SEVENTH FRAMEWORK PROGRAMME
Information and Communication Technologies

Grant agreement for**: Small or medium-scale focused research project**

Annex I - "Description of Work"

Project acronym: *MOLTO*
Project full title: Multilingual On-Line Translation
Grant agreement no.: 247914

Version number: 3 Revision 1 (21/01/2011)
Date of preparation of Annex I: 27/10/2009
Date of approval of Annex I by Commission: 27/10/2009

Beneficiary Number	Beneficiary name	Beneficiary short name	Country
1 (coordinator)	Goeteborgs universitet	UGOT	Sweden
2	Helsingin yliopisto	UHEL	Finland
3	Universitat Politècnica de Catalunya	UPC	Spain
4	Ontotext AD	Ontotext	Bulgaria
5	Matrixware GmbH[1]	MXW	Austria

[1] MXW left the Consortium in April 2010.

Table of Contents

1 Overall budget breakdown for the project

Participant	**RTD**	**MGT**	**Total Costs**	**Estimated EC funding**
UGOT	713,700	288,732	1,002,432	824,007
UHEL	598,198	51,673	649,871	500,321
UPC	732,915	52,836	785,851	602,522
Ontotext	515,923	54,843	570,766	441,785
Mxw	0,	6,365	6,365	6,365
Total	**2,560,736**	**454,449**	**3,015,185**	**2,375,000**

2 Project summary

MOLTO's goal is to develop a set of tools for translating 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.

As its main technique, MOLTO uses domain-specific semantic grammars and ontology-based interlinguas. These components are implemented in GF (Grammatical Framework), which is 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.

A part of the scale-up is to increase the size of domains and the number of languages. A more substantial part is to make the technology accessible for domain experts without GF expertise and minimize the effort needed for building a translator. Ideally, this can be done by 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. The statistical methods will add robustness to the system when desired. New methods will be developed for combining GF grammars with statistical translation, to the benefit of both.

MOLTO technology will be released as open-source libraries which can be plugged in to standard translation tools and web pages and thereby fit into standard workflows. It will be demonstrated in web-based demos and applied in three case studies: mathematical exercises in 15 languages, patent data in at least 3 languages, and museum object descriptions in 15 languages.

3 Concept and objectives, progress beyond state of the art, S/T methodology and work plan

3.1 Concept and project objectives

The MOLTO project is rooted in two lines of research. One is the GF approach to multilingual grammars and interlingua-based translation pioneered by the UGOT site since the early 1990's. The other line is semantic web technology, providing structured data that can be used as the basis of GF translation. The time is ripe to put these lines together and develop a solution to the increasingly urgent problem of real-time multilingual translation of web documents with high quality. This requires a consortium with a variety of competences. While UGOT stands for the multilingual GF technology, web technology is represented by Ontotext. UPC is the main responsible for scaling up GF translation with statistical methods. UHEL contributes with the integration of MOLTO techniques with standard translation tools and workflows. To show the generality of the techniques, three very different case studies are performed: mathematical exercises (main responsible UPC), patents (Mxw), and cultural heritage (UGOT).

MOLTO builds on the results of several earlier projects, in particular the following European projects:

1. TYPES, a series of networks of excellence, developing semantic representations and interactive systems based on type theory and also GF (UGOT)
2. TALK, Tools for Ambient Linguistic Knowledge, developing GF and the resource grammar library (UGOT)
3. WebALT, Web Advanced Learning Technologies, developing GF and multilingual translation in the mathematics domain (UHEL, UPC)
4. JEM, Joining Educational Mathematics, dissemination and further development of GF and multilingual translation in the mathematics domain (UHEL, UPC, UGOT)
5. TAO, Transitioning Applications to Ontologies, developing tools for transitioning legacy web applications to the semantic web (Ontotext)
6. TC-STAR, Technology and corpora for speech to speech translation, integrating human knowledge in data-driven translation systems (UPC)

The following table shows the main achievements of the named project from the MOLTO point of view and how MOLTO builds on them.

Project	Result	Advancement
TYPES	semantics and interaction	natural language interface
TALK	domain grammars	scaling up domain grammars
WebALT	multilingual mathematics	enhanced grammar and tools
JEM	dissemination of WebALT	extended domains and user base
TAO	adaptation of ontologies	adaptation of ontology-based grammars
TC-STAR	hybrid systems	new kinds of hybrid systems

The mission of the MOLTO project is thus to enable multilingual translation with high quality, and with a level of speed and automation sufficient for real-time translation tasks. An extreme use case for the task is a **multilingual wiki page**, such as seen in Wikipedia[2]. This use case is characterized by the following desired features:

1. **many languages** (currently 264 languages in Wikipedia)
2. **many contributors** (hundreds of thousands in Wikipedia)
3. **frequent updates** (average in Wikipedia close to 20 per article)
4. **synchrony between languages** (the same information in different languages; updates in one language propagated to the others)
5. **high quality** (grammatically and stylistically flawless text)

The goal of synchrony is where the need of translation comes in. Wikipedia is based on the voluntary work of human translators. but the frequency of updates and the multitude of languages make it impossible to achieve full synchrony by human translation. Consequently, a vast majority of the articles can only be found in one language: there are 2.8 million articles in English, but only 0.9 million in the second-largest Wikipedia language, German. Only 25 languages have more than 0.1 million articles. Automatic translation is the only conceivable way to maintain any kind of synchrony through languages and updates.

The above use case is of course highly relevant to the European reality, a union of countries with 23 official languages, where information from all aspects of life needs to be freely exchanged for mutual benefit.

The best state-of-the-art translation tools, Google translation[3] and Systran[4] are far from being capable of tasks like the translation of Wikipedia. One problem is the number of languages covered by them, which is way below 264 (currently 41 in Google and 15 in Systran). The essential problem, however, is quality. Even though Google and Systran translations are usually good enough to give an idea of the contents of a text, they are often grammatically and semantically flawed. Thus they cannot be used in tasks where reliability is required. While machine translation is occasionally performed on Wikipedia articles for purposes of information search[5], it is never used for the purpose of creating Wikipedia content, except perhaps as an aid for human translators.

The MOLTO project aims to provide technology which can simultaneously achieve the five goals stated above. We do not promise to scale up to the dimensions of the entire Wikipedia, but we aim to produce, as one demonstration of MOLTO technology, a set of articles in the domain of cultural heritage. The number of languages we aim to cover simultaneously is 15, which will include 12 of the 23 official languages of the European Union. The 12 EU-languages are Bulgarian, Danish, Dutch, English, Finnish, French, German, Italian, Polish, Romanian, Spanish, and Swedish, and the 3 non-EU languages are Catalan, Norwegian, and Russian.

The main respect in which the MOLTO technology does not reach all the way up to the Wikipedia task is its use of **restricted language**. This is the way in which we can achieve the goals stated. The reason is that it is impossible to combine large coverage with high precision in automatic translation. This dilemma was first noted by Bar-Hillel (1964). The main-stream systems like Google translation and Systran opt for coverage, but the choice of precision via restriction of language is not new to MOLTO; the most successful and influential example is perhaps the METEO system, which translates weather reports between English and French with high quality (Chandioux 1977). What MOLTO adds to the state of the art is to make restricted language translation much more practical and scalable than

[2] wikipedia.org

[3] www.google.com/translate

[4] www.systransoft.com

[5] semanticcompositions.typepad.com/index/2006/02/a_translationse.html

ever before.

The main limitation of restricted language translation is obviously that it cannot cope with all text. It is therefore not well adapted for translating already existing documents, but should target tasks in which the translatable content is created in the first place. Even in such tasks, the current state of the art poses two severe problems:

- The **development cost problem**: a large amount of work is needed for building translators for new domains and new languages.
- The **authoring problem**: since the method does not work for all input, the author of the source text of translation—for instance, a person writing or updating Wiki articles—may need special training to write in a way that can be translated at all.

