corpus will be also included in order to be compared with the general open domain systems Google and Bing.

	DE2EN			**EN2DE**		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1-WER	0.52	0.64	***0.72***	0.42	0.51	***0.69***
1-PER	0.66	0.76	***0.82***	0.56	0.64	***0.77***
1-TER	0.59	0.67	***0.76***	0.45	0.53	***0.71***
BLEU	0.43	0.58	***0.65***	0.33	0.45	***0.58***
NIST	8.25	9.67	***10.12***	6.53	8.05	***9.40***
ROUGE-W	0.40	0.48	***0.52***	0.34	0.41	***0.48***
GTM-2	0.30	0.40	***0.47***	0.25	0.32	***0.43***
METEOR-pa	0.60	0.69	***0.74***	0.36	0.45	***0.57***
ULC	0.09	0.29	***0.41***	0.03	0.19	***0.43***

Table 5.2. Set of lexical metrics applied to the translation of the three SMT systems in the English-German language pair.

[4] http://www.microsofttranslator.com/

[5] http://translate.google.com

GF alignments for SMT usage

As in the case of the baselines, several hybridization approaches have been defined:

1. Hard integration: force fixed GF fragment translations within a SMT system.
2. Soft integration led by SMT: make available GF fragment translations to a SMT system.
3. Soft integration led by GF: complement with SMT options the GF translation structure.

MOLTO's work up to now has been devoted to the second approach to the hybridization. In order to make available GF translations to a SMT system one mainly needs to be able to feed an SMT decoder with translation pairs, that is, with fragments of texts aligned in both languages. In this way, if GF is able to generate Giza-like alignments, phrases can be extracted in the SMT style and combined in translation tables.

During the first six months of this WP, GF has been adapted so that now it is able to generate both alignments in Graphviz format and with a text Giza-like nomenclature. Given this new functionality we have used the Phrasebook grammar and the Resource grammar to generate synthetic corpora and Giza-like alignments for this parallel English-Spanish corpus. The synthetic corpora generated from these grammars are still too small for SMT, but this experiment served us as a proof of concept and showed that we can include GF extracted phrases into the SMT translation pipeline. The following step regarding this task is the application of the method to the use case domain: patents. This is tightly linked to the creation of a patents grammar in WP7.

Robust parsing

For both kinds of soft integration, one also needs GF to be able to parse general text robustly, it must be able to skip those structures not covered by the grammar and give some general information so that the statistical component of the engine takes care of the fragments.

The first work on this task is the robust parser being developed for GF. Current experiments use shallow parsing as a first approximation and have been tested against the basic noun phrases from the Penn Treebank. 75% of the phrases appearing in this corpus were successfully parsed, and efforts are being made to increase the coverage by a better definition of the syntax of named entities, the creation of a grammar for dates and an improvement of the lexicon. The following step in this case is the analysis of the performance for verb phrases and the subsequent application to the patent corpus.

Related Publications

SMatxinT, the Spanish-to-Basque hybrid translator. Cristina Espa~na-Bonet, Gorka Labaka, Lluis Marquez and
Kepa Serasola. UPC Internal Report.

2.2.5 WP6 Case Study: Mathematics – M12

We are working to complete deliverable D6.1 (*Simple drill grammar library*). The *Simple drill grammar library* grows from a GF library that was developed as part of the WebALT project (EDC-22253). While using the *abstract* structure of the former library, we made progress in:
Upgrading code to comply with the actual GF resource library: that is, from GF version 1.4 to GF version 3.2.1. This will make Simple drill grammar library compatible with the present GF state of the art and, therefore, allows code simplification.
Code cleaning and modularization: for better usability, we split the former grammar library into a different module for each content dictionary as shown below.

Extension to other languages: we have extended the targeted languages from 7 to 10 languages as shown below.

For each of the languages (Bulgarian, German, Catalan, Italian, English, Romanian, Finnish, Spanish, French, Swedish) we split the former monolithic library into the following modules (modeled along the lines of OpenMath *content dictionaries*): `Arith1`, `Arith2`, `Calculus1`, `Complex1`, `Fns1`, `Integer1`, `Integer2`, `Interval1`, `Limit1`, `LinAlg1`, `LinAlg2`, `Logic1`, `MinMax1`, `Nums1`, OpenMath, `PlanGeo1`, `Quant1`, `Relation1`, `Rounding1`, `SData1`, `Set1`, `SetName1`, `Transc1` and `VecCalc1`.

The following table displays some figures summarizing what has been done so far:

Concept	WebALT	MOLTO
Lines of code	20,731	12,544
Modules	158	425
Languages	7	10+1

Notice that though the number of languages has increased, we have now fewer lines of code. This is the effect of code cleaning and the usage of the up-to-date productions of the GF resource library. Since we have now a module for each content dictionary, the number of modules has increased accordingly. At this moment, the library contains 10 natural languages and LaTeX.

The Mathbar demo available online[6] is a version of the Minibar demo from WP2 with LaTeX rendering of mathematical formulas. The work done in WP6 will be presented by a poster and a demo at the 15th JAEM[7] days, 15 Jornadas para el Aprendizaje y la Enseñanza de las Matemáticas, and in addition during the GF tutorial accepted at CADE 2011. It will demonstrate the *Simple drill grammar library*.

2.2.6 WP7 Case Study: Patents – M12

The first deliverable of the work package is later in the project, but several tasks started from Month 10.

Corpus preparation

A parallel corpus in three languages (English, French and German) has been gathered from the patent corpus given for the CLEF-IP track in the CLEF 1010 Conference[8]. These data are an extract of the MAREC corpus, containing over 2.6 million patent documents pertaining to 1.3 million patents from the European Patent Office. Our parallel corpus is a subset with those patents with translated claims and abstracts into the three languages. From this first subset we selected those patents that deal with *Specific therapeutic activity of chemical compounds or medical preparations*, corresponding to the IPC code A61P. This is the domain defined in the DoW.

The final corpus built this way covers 56,000 patents out of the 1.3 million. That corresponds to 279,282 aligned parallel fragments as it can be seen in Table 5.3. Two small sets for development and test purposes have also been selected: 993 fragments for development and 1008 for test.

SET	Segments	EN tok	DE tok	FR tok

[6] http://www.grammaticalframework.org/demos/minibar/mathbar.html

[7] http://www.15jaem.org/

[8] http://clef2010.org/

Training	279,282	7,954,491	7,346,319	8,906,379
Development	993	29,253	26,796	33,825
Test	1,008	31,239	28,225	35.26

Table 5.3. Size in segments and number of tokens (tok) for the three sets defined in the patents domain.

Compound detector

The patent corpus is written in a lawyer-like style and uses a very specific vocabulary of chemistry full of compound names. The treatment of these compounds is crucial for the translation. We are therefore developing a pipeline to detect, properly tokenize and translate them.

The first component, the detector and tokenizer, is based on affix detection. A list with 150 affixes has been compiled and it is used to select the candidate tokens to be a compound from the corpus. These candidates are matched against a dictionary and those without a match are considered to be a compound. These compounds do not get an internal tokenization. The results of WP5 show that our SMT baseline with the compound tokenization improves two state-of-the-art systems such as Google and Bing and this confirms the importance of the detection for an adequate translation.

More than 100,000 compounds are found with the procedure stated above. However, this list of compounds contains some noise, that is, there are some regular words in the list. Since the amount of noise is considerable and although the extra words do not in general imply a wrong tokenization, the selection must be stricter in future versions of the detector in order to build a reliable compound dictionary.

2.2.7 WP8 Case Study: Cultural Heritage – M12

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

2.2.8 WP9 User Requirements and Evaluation – M12

The objective for this period was the deliverable D 9.1 MOLTO test criteria, methods and schedule, due at M6. An internal draft was circulated during period M6 to M8. After commentary and refinement, the final version was delivered in 12/2010 (M9).

A significant progress in WP 9 was brought in by M. Koponen's work with machine translation quality and error analysis.

Related to the automatic evaluation that will be mainly used in the Patents use case to analyse different linguistic dimensions in a fast way, the Asiya software has been extended (^Asiya is publicly available at http://www.lsi.upc.edu/~nlp/Asiya).

In short, Asiya is a common interface to a compiled collection of evaluation and meta-evaluation methods. The metric repository incorporates the latest versions of most popular metrics, operating at different linguistic dimensions (lexical, syntactic, and semantic) and based on different similarity assumptions (precision, recall, overlap, edit rate, etc.). Asiya also incorporates schemes for metric combination, i.e., for integrating the scores conferred by different metrics into a single measure of quality. The meta-metric repository includes both measures based on human acceptability (e.g., correlation with human assessments), and human likeness, such as Orange and King, as well as several statistical significance tests.

As a summary of the work, some meta-evaluation measures and the latest evaluation measures have been incorporated. Also, a family of genuine document-level evaluation measures based on discourse representations has been implemented. This makes available to MOLTO a very rich set of metrics to be used to evaluate English translations and also a considerable amount of metrics to be applied on French and German translations (i.e., the languages defined in WP7).