These two problems have probably been the main obstacles to making high-quality restricted language translation more wide-spread in tasks where it would otherwise be applicable. The main tenets of MOLTO concern solving these problems:

- Development: we can decrease the effort of developing restricted language translators radically.
- Authoring: we can make it possible to translate restricted language without preparatory training and without changing the work flow of content production.

MOLTO addresses these problems by creating tools that help developers of translation systems on the one hand, and authors and translators—i.e. the users of the systems—on the other. We believe that we can improve both the development and use of restricted language translation by an order of magnitude, as compared with the state of the art. As for development costs, this means that a system for many languages and with adequate quality can be built in a matter of months rather than years. As for authoring, this means that content production does not require the use of manuals or involve trial and error, both of which can easily make the work ten times slower than normal writing.

Besides creating translation tools, MOLTO will also explore the two-way interoperability of grammars with Semantic Web[6] conceptual models (**ontologies**) and knowledge bases. In the last years, a rapidly increasing amount of various data sets has been made available in a machine readable form, through W3C[7] standards like the Resource Description Framework (RDF[8], the Web Ontology Language (OWL[9]) and initiatives like Linked Open Data (LOD [10]). LOD alone points to almost one hundred data sets, semantically aligned between each other, capturing various areas of life, from Wikipedia structured exports, through to FOAF profiles, thesauri like WordNet, movie and music databases, and all the major scientific bio-medical data sets. A part of these riches will be used in MOLTO through a highly scalable semantic data representation infrastructure, to provide MT tools with data sets containing named entity profiles and lexical knowledge.

The grammar-based MT will thereby benefit from semi-automatic creation of abstract grammars from ontologies, and potentially use the knowledge base for disambiguation on the lexical level. In the opposite direction of interoperability, from grammar to ontology, the knowledge sets will be enriched with the conceptual models captured in the grammars and the capability to render natural language as machine readable knowledge on the level of concepts, entity instances and relationships, for the purposes of both knowledge acquisition and retrieval. This interoperability will heavily effect the internal and presentation layers of the use case prototypes, providing the general user with the possibility to type in natural language to query the knowledge base, and get back grammatically sound textual representations of the resulting structured knowledge. The query functionality will be available in all languages covered by the corresponding document translation system.

Extensive case studies will be carried out to test and evaluate the tools on sufficiently different areas to show that the technology is generally applicable: mathematical teaching material, descriptions of museum objects, and patents. On these areas, we will show that

- translators can be created with reasonable effort,
- the translation tools are easy to use and fit within normal workflows,
- translation quality is significantly improved in comparison to earlier tools,
- translations quality can reach perfection in conveying the information contained in the source, in a grammatically

[6] http://en.wikipedia.org/wiki/Semantic_Web/

[7] http://www.w3.org/

[8] http://en.wikipedia.org/wiki/Resource_Description_Framework

[9] http://en.wikipedia.org/wiki/Web_Ontology_Language/

[10] http://linkeddata.org/

flawless target language,

- domain specific background structured knowledge allows rapid translator creation, improves translation quality, and provides cross-language retrieval,
- NL (natural language) querying and results dramatically improve the usability of the systems.

The translators for mathematics and museum objects will build upon existing formalized knowledge representations. They will use ontologies as a natural starting point for meaning-preserving restricted language translation, and use ontology-based technology for semantic information retrieval and natural language querying (in any target language) on the translated documents and domain knowledge bases.

The patent translation task is an opening to non-restricted language. There is a database of legacy documents, and no ready-made ontology is available with sufficient coverage of the domain. This is where **robustness** has to be introduced in the MOLTO tools. This problem will be studied by extending MOLTO's **rule-based** translation methods with **statistical** translation. Focusing on patents from the bio-medical and pharmaceutical industries, the machine translation (MT) and information retrieval in this use case will benefit from

existing structured knowledge bases like Linked Life Data[11] (LLD), aligning EntrezGene, Gene Ontology, Medical Subject Headings and almost 20 others from the domain covering symptoms, side effects, pathway interactions and drugs; patent classification taxonomies like IPC[12]; generic patent ontology PROTON Patents (currently under development by Ontotext and Matrixware); and DBPedia for open domain entity descriptions.

Statistical methods have a dominating role in today's machine translation research. Their advantages include robustness (any input can be translated) and productivity (manual rule writing is avoided). While MOLTO has a rule-based approach to both these issues, we are also interested in combining rule-based and statistical methods in optimal ways. We try to find new methods to improve robustness without sacrificing quality. Using these methods, we aim to provide a continuous scale of choices on how much manual intervention is involved to improve the quality.

3.2 Progress beyond the state of the art

The single most important S&T innovation of MOLTO will be a mature system for multilingual on-line translation, scalable to new languages and new application domains.

The following table gives an overview of how MOLTO advances the state of the art. The baseline is the current capability of systems that permit automatic publishing quality translation, such as the WebALT mathematics translation and other comparable systems based on GF or other techniques. We are not comparing the progress with Google Translate and Systran here, because these systems don't achieve the desired translation quality.

Feature	**Current**	**Projected**
Languages	up to 7	up to 15
Domain size	100's of words	1000's of words
Robustness	none	open-text capability
Development per domain	months	days
Development per language	days	hours
Learning (grammarians)	weeks	days
Learning (authors)	days	hours

The single most important tangible product of MOLTO is a software toolkit, available via the MOLTO website. The toolkit is a family of open-source software products:

1. 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
2. a translator's tool, available as an API and some interfaces in web browsers and translation tools
3. a grammar library for linguistic resources
4. a grammar library for the domains of mathematics, patents, and cultural heritage

All of these tools are portable to different platforms (operating systems, web browsers, small devices). All except the

[11] www.linkedlifedata.com

[12] http://www.wipo.int/classifications/ipc/en/

last are generic and portable to new domains and languages, as shown by the following table.

The number 18 of grammar library languages is the minimum number of languages we expect to be available at the end of MOLTO. The number 3 to 15 is the number of languages actually implemented in MOLTO's domain grammars (3 in WP7, 15 in WP6 and WP8).

Component portability	**Devices**	**Domains**	**Languages**
Grammar development	PC-size, web-based	any	any
Translation	PC-size, small device, web-based	any	any
Grammar library	PC-size, web-based	any	18
Domain grammars	PC-size, small device, web-based	specific	3 to 15

The main impact is expected to be on how the possibilities of translation are viewed in general. The field is currently dominated by open-domain browsing-quality tools (Google translate and Systran), and domain-specific high-quality translation is considered expensive and cumbersome. MOLTO will change this view by making it radically easier to provide high-quality translation on its scope of application—that is, where the content has enough semantic structure—and it will also widen this scope to new domains. Socioeconomically, this will make web content more available in different languages, including interactive web pages. At the end of MOLTO, the technology will be illustrated in case studies that involve up to 15 languages with a vocabulary of up to 2,000 special terms (in addition to basic vocabulary provided by the resource grammar).

The generic tools developed MOLTO will moreover make it possible for third parties to create such translation systems with very little effort. Creating a translation *system* for a new language covering an unlimited set of documents in a domain will be as smooth (in terms of skill and effort) as creating an individual translation of *one* document.

Here are the principal measurable expected outcomes:

1. languages treated simultaneously: up to 15
2. domains with substantial applications: 4
3. translation quality: "complete" or "useful" on the TAUS scale (Translation Automation Users Society[13])
4. source authoring: the MOLTO tool for writing translatable controlled text can be learned in less than one hour, the speed of writing translatable controlled text is in the same order of magnitude as writing unlimited plain text
5. localization of systems: the MOLTO tool for adding a language to a system can be learned in less than one day, and the speed of its use is in the same order of magnitude as translating an example text where all the domain concepts occur

The measurements of all these features are performed within WP9 in connection to the project milestones. The advisory group will confirm the adequacy and accuracy of the measurements.

Here are the links between the main objectives and the tasks in WP's:

1. adaptability of translation systems: WP2
2. user friendliness and integration in workflows: WP3
3. integration with semantic web technology: WP4
4. usefulness on different domains: WP6,7,8
5. scaling up towards more open text: WP5,7
6. quality of translation: WP9
7. wide user adaptation and exploitability: WP10

3.2.1 Multilingual grammars

The main technology behind MOLTO is **GF**, **Grammatical Framework**[14] (Ranta 2004). GF is a **grammar formalism**, akin to HPSG (Head-Driven Phrase Structure Grammar, Pollard and Sag 1994), LFG (Lexical Functional Grammar, Bresnan 1982) or TAG (Tree Adjoining Grammar, Joshi 1985)—that is, a mathematical model

[13] http://www.translationautomation.com/best-practices/quality-evaluation-and-ta.html

[14] digitalgrammars.com/gf

of natural language, equipped with a formal notation for writing grammars and a computer program implementing parsing and generation which are declaratively defined by grammars. The novel, and so far unique, feature of GF is the notion of **multilingual grammars**, which describe several languages simultaneously by using a common representation called **abstract syntax**; see Figure 1.

The core of a MOLTO translation system is a multilingual GF grammar, where meaning-preserving translation is automatically provided as a composition of parsing and generation via the abstract syntax, which works as an **interlingua**. This model of translation is different from approaches based on other comparable grammar formalisms, such as synchronous TAGs (Shieber and Schabes 1990), Pargram (Butt & al. 2002, based on LFG), LINGO Matrix (Bender and Flickinger 2005, based on HPSG), and CLE (Core Language Engine, Alshawi 1992). These approaches use **transfer rules** between individual languages, separate for each pair of languages.

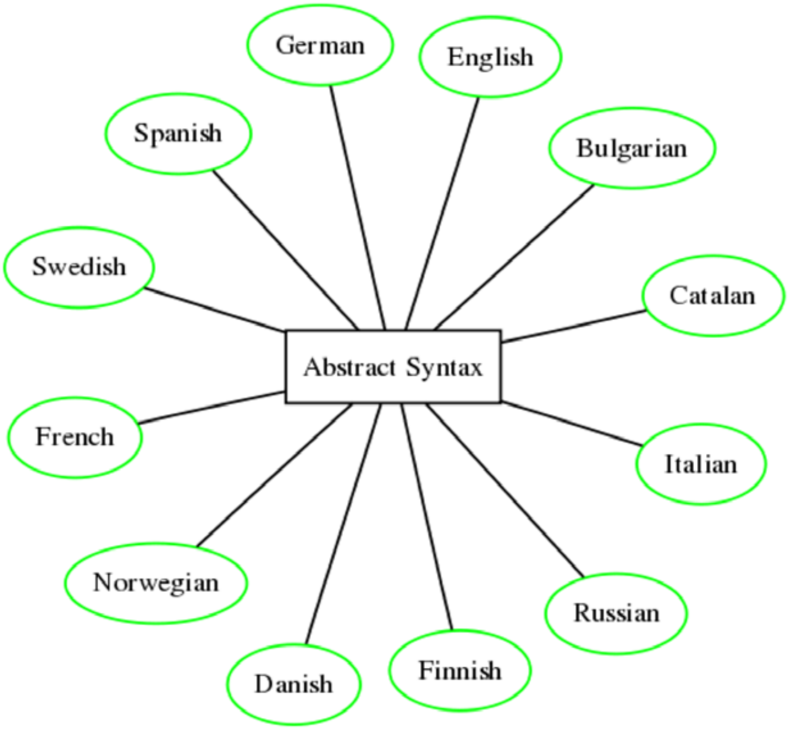


Figure 1: A multilingual GF grammar with reversible mappings from a common abstract syntax to the 12 languages currently available in the GF Resource Grammar Library.

Being interlingua-based, GF translation scales up linearly to new languages without the quadratic blow-up of transfer-based systems. In transfer-based systems, as many as $n(\tilde{n}1)$ components (transfer functions) are needed to cover all language pairs in both directions. In an interlingua-based system, $2n+1$ components are enough: the interlingua itself, plus translations in both directions between each language and the interlingua. However, in GF, $n+1$ components are sufficient, because the mappings from the abstract syntax to each language (the **concrete syntaxes**) are **reversible**, i.e. usable for both generation and parsing.

The idea of multilingual GF grammars arose as an implementation of Curry's distinction between **tectogrammatical** and **phenogrammatical** structure (Curry 1963). In GF, the tectogrammatical structure is called abstract syntax, following standard computer science terminology. It is defined by using a **logical framework** (Harper & al. 1993), whose mathematical basis is in the **type theory** of Martin-Löf (1984). Two things can be noted about this architecture, both showing improvements over state-of-the-art grammar-based translation methods.

First, the translation interlingua (the abstract syntax) is a powerful logical formalism, able to express the finest semantic structures such as context-dependencies and anaphora (Ranta 1994). In particular, it is more expressive than the simple type theory used in Montague grammar (Montague 1974) and employed in the Rosetta translation project (Rosetta 1998), which as a logic-based system has many similarities with MOLTO.

Second, GF uses a **framework for interlinguas**, rather than one universal interlingua. This makes the interlingual approach more light-weight and feasible than in systems assuming one universal interlingua, such as Rosetta and UNL, Universal Networking Language[15]. It also gives more precision to special-purpose translation: the interlingua of a GF translation system (i.e. the abstract syntax of a multilingual grammar) can encode precisely those structures and distinctions that are relevant for the task at hand. Thus an interlingua for mathematical exercises (Caprotti 2006) is different from one for commands for operating an MP3 player (Perera and Ranta 2007). The expressive power of the logical framework is sufficient for both kinds of tasks.

3.2.2 Grammar-ontology interoperability for translation and retrieval

Parallel to the first development efforts of GF in the late 1990's, another framework idea was emerging in web technology: XML, Extensible Mark-up Language, which unlike HTML is not a single mark-up language but a framework for creating custom mark-up languages. The analogy between GF and XML was seen from the beginning, and GF was designed as a formalism for multilingual rendering of semantic content (Dymetman and al. 2000). XML originated as a format for structuring documents and structured data serialization, but a couple of its descendants, RDF(S) and OWL, developed its potential to formally express the **semantics** of data and content, serving as the fundaments of the emerging Semantic Web.

Both RDF(S) and OWL have been initially designed to express formal meaning representations of data and content in machine readable form. This approach needs, as it complements, GF-like techniques for rendering information, especially in regard to rendering natural language to machine readable semantic models (ontologies) and vice versa—grammatically sound textual representations of the formal knowledge.

Almost any meaning representation format is easy to convert into GF's abstract syntax, which can then be mapped to different target languages. In particular the OWL language could be seen as a syntactic sugar for a subset of Martin-Löf's type theory so it is trivial to embed it in GF's abstract syntax. The opposite is not always feasible, but it is possible if the abstract syntax follows some restrictions. These restrictions also have the positive aspect that reasoning with the OWL subset is more efficient.

The translation problem defined in this way is radically different from the problem of translating plain text from one language to another. Many of the projects in which GF has been used involve precisely this: a meaning representation formalized as GF abstract syntax. Some projects build on previously existing meaning representation and address mathematical proofs (Hallgren and Ranta 2000), software specifications (Burke and Johannisson 2005, Beckert & al. 2007), and mathematical exercises (Caprotti 2006, in the European project WebALT[16]). Other projects start with semantic modelling work to build meaning representations from scratch, most notably ones for dialogue systems (Ranta and Cooper 2004, Bringert & al. 2005, Perera and Ranta 2007) in the European project TALK[17]. Yet another project, and one precisely corresponding to the introductory scenario of this proposal, is the multilingual Wiki system presented in (Meza Moreno and Bringert 2008). In this system, users can add and modify reviews of restaurants in three languages (English, Spanish, and Swedish). Any change made in any of the languages gets automatically translated to the other languages.