The use of resources shows no major deviation from the Annex I (Description of Work). However, we have realized that Deliverable D9.1 due date has been planned too early, in view of the delays in internal project communication. Thus, the delivery of D9.1 had to be delayed from the planned deadline, mostly because the case study and content-related evaluation targets were under refinement, and secondly for collecting feedback from distinct WP owners to reflect their views on the evaluation, scope and metrics. Further communication between WP9 and other sites was necessary to clarify the intended goals for case studies. As a corrective action to the delay, we are carrying out an update round on the test criteria to reflect the case studies. Also, related to the following deliverable D9.2, we are working on a preliminary evaluation checkup round across the MOLTO sites.

Work towards Deliverable D9.2. MOLTO evaluation and assessment report (M36)

We still see a need for refining the test criteria and methods described in D9.1. The criteria are to be refined to better reflect the case studies data. This is a required, additional step in order to proceed with D9.2 MOLTO evaluation and assessment report. Before the D9.2 due date, there will be a preliminary checkup with each associated site in good time before their WP due dates. The aim of this checkup is to ensure that the evaluation metrics and criteria are valid in the evolving, concurrent project environment.

Related publications

Koponen, M. *Assessing Machine Translation Quality with Error Analysis*. MikaEL: Electronic proceedings of the KäTu symposium on translation and interpreting studies; Volume 4, 2010
Jesús Giménez and Lluís Màrquez. *Linguistic Measures for Automatic Machine Translation Evaluation*. To appear in Machine Translation, Springer Netherlands, 2010.

Jesús Giménez and Lluís Màrquez. *Asiya: An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation*. The Prague Bulletin of Mathematical Linguistics, No. 94, 2010

Elisabet Comelles, Jesús Giménez, Lluís Màrquez, Irene Castellón and Victoria Arranz. *Document-level Automatic MT Evaluation based on Discourse Representations*. In Proceedings of the 5th Workshop on Statistical Machine Translation (IWMT, 2010), ACL-2010, Uppsala, Sweden, 2010.

2.2.9 WP10 Dissemination and Exploitation – M12

The stated objectives of this workpackage are to:

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

In order to promote the project we have attended a number of meetings, delivered tutorial, systems demonstrations and posters at:

- GF tutorial, LREC Malta 17-23 May 2010
- EAMT 2010, 27-28 May 2010, poster
- ACL 2010, Uppsala 11-16 July 2010, poster and system demo
- CNL 2010, 13-15 September 2010, Marettimo. Talk and GF tutorial
- SLTC 2010, Linköping 28-29 October 2010, poster and system demo
- META-NET Forum, November 2010
- FreeRBMT, 20-21 January 2011, Barcelona, MOLTO (invited talk)

The second MOLTO project meeting took place in Varna, hosted by Ontotext, in September 2010. A week-long project internal meeting has taken place in Gothenburg during the week 1-5 November 2010 to study how to make GF more robust using SMT. The lectures, mainly delivered by the UPC team, have been open to the general audience and could be considered extra courseware by regular students to obtain credit. MOLTO events lectures and slide presentations are archived on the website (http://www.molto-project.eu/biblio).

The planned events for the coming year include:

- GF Summer School: Frontiers of Multilingual Technologies 15-26 August 2011, Barcelona
- CADE 2011, 31 July- 5 August 2011, GF Tutorial including a demo of the case study on Mathematics
- FreeRBMT, to be organized by UGOT in June 2012

Besides the online demonstration of a multilingual travel phrasebook, described online in Deliverable D10.2, the Knowledge Representation Infrastructure (MOLTO KRI) is demoed as well and shows information retrieval in English and Swedish using natural language.

As part of this workpackage we also produced the public document "*MOLTO - Multilingual On-line Translation - Annual Report 2010-2011*" listed in Appendix 10 to Annex I[9].

2.3 Project management during the period

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, with numerous exchanges of emails, finally leading to the decision of EPO of not entering the MOLTO Consortium in the middle of November. EPO however agreed to provide the patent corpus and to assist in dissemination and exploitation. During the month of November 2010, we amended the Grant Agreement, the Description of Work and the Budget, by redistributing the work and the funding among the remaining partners. Finally the contract amendment was signed on 1 March 2011, effective from 23 April 2010.

MXW's tasks and effort are distributed to other partners in the following way:

- WP7 work distributed to UGOT, Ontotext, UPC
- WP7 leader is UPC, with Cristina España as the WP leader
- WP7 starts Month 10 and ends Month 33
    - deliverables postponed by 4 months
    - Milestone 9 postponed to Month 33

All other WP's, deliverables, and milestones retain their schedule from the original DoW.

UGOT has submitted a proposal to extend the Consortium, MOLTO - Enlarged EU (FP7-ICT-2011-7), in the framework of the EU programme *Small or medium-scale focused research project (STREP) ICT Call 7*, ICT 2011.11.3. The leader will still be UGOT with new participants: University of Zurich, and the Dutch company Be Informed. The former MOLTO partner UHEL is also part of the new project. The planned activities would start on 1 Aug 2011 and run for 18 Months until 28 February 2013, end of both projects. The total effort planned is 73Person Months (10 MGT, 63 RTD) to carry out two main objectives: Multilingual wiki system (wiki scenario), and Localization of interactive knowledge-based systems.

[9] http://www.molto-project.eu/sites/default/files/AnnualReport2011.pdf

2.4 Deliverables and milestones tables

Deliverables											
Del. no.	**Deliverable name**	**Version**	**WP no.**	**Lead beneficiary**	**Nature**	**Dissemination level**[10]	**Delivery date from Annex I (proj month)**	**Actual / Forecast delivery date Dd/mm/yyyy**	**Status No submitted/ Submitted**	**Contractual Yes/No**	**Comments**
D1.1	Workplan for MOLTO	1	WP1	UGOT	R	CO	M1		Submitted	Yes	Has to be constantly kept up to date by the partners online
D10.1	Dissemination plan with monitoring and assessment	1	WP10	UGOT	R	CO	M3		Submitted	Yes	Has to be constantly kept up to date by the partners online
D10.2	MOLTO web services	1	WP10	UGOT	P	PU	M3		Submitted	Yes	
D9.1	MOLTO test criteria, methods and schedule	1	WP9	UHEL	R	CO	M7		Submitted	Yes	Has to be revised
D1.2	Periodic management report 1	1	WP1	UGOT	R	CO	M7		Submitted	Yes	
D4.1	Knowledge Representation Infrastructure	1	WP4	Ontotext	RP	PU	M8		Submitted	Yes	Has to be amended to include a glossary
D2.1	GF Grammar Compiler API	1	WP2	UGOT	P	PU	M12		Submitted	Yes	

Table 2.1. Deliverables

[10] **PU** = Public
PP = Restricted to other programme participants (including the Commission Services).
RE = Restricted to a group specified by the consortium (including the Commission Services).
CO = Confidential, only for members of the consortium (including the Commission Services).
Make sure that you are using the correct following label when your project has classified deliverables.
EU restricted = Classified with the mention of the classification level restricted "EU Restricted"
EU confidential = Classified with the mention of the classification level confidential " EU Confidential "
EU secret = Classified with the mention of the classification level secret "EU Secret "

Milestones							
Milestone no.	**Milestone name**	**Work package no**	**Lead beneficiary**	**Delivery date from Annex I dd/mm/yyyy**	**Achieved Yes/No**	**Actual / Forecast achievement date dd/mm/yyyy**	**Comments**
MS1	15 Languages in the Library	WP2, WP10	UGOT	M6	Yes		
MS2	Knowledge representation infrastructure	WP4	Ontotext	M6	Yes		
MS3	Web-based translation tool available	WP3, WP10	UHEL	M12	Yes		

Table 2.2. Milestones

3 Use of the resources

<table>
<tr><th colspan="4">Personnel, subcontracting and other major cost items for UGOT for the period</th></tr>
<tr><th>Work Package</th><th>Item description</th><th>Amount in € with 2 decimals</th><th>Explanations</th></tr>
<tr><td>WP1, WP2, WP4, WP5, WP7, WP10</td><td>Personnel direct costs</td><td>126378,74</td><td>Aarne Ranta (WP1:2PM, WP2:1PM), Olga Caprotti (WP1:2PM, WP7:1PM, WP10:3PM), Ramona Enache (WP2:8PM, WP4:1PM, WP5:1PM)</td></tr>
<tr><td>WP1, WP10, WP3</td><td>Travel costs</td><td>17272,30</td><td>MOLTO project meetings, CICling2010, LREC 2010, ACL2010, Google London Workshop, CNL 2010, SLTC 2010, META-NET Forum 2010,</td></tr>
<tr><td></td><td>Remaining direct costs</td><td>2049,26</td><td>Poster printing, 2 laptop computers, editing software</td></tr>
<tr><td></td><td>Indirect costs</td><td>87420,18</td><td></td></tr>
<tr><td colspan="2"><b>TOTAL COSTS</b></td><td>233120,48</td><td></td></tr>
</table>