At the time of the TALK project, an emerging topic was the derivation of dialogue system grammars from OWL ontologies. A prototype tool for extracting GF abstract syntax modules from OWL ontologies was thereby built by Peter Ljunglöf at UGOT. This tool was implemented as a plug-in to the Protégé system for building OWL ontologies[18] and intended to help programmers with OWL background to build GF grammars. Even though this tool remained as a prototype within the TALK project, it can be seen as a proof of concept for the more mature tools to be built in the MOLTO project.

In slightly simplified terms, the OWL-to-GF mapping translates OWL's classes to GF's categories and OWL's properties to GF's functions that return propositions. As a running example in this and the next section, we will use the class of integers and the two-place property of being divisible ("*x* is divisible by *y*"). The correspondences are as follows:

```
Class(pp:integer ...)  <==>       cat integer ;

ObjectProperty(pp:div  <==>       fun div :
```

[15] www.undl.org

[16] EDC-22253, 2005–2007, webalt.math.helsinki.fi

[17] IST-507802, 2004–2006, www.talk-project.org

[18] protege.stanford.edu

domain(pp:integer)
range(pp:integer))
	integer -> integer -> prop ;

The GF-Protégé plug-in brings us to the development cost problem of translation systems. We have noticed that in the GF setting, building a multilingual translation system is equivalent to building a multilingual GF grammar, which in turn consists of two kinds of components:

- a language-independent abstract syntax, giving the semantic model via which translation is performed;
- for each language, a concrete syntax mapping abstract syntax trees to strings in that language.

In MOLTO, GF abstract syntax can also be derived from sources other than OWL (e.g. from OpenMath[19] in the mathematical case study) or even written from scratch and then possibly translated into OWL ontologies, if the inference capabilities of OWL reasoning engines are desired. The CRM ontology (Conceptual Reference Model[20]) used in the museum case study is already available in OWL[21].

MOLTO's ontology-grammar interoperability engine will thus help in the construction of the abstract syntax by automatically or semi-automatically deriving it from an existing ontology. The mechanical translation between GF trees and OWL representations then forms the basis of using GF for translation in the Semantic Web context, where huge data sets become available in RDF and OWL in initiatives like Open Linked Data (LOD).

The interoperability between GF and ontologies will also provide humans with natural ways of interaction with knowledge based systems in multiple languages, expressing their need for information in NL and receiving the matching knowledge expressed in NL as well:

Human -> NL -> GF -> ontology -> GF -> NL -> Human

providing an entirely new dimension to the usability of semantics-based retrieval systems, and opening extensive structured bodies of knowledge in human understandable ways.

Previous work includes systems like QuestIO, AquaLog, CLONE, CLIE (Damljanovic and Bontcheva 2008) for controlled language—ontology interaction, which are limited mostly to one language and can benefit from deeper language analysis. In contrast MOLTO will expose language-ontology interoperability in all the target languages and additionally experiment with improving cross-language retrieval robustness through hybrid grammar-statistical methods resulting in the evaluation of several alternative paths in the knowledge graph, instead of failing to match results.

The semantic infrastructure in MOLTO will also act as a central multi-paradigm index for (i) conceptual models—upper-level and domain ontologies; (ii) knowledge bases; (iii) content and metadata as needed by the use cases (mathematical problems, patents, museum artefact descriptions); and provide NL-based and semantic (structured) retrieval on top of all modalities of the data modelled. In addition to the traditional triple model for describing individual facts,

<subject, predicate, object>

the semantic infrastructure, will build on quintuple-based facts,

<subject, predicate, object, named graph, triple set>

The infrastructure will include: inference engine (TRREE[22]), semantic database (OWLIM[23]), semantic data integration framework (ORDI[24]) and a Multi-paradigm semantic retrieval engine, all of which are previous work, resulting from private (Ontotext) and public funding (TAO[25]. TripCom[26]). This approach will enable MOLTO's baseline and use case driven knowledge modelling with the necessary expressivity of metadata-about-metadata

[19] www.openmath.org

[20] http://cidoc.mediahost.org/standard_crm(en)(E73)print.xml

[21] http://www8.informatik.uni-erlangen.de/IMMD8/Services/cidoc-crm/versions.html

[22] http://www.ontotext.com/trree/

[23] http://www.ontotext.com/owlim/

[24] http://www.ontotext.com/ordi/

[25] IST-2004-026460 http://www.tao-project.eu/

[26] IST-4-027324-STP http://www.tripcom.org/

descriptions for provenance of the diverse sources of structured knowledge (upper-level, domain specific and derived (from grammars) ontologies; thesauri; domain knowledge bases; content and its metadata).

3.2.3 Grammar engineering for new languages

While abstract syntax construction is an extra task compared to many other kinds of translation methods, it is technically relatively simple, with cost moreover amortized as the system is extended to new languages. Concrete syntax construction can be much more demanding in terms of programming skills and linguistic knowledge, due to the complexity of natural languages. This task is where GF claims perhaps the highest advantage over other approaches to special-purpose grammars. The two main assets are:

- Programming language support: GF is a modern functional programming language, with a powerful type system and module system supporting modular and collaborative programming and reuse of code.
- **RGL**, the **GF Resource Grammar Library**, implementing the basic linguistic details of languages: **inflectional morphology** and **syntactic combination functions**.

The RGL covers twelve languages at the moment, shown in Figure 1; see also Khegai 2006, Ranta 2007, El Dada and Ranta 2007, and Angelov 2008. To give an example of what the library provides, let us first consider the inflectional morphology. It is presented as a set of lexicon-building functions such as, in English,

```
mkV : Str -> V
```

i.e. function mkV, which takes a string (Str) as its argument and returns a verb (V) as its value. The verb is, internally, an inflection table containing all forms of a verb. The function mkV derives all these forms from its argument string, which is the infinitive form. It predicts all regular variations: (mkV "walk") yields the purely agglutinative forms *walk-walks-walked-walked-walking* whereas (mkV "cry") gives *cry-cries-cried-cried-crying*, and so on. For irregular English verbs, RGL gives a three-argument function taking forms such as *sing,sang,sung*, but it also has a fairly complete lexicon of irregular verbs, so that the normal application programmer who builds a lexicon only needs the regular mkV function.

Extending a lexicon with domain-specific vocabulary is typically the main part of the work of a concrete syntax author. Considerable work has been put into RGL's inflection functions to make them as "intelligent" as possible and thereby ease the work of the users of the library, who don't know the linguistic details of morphology. For instance, even Finnish, whose verbs have hundreds of forms and are conjugated in accordance with around 50 conjugations, has a one-argument function mkV that yields the correct inflection table for 90% of Finnish verbs (Ranta 2008).

As an example of a syntactic combination function of RGL, consider a function for predication with two-place adjectives. This function takes three arguments: a two-place adjective, a subject noun phrase, and a complement noun phrase. It returns a sentence as value:

```
pred : A2 -> NP -> NP -> S
```

This function is available in all languages of RGL, even though the details of sentence formation are vastly different in them. Thus, to give the concrete syntax of the abstract (semantic) predicate div x y ("x is divisible by y"), the English grammarian can write

```
div x y = pred (mkA2 "divisible" "by") x y
```

The German grammarian can write

```
div x y = pred (mkA2 "teilbar" durch_Prep) x y
```

which, even though superficially using different forms from English, generates a much more complex structure: the complement preposition durch_Prep takes care of rendering the argument y in the accusative case, and the sentence produced has three forms, as needed in grammatically different positions (x ist teilbar durch y in main clauses, ist x teilbar durch y after adverbs, and x durch y teilbar ist in subordinate clauses).

The syntactic combinations of the RGL have their own abstract syntax, but this abstract syntax is not the interlingua of translation: it is only used as a library for implementing the semantic interlingua, which is based on an ontology and abstracts away from syntactic structure. Thus the translation equivalents in a multilingual grammar need not use the same syntactic combinations in different languages. Assume, for the sake of argument, that x is divisible by y is expressed in Swedish by the transitive verb construction y delar x (literally, "y divides x"). This can be expressed easily by using the transitive verb predication function of the RGL and switching the subject and object,

```
div x y = pred (mkV2 "dela") y x
```

Thus, even though GF translation is interlingua-based, there is a component of transfer between English and Swedish. But this transfer is performed when the grammar is compiled. In general, the use of the large-coverage RGL as a library for restricted grammars is called grammar specialization. The way GF performs grammar specialization is based on techniques for optimizing functional programming languages, in particular partial evaluation (Ranta 2004, 2007). GF also gives a possibility to run-time transfer via semantic actions on abstract syntax trees, but this option has rarely been needed in previous applications, which helps to keep translation systems simple and efficient.

As shown in Figure 1, the RGL is currently available for 12 languages, of which 9 are official languages of the European Union: Bulgarian, Danish, English, Finnish, French, German, Italian, Spanish, and Swedish. The other 3 are Catalan, Norwegian, and Russian. Work is in progress for several more languages, so that a complete inflectional morphology and large parts of syntax are already available for two EU languages (Polish and Romanian) as well as for Arabic and Hindi/Urdu. A collaborative project has been started for extending RGL to new languages: the GF Resource Grammar Summer School was held in August 2009.[27].

Grammars for 16 new languages were started during the Summer School. Many of these are expected to deliver results by the end of 2009: for instance, the EU languages Dutch, Maltese, and Portuguese, the prospective EU languages Icelandic and Turkish, as well as Afrikaans and Japanese. Some of these languages will be integrated in the show-case web service in WP10.

In the MOLTO project, grammar engineering in GF will be further improved in two ways:

- An IDE (**Integrated Development Environment**), helping programmers to use the RGL and manage large projects.
- **Example-Based Grammar Writing**, making it possible to bootstrap a grammar from a set of example translations.

The former tool is a standard component of any library-based software engineering methodology. The latter technique uses the large-coverage RGL for parsing translation examples, which leads to translation rule suggestions. For example, the German rule for divisibility shown above can be derived from the example

```
div x y = "x ist teilbar durch y"
```

This technique has similarities with the explanation-based learning of the CLE and Regulus projects (Alshawi 1992, Rayner 2006). GF’s grammar specialization by partial evaluation has the advantage of mapping examples directly with the semantic structures of the interlingua.

3.2.4 Translator’s tools

For the translator’s tools, there are three different use cases:

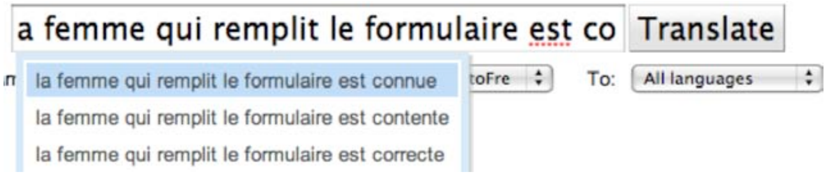
- restricted source
    - production of source in the first place
    - modifying source produced earlier
- unrestricted source

Working with restricted source language recognizable by a GF grammar is straightforward for the translating tool to cope with, except when there is ambiguity in the text. The real challenge is to help the author to keep inside the restricted language. This help is provided by predictive parsing, a technique recently developed for GF (Angelov 2009). Incremental parsing yields word predictions, which guide the author in a way similar to the T9 method[28] in mobile phones. The difference from T9 is, however, that GF’s work prediction is sensitive to the grammatical context. Thus it does not suggest all existing words, but only those words that are grammatically correct in the context. Figure 2 shows an example of the parser at work. The author has started a sentence as la femme qui remplit le

[27] digitalgrammars.com/gf/doc/gf-summerschool.html

[28] www.t9.com

formulaire est co ("the woman who fills the form is co"), and a menu shows a list of words beginning with co that are given in the French grammar and possible in the context at hand; all these words are adjectives in the feminine form.



a femme qui remplit le formulaire est co

la femme qui remplit le formulaire est connue
la femme qui remplit le formulaire est contente
la femme qui remplit le formulaire est correcte

Figure 2: French word prediction in GF parser, suggesting feminine adjectives that agree with the subject *la femme.*

Notice that the very example shown in Figure 2 is one that is difficult for n-gram-based statistical translators: the adjective is so far from the subject with which it agrees that it cannot easily be related to it.

Predictive parsing is a good way to help users produce translatable content in the first place. When modifying the content later, e.g. in a wiki, it may not be optimal, in particular if the text is long. The text can contain parts that depend on each other but are located far apart. For instance, if the word femme ("woman") in the previous example is changed to homme, the preceding article la has to be changed to l', and the adjective has to be changed to the masculine form: thus connue ("known") would become connu, and so on. Such changes are notoriously difficult even for human authors and translators, and can easily leave a document in an inconsistent state. This is where another utility of the abstract syntax comes in: in the abstract syntax tree, all that is changed is the noun, and the regenerated concrete syntax string automatically obeys all the agreement rules. The process is shown in Figure 3. The one-word change generating the new set of documents can be performed by editing any of the three representations: the tree, the English version, or the French version. This functionality is implemented in the GF syntax editor (Khegai & al. 2003).

Pred known_A (Rel **woman_N** (Compl fill_V2 form_N))
the woman who fills the form is known
la femme qui remplit le formulaire est connue
–>
Pred known_A (Rel **man_N** (Compl fill_V2 form_N))
the **man** who fills the form is known
l' **homme** qui remplit le formulaire est *connu*

Figure 3: Change in one word (boldface) propagated to other words depending on it (italics).

Restricted languages in the sense of MOLTO are close to controlled languages, such as Attempto (Fuchs & al. 2008); the examples shown in this section are actually taken from a GF implementation of Attempto Controlled English generalized to five languages (Ranta and Angelov 2009). However, unlike typical controlled languages, MOLTO does not require the absence of ambiguity. In fact, when a controlled language is generalized to new languages, lexical ambiguities in particular are hard to avoid.

The predictive parser of GF does not try to resolve ambiguities, but simply returns all alternatives in the parse chart. This is not always a problem, since it may be the case that the target language has exactly the same ambiguity and then it remains hidden in the translation. In practise this happens often in closely related languages. But if the ambiguity makes a difference in translation, it has to be resolved. There are two complementary approaches: using statistical models for ranking or using manual disambiguation. The statistical model can be used to compute the default most likely alternative, but as it may fail to produce the right prediction, the possibility of manual intervention is necessary. The syntax editor is very powerful in this case because it shows the entire abstract syntax tree, allowing a user to make such adjustments relatively easily. For users less versed in abstract syntax, however, a better choice is to show the ambiguities as different translation results. Then the user just has to select the right alternatives. The choice is propagated back in the abstract syntax, which has the cumulative effect that a similar ambiguity in a third language gets fixed as well. This turns out to be very useful in a collaborative environment such as Wikipedia.

Both predictive parsing and syntax editing are core functionalities of GF and work for all multilingual grammars. While the MOLTO project will exploit these functionalities with new grammars, it will also develop them into tools fitting better into users' work flows. Thus the tools will not require the installation of specific GF software: they will work as plug-ins to ordinary tools such as web browsers, text editors, and professional translators' tools such as

SDL[29] and WordFast[30]. The snapshot in Figure 2 is from an actual web-based translation prototype using GF. It shows a slot in an HTML page, built by using JavaScript via the Google Web Toolkit (Bringert & al. 2009). The translation is performed in a server, which is called via HTTP. Also client-side translators, with similar user interfaces, can be built by converting the whole GF grammar to JavaScript (Meza Moreno and Bringert 2008).