Table 3.1. Use of resources for UGOT for the period

<table>
<tr><th colspan="4">Personnel, subcontracting and other major cost items for UHEL for the period</th></tr>
<tr><th>Work Package</th><th>Item description</th><th>Amount in € with 2 decimals</th><th>Explanations</th></tr>
<tr><td></td><td>Personnel direct costs</td><td>110568</td><td>Lauri Alanko (4.6PM), Lauri Carlson (1.4PM), Hyvärinen Mirka (4), Laxström Niklas (1.7PM), Listenmaa Inari (5.8PM), Nyrkkö Seppo (6.6PM), Shen Junyou (4.3PM)</td></tr>
<tr><td>WP9, WP5, WP3</td><td>Travel costs</td><td>7207</td><td>MOLTO kick-off, Barcelona, 8.-11.3.2010 /Carlson, Nyrkkö ESSLLI-kesäkoulu, Kööpenhamina, 8.-13.8.2010/Koponen MOLTO projektikokous, Varna, 7.-10.9.2010 / Carlson, Alanko, Listenmaa MOLTO hankkeen tapaaminen GF meets SMT/31.10.-2.11.2010/Koponen Aarne Ranta accommodation expenses of 4-05/05/2010</td></tr>
<tr><td></td><td>Remaining direct costs</td><td></td><td></td></tr>
<tr><td></td><td>Indirect costs</td><td>70664</td><td></td></tr>
<tr><td colspan="2"><b>TOTAL COSTS</b></td><td>188439</td><td></td></tr>
</table>

Table 3.2. Use of resources for UHEL for the period

Personnel, subcontracting and other major cost items for UPC for the period			
Work Package	**Item description**	**Amount in € with 2 decimals**	**Explanations**
WP1, WP2, WP3, WP5, WP6, WP7, WP9, WP10	Personnel direct costs	110.675,37	SALUDES CLOSA JORDI (WP1:0.5PM, WP2:0.57PM, WP6:3.81PM), XAMBO DESCAMPS SEBASTIAN (WP6:2.27PM), CARRERAS PEREZ XAVIER (WP2:0.57PM,WP5:1.37PM), ESPAÑA BONET CRISTINA (WP2:085PM, WP3:1PM, WP5:3.05PM, WP7:0.06PM, WP9:2PM), FARWELL DAVID LORING (WP5:0.46PM), MARQUEZ VILLODRE LUIS (WP1:0.50PM, WP5:0.94PM, WP7:0.61PM, WP9:0.22PM, WP10:1PM), PADRO CIRERA LLUIS (WP5:1.84PM, WP7:0.60PM, WP9:0.23PM), RODRIGUEZ HONTORIA HORACIO (WP5:1.84PM, WP7:0.60PM, WP9:0.22PM)
WP1, WP2, WP3, WP5, WP6, WP7, WP9, WP10	Other direct costs	6.911,34	**2nd Progress Meeting** (Goteborg) 10th March 2011 => Jordi Saludes; Cristina España **GF meets SMT** (Goteborg) 1st-5th November 2011 => Xavier Carreras, Cristina España, Lluís Marquez, Sebastià Xambó, Jordi Saludes **1st Molto Project** Meeting (Varna) 8th-10th Sept2011 =>Cristina España, Jordi Saludes, Sebastià Xambó **freeRBMT 2011-** 20-21 January 2011 (Barcelona)- Cristina España **EAMT 2010-**27-28th May 2010 (Saint Raphael) – Cristina España
WP1, WP10	subcontracting	1.021,74	minor tasks subcontracting: coffee breaks, dissemination poster
WP1, WP2, WP3, WP5, WP6, WP7, WP9, WP10	Indirect costs	77.963,03	
	TOTAL COSTS	196.571,48	

Table 3.3. Use of resources for UPC for period

Personnel, subcontracting and other major cost items for Ontotext for the period			
Work Package	**Item description**	**Amount in € with 2 decimals**	**Explanations**
WP 1,2,3,4,6,7,8,9,10	Personnel direct costs	67376	costs for 1PM, 3 senior developers and 4 developers
	Subcontracting		
	Equipment costs (annual depreciation)	3518	annual depreciation for notebooks, 1 PC and 1 server for testing development
	Travel costs	1926	1 kick-off and 1 project meetings, and 1 two-days conference
	Remaining direct costs	508	project meeting in Varna
	Indirect costs	117326	
	TOTAL COSTS	90002	

Table 3.4. Use of resources for Ontotext for period

4 Financial statements

Signed copies of the financial statements in original are sent separately.