To deal with unrestricted legacy input, such as in the patent case study, predictive parsing and syntax editing are not enough. The translator will then be given two alternatives: to extend the grammars, or to use statistical translation. For grammar extension, some functionalities of the grammar writer's tools are made available to the translator—in particular, lexicon extension (to cope with unknown words) and example-based grammar writing (to cope with unknown syntactic structures). In statistical translation, the worst-case solution is to fall-back to phrase-based statistical translation. In MOLTO, we will study the ways to specialize this to translation in limited domains, so that the quality is higher than in general-purpose phrase-based translation. We will also study other methods to help translators with unexpected input.

1.2.5 Multilingual services

MOLTO will provide a unique platform for multilingual document management, satisfying the five desired features listed in Section 1.1. It will enable truly collaborative creation and maintenance of content, where input provided in any language of the system is immediately ported to the other languages, and versions in different languages are thereby kept in synchrony. This idea has had previous applications in GF (Dymetman & al. 2000, Khegai & al. 2003, Meza Moreno and Bringert 2008). In MOLTO, it will be developed into a technology that can be readily applied by non-experts in GF to any domain that allows for an ontology-based interlingua.

The methodology will be tested on three substantial domains of application: mathematics teaching material, patents, and museum object descriptions. These case studies are varied enough to show the generalisability of the MOLTO technology, and also extensive enough to produce useful prototypes for end users of translations: mathematics students, intellectual property researchers, and visitors to museums. End users will have access in their own languages to information that may be originally produced in other languages.

The MOLTO set-up not only produces translations of documents, but it can also enhance queries about them. The idea of controlled-language queries on semantic web documents has been developed previously (Damljanovic and Bontcheva 2008, Fuchs & al. 2008), but almost exclusively for English. In the GF setting, all query technology developed for one language becomes automatically usable in other languages as well. Queries in natural language can be interpreted by the same grammars that perform translation: even in the cases where the translation grammars themselves don't cover questions, they do provide all domain-specific vocabulary, and question forms can be inherited from the RGL.

Regarding education in the EU, every day around 25 million students take science and engineering lessons at high school and university (source: Eurostat[31]). All these require mathematical training and, more specifically, problem-solving training. Learning-by-doing is usually imparted in math courses through exercises.

In the last years, several web-based systems have appeared which allow assigning a different exercise to each student and having it assessed automatically. Nevertheless, these systems are limited in scope to the simplest problems: The ones requiring the student to carry out an algorithm; besides, they force the student to use a proprietary syntax.

According to the idea that the point of mathematical education is insight, not numbers, one of the main goals of such training should be to provide the students with the skills to deal with real world situations that require modeling as much as solving. Word problems describe a simplified real-world situation where some unknown quantity is to be deduced by the student by using his/her mathematical skills.

Students will benefit from having a dialog system that assists in building such a model by pointing out inconsitencies and this system can be implemented as a multilingual query interface to a Computer Algebra System and/or Proof Assistant.

3.2.5 Robust and statistical translation methods

Grammar-based translation works only for a language fragment determined by a grammar, but in real-life translation

[29] www.sdl.com
[30] www.wordfast.net
[31] http://epp.eurostat.ec.europa.eu/

it may not be guaranteed that all input lies within this fragment. Moreover, in real-life scenarios one should be able to cope with incorrect, ungrammatical and non-formal (e.g., lack of punctuation, use of shortened word forms, etc.) language. This challenge will be approached by investigating the use of robust parsing and statistical translation in continuum with grammar-based translation. The robust methods can be applied on two levels: directly on the source text as a fall-back to grammar-based translation, and also as a method of improving the grammar on the fly, possibly in interaction with a human translator.

Statistical Machine Translation (SMT) is a common paradigm for Machine Translation which offers robustness and flexibility, especially when one has a large amount of parallel texts available. From the first works on SMT by Brown et al. Brown et al. (1990), the field has experienced notable enhancements. It was soon noticed that translation is not a word to word process, that the information of surrounding words would help and that one word could be translated into more than one element. This motivated the usage of phrases as translation units in the so-called Phrase-Based SMT Och and Ney (2004); Koehn et al. (2003). In SMT, the best translation for a given source sentence is the most probable one, and the probability is expressed as the sum of different components. The log-linear model Och and Ney (2002), a generalisation of the original noisy-channel approach, estimates the probability as the logarithmic sum of several terms. Two of them, the language model and the translation model, are the core of the approach, but other probabilistic terms, such as distortion, word penalty, etc. are usually in the recipee. The search for the most probable translation is often referred to as decoding. State of the art decoders (e.g., Koehn et al. (2007)) make use of dynamic programming and approximate search to explore the huge space of possible translations efficiently.

Moses Koehn et al. (2007) is a widely used phrase-based SMT system, implementing the above mentioned log-linear approach (also known as factored models). Moses setting is designed to be especially flexible at using any probabilistic component defined by the user (called model features) in the log-linear estimation of the translation probability. Moses has become a de facto standard for phrase-based SMT systems and is typically used for comparison in any new proposed SMT method. The concrete objectives in this proposal around robust and statistical MT are:

- Extend the grammar-based approach by introducing probabilistic information and confidence scored predictions.
- Construct a GF domain grammar and a domain-adapted state-of-the-art SMT system for the *Patents* use case.
- Develop combination schemes to integrate grammar-based and statistical MT systems in a hybrid approach.
- Fulfil the previous objectives on a variety of language pairs of the project (covering three languages at least).

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 (1 and 2). We will compile and annotate general-purpose large bilingual and monolingual corpora for training basic SMT systems. This compilation will rely on publicly available corpora and resources for MT (e.g., the multilingual corpus with transcriptions of European Parliament Sessions).

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). This corpora will come from the compilation to be made at WP7, leaded by Mxw. The UPC team has experience at performing adaptation of SMT systems Giménez and Màrquez (2006); Garcì a et al. (2009).

Another source for domain specific corpora is the automatic synthesis of aligned translations generated with the GF grammars in this domain. Grammar induction is based partly on traditional phrase alignment techniques, partly on the GF Resource Grammar Library (RGL). In grammar-based MOLTO translation, RGL is specialized to domain-specific tasks to maximize efficiency, reduce ambiguity, minimize the need of transfer, and guarantee idiomatic translation. However, such domain-specific grammars tend to have restricted coverage, which leads to reasonable input being out-of-grammar. To fill the gaps, statistically based smoothing can be used. The method is inspired by the technique used in the TALK project to improve the robustness of spoken language models via a synthetic corpus (Jonson 2006). Its usage for translation is even more promising than for dialogue systems, because we do not need to return semantic values but just translations, which are always guaranteed.

Combination of grammar-based and statistical paradigms is a novel and active research line in MT. In MOLTO, we depart from three key assumptions when facing the combination of paradigms: 1) the quality of a completely translated sentence by a GF-based system will be always better than the translation obtained with SMT; 2) When the GF-based systems fails at producing a complete translation it can probably produce a set of partial translations (phrases) with confidence scores or probabilities; 3) The SMT system is always capable of generating an output translation (although the quality can be very low at certain extreme cases. Assumption number one implies that our combination setting will be set as a fallback strategy, i.e., SMT will be seen as a back-off for GF-based MT. Assumption number two makes it possible to combine partial outputs from GF with the SMT system in a real hybrid approach. Assumption number three guarantees that a translation will be always output by the combined system. We plan explore several instantiations of the fallback approach. From simple to complex:

- *Independent combination*: in this case, the combination is set as a cascade of independent processors. When Grammar-based MT does not produce a complete translation, the SMT system is used to translate the input sentence. This external combination will be set as the baseline for the rest of combination schemes.
- Construction of a *hybrid system* based on both paradigms. In this case, a more ambitious approach will be followed, which consists of constructing a truly hybrid system which incorporates an inference procedure able to deal with multiple proposed fragment translations, coming from grammar-based and SMT systems. Again we envision several variants:
  - Fix translation phrases produced by the partial GF analyses in the SMT search. In this variant we assume that the partial translations given by GF are correct so we can fix them and let SMT to fill the remaining gaps and do the appropriate reordering. This hard combination is easy to apply but not very flexible.
  - Use translation phrase pairs produced by the partial GF analyses, together with their probabilities, to form an extra feature model for the Moses decoder (probability of the target sentence given the source).
  - Use *tree fragment pairs* produced by the partial GF analyses, together with their probabilities, to feed a *syntax* based SMT model, such as the one by Carreras and Collins (2009) . In this case the search process to produce the most probable translation is a probabilistic parsing scheme.

Some work can be found in the MT literature regarding the combination of systems, under the Multi-Engine–MT label Chen et al. (2007); Matusov et al. (2006); Macherey and Och (2007); Mellebeek et al. (2006); Huang and Papineni (2007); Rosti et al. (2007); Karakos et al. (2008). All the papers on Multi-engine MT reach similar conclusions: combining the outputs results in a better translation. Most of the approaches generate a new consensus translation combining different SMT systems using different language models and in some cases combining also with rule-based MT systems. Some of the approaches require confidence scores for each of the outputs. The improvement in translation quality is around 18% relative increasing in BLEU score. Also remarkable is the work on training SMT systems for post-editing the output of a rule-based MT system Terumasa (2007); Simard et al. (2007). Significant improvements are obtained, especially in out-of-domain test corpora.

In MOLTO, we expect to progress beyond state-of-the-art in several aspects. The main novelties presented in this project regarding MT system combination are the following:

- The GF-based system will be used to help adapting the SMT system to the particular textual domain.
- The combination is pivoting on the interlingua GF approach (focusing on translation quality as the main aspect). SMT is used as the framework for combining partial GF-based analyses with pure statistical features.
- The hybrid combination approach will allow to have the individual MT systems making *on-line* confidence-rated translation predictions on the sentence under a unified search scheme (decoder).
- SMT will not be restricted to phrase-based models. Syntax-based SMT models will be included in the hybridization.

MOLTO shares the character of a hybrid approach with the project EuroMatrix[32] and its successor EuroMatrixPlus[33], and will make use of tools created in these projects, in particular the Moses system. The starting point, however, is a complete opposite: in EuroMatrix and EuroMatrixPlus, the starting point is large-coverage statistical translation whose quality is increased by adding linguistic rules. MOLTO's starting point is high-quality translation whose coverage is increased by adding statistical components.

3.2.6 Productivity and usability

Our case studies will show that it is possible to build a completely functional high-quality translation system for a new application in a matter of months—for small domains in just days. The effort to create a system dynamically applicable to an unlimited number of documents will be essentially the same as the effort it currently takes to manually translate a set of static documents. The expertise needed for producing a translation system will be low, essentially amounting to the skills of an average programmer who has practical knowledge of the targeted language

[32] www.euromatrix.net

[33] www.euromatrixplus.net

and of the idiomatic vocabulary and syntax of the domain of translation. The expertise needed for using the translation system will be minimal, due to the guidance provided by MOLTO.

3.2.7 Translation quality

We will compare the results of MOLTO to other translation tools, by using both automatic metrics (BLEU, Bilingual Evaluation Understudy, Papineni & al. 2002) and, in particular, the human evaluation of “utility”, as defined by TAUS. The comparison is performed with the freely available general-purpose tools Google translate and Systran. While the comparison is “unfair” in the sense that MOLTO is working with special-purpose domain grammars, we want to perform measurements that confirm that MOLTO’s quality really is essentially better. Comparisons with domain-specific systems will be performed as well, if any such systems can be found. Domain-specific translation systems are still rare and/or not publicly available.

Regarding automatic metrics for MT, the usage of lexical n-gram based metrics (WER, PER, BLEU, NIST, ROUGE, etc.) represents the usual practice in the last decade. However, recent studies showing some limitations of lexical metrics at capturing certain kind of linguistic improvements and making appropriate rankings of heterogeneous MT systems Callison-Burch et al. (2006); Callison-Burch et al. (2007); Callison-Burch et al. (2008); Giménez (2008) have fostered research on more sophisticated metrics, which can combine several aspects of syntactic and semantic information. The IQmt suite[34], developed by the UPC team, is one of the examples in this direction Giménez and Amigó (2006); Giménez and Màrquez (2008). In IQmt, a number of automatic metrics for MT, which exploit linguistic information from morphology to semantics, are available for the English language and will be extended to other languages (e.g., Spanish) soon. These metrics are able to capture more subtle improvements in translation and show high correlation with human assessments Giménez and Màrquez (2008); Callison-Burch et al. (2008). We plan to use IQmt in the development cycle whenever it is possible. For languages not covered in IQmt, we will rely on BLEU (Papineni et al. 2002).

Regarding human evaluation, the TAUS method is the more appropriate one for the MOLTO tasks, since we are aiming for reliable rendering of information. It consists of inspection of a significant number of source/target segments to determine the effectiveness of information transfer. The evaluator first reads the target sentence, then reads the source to determine whether additional information was added or misunderstandings identified. The scoring method is as follows:

4. Complete: All of the information in the source was available from the target; reading the source did not add to information or understanding.
3. Useful: The information in the target was correct and clear, but reading the source added some additional information or understanding.
2. Marginal: The information in the target was correct, but reading the source provided significant additions or clarifications.
1. Poor: The information in the target was unclear and/or incorrect; reading the source would be necessary for understanding.

We aim to reach “complete” scores in mathematics and museum translation, and “useful” scores in patent translation.

Dimensions not mentioned in the TAUS scoring are “grammaticality” and “naturalness” of the produced text. The grammar-based method of MOLTO will by definition guarantee grammaticality; failures in this will be fixed by fixing the grammars. Some naturalness will be achieved in the sense of “idiomaticity”: the compile-time transfer technique presented in Section 1.2.3 will guarantee that forms of expression which are idiomatic for the domain are followed. The higher levels of text fluency reachable by Natural Language Generation techniques such as aggregation and referring expression selection have been studied in some earlier GF projects, such as (Burke and Johannisson 2005). Some of these techniques will be applied in the mathematics and cultural heritage case studies, but the main focus is just on rendering information correctly. On all these measures, we expect to achieve significant improvements in comparison to the available translation tools, when dealing with in-grammar input.

[34] http://www.lsi.upc.edu/ nlp/IQMT/

3.3 S/T Methodology and associated work plan

3.3.1 Overall strategy and general description

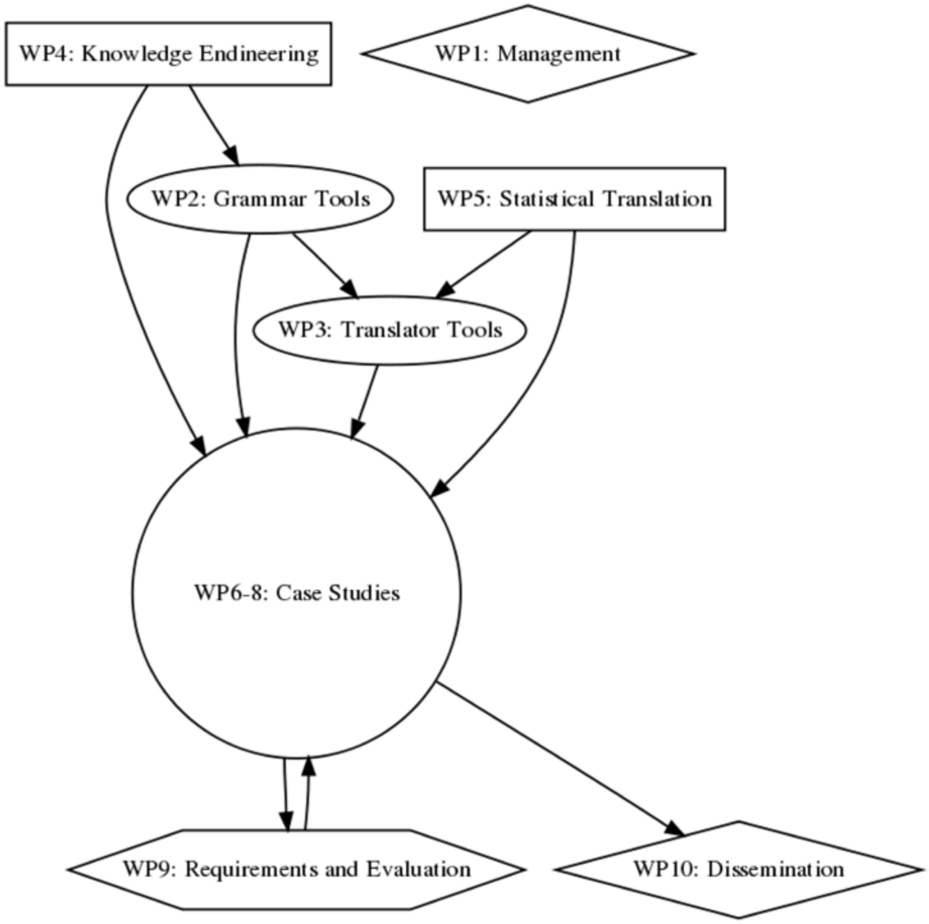
The project will develop tools and applications in parallel. The leading idea is to have working prototypes from the beginning, and deliver updates frequently. The work is divided into four kinds of packages:

- Management and dissemination: WP1 and WP10. These run throughout the project.
- Generic tools: WP2–5. These start early in the project.
- Case studies: WP6–8. These start later than the tools, because they assume some maturity of the tools. However, some of them also involve data collection, which can be started earlier.
- Requirements and evaluation: WP9. This runs throughout the project. Its purpose in the beginning is to define the requirements for both the generic tools and the case studies in a coherent way that can lead to maximal synergy between work packages, (the case studies are otherwise independent of each other). Later in the project, WP9 performs evaluation and delivers feedback. In the last phase of the project, when the development of new functionalities in tools and case studies has stopped (month 30), WP9 takes care of bug fixing and consolidation of the tools and case studies, so that everything remains coherent.

The dependencies among work packages are shown in Section 1.3.4 below. Since the dependencies are few, and localized in well-defined deliverables, many of the work packages run in parallel, as shown in the Gantt chart in the next section.

3.3.1.1 Dependencies among work packages

The following figure shows the dependences between work packages. Work packages of different types are shown by using different forms: rectangle = basic technology and research, ellipse = generic tools, circle = case study, hexagon = requirements and evaluation, diamond = management and dissemination.



The two-way dependency between WP6-8 and WP9 is due to the two different functions of WP9: it identifies user requirements for the case studies in the beginning of the project, and evaluates their results later in the project.

3.3.1.2 Risk assessment and contingency plan

One strength of the MOLTO project is that the core technology used in it is owned by the partners. Therefore we are not vulnerable to typical risks arising from sudden changes in the functionality or availability of external tools.

But here are some other risks we have identified, connected with the project's milestones (Section 1.3.7).

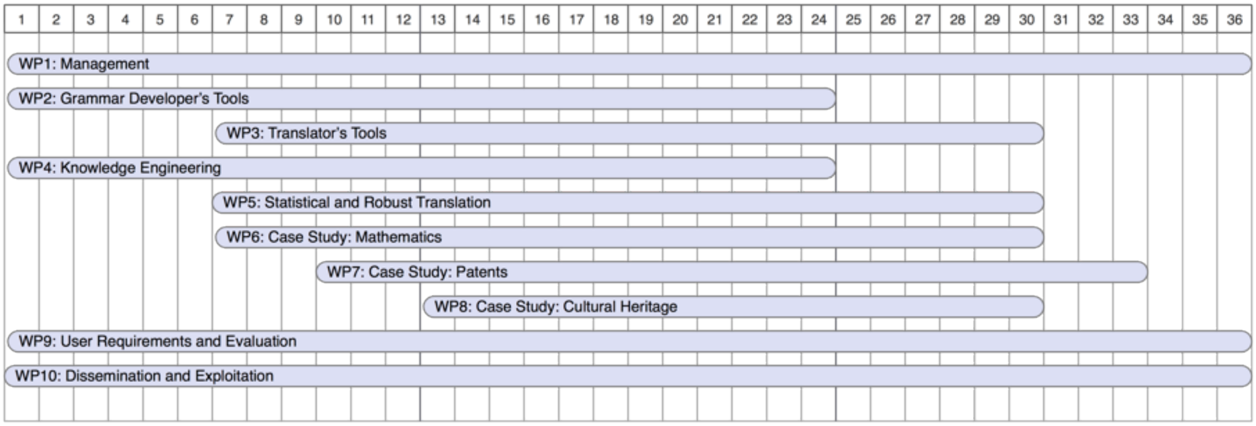
- **MS1: 15 languages in the library.** *We may have difficulties in reaching the goal of 15 languages.* Since the RGL will be available in more than 15 languages, the risk concerns our capacities to develop the domain grammars and lexica. As shown in Section 2.3, the key persons of the consortium already cover 10 languages; for the remaining ones, we have the possibility to hire short-time project workers from the wide student base of UGOT, UHEL, and UPC. Since we have not specified exactly which languages we cover, this will be possible to arrange.
- **MS2: Knowledge representation infrastructure.** *Retrieval access for the consortium may not be satisfactory.* Retrieval access is mainly needed in the case studies. But it is not needed in the beginning of the case studies, so there is a few months time to solve the problems after the projected M6.
- **MS3: Web-based translation tool available.** *The tool may not be satisfactory for all uses users.* There is plenty of time to improve the tool during the lifetime of WP3; the first release of the tool is on purpose made early, so that we can collect user feedback and solve remaining problems.
- **MS4: Grammar-ontology interoperability.** *The OWL-to-GF mapping may not be adequate for all uses of OWL.* We may have to rule out some legacy uses of OWL from the scope of the tool; as a further support for this, the manual of the tool will specify the best practices to guarantee that ontologies interoperate with GF.
- **MS5: First prototypes of the cascade-based combination models.** *The model does not show significant improvements in evaluation.* This cascade-based model is the most modest of the statistical techniques, and if its performance is weak, the more advanced techniques developed later are hoped to replace it.

- **MS6: Grammar tool complete.** *Example-Based Grammar Writing may prove not good enough to infer the resource grammar constructs from examples alone.* This problem can be solved by recourse to predictive parsing when producing the examples, and by using the IDE in a traditional manner, for browsing the resource grammar library.
- **MS7: First prototypes of hybrid combination models.** *The model does not show significant improvements in evaluation.* This model is aimed to improve upon the more basic cascade-based model in M5.1. If the performance is still too weak, we have 6 months to improve the hybrid translator before the final version in M5.3.
- **MS8: Translation tool complete.** *The model does not show significant improvement in comparison to pure grammar-based models.* If the most advanced hybrid model doesn't fulfil the expectations, MOLTO will have to rely its starting point, purely grammar-based models. These will be sufficient for the case studies in WP6 and WP8, but we may have to cut down the ambitions in WP7. The reasons why our hybrid models failed will be an interesting scientific result anyway, since such models are a focus area in the forefront of machine translation.
- **MS9: Case studies complete.** *In the mathematics case study, an exercise may be too complex for the reasoner, or it can be solved in just one step (which is useless to the student).* In the first case, we should restrict the kind of exercises to consider or use a more complex reasoner that can be driven by tactics. For the remaining exercises in which semi-automatic solving fail, we should provide a step-by-step solution method.

  In the patent case study, there is a risk as to whether there are enough examples of a sufficient quality in a particular language to be useful for training the SMT engine. To neutralize this risk, we keep the set of languages flexible with phrases such as "at least 3 languages" and " candidate languages" in this case study.

  In the museum case study, the fact database of Gothenburg City Museum might not provide sufficient information for the texts we want to cover. The database currently has descriptions of 30,000 artefacts. We can spend some time in the project to enrich the data if needed.

3.3.2 Timing of work packages and their components



3.3.3 Work package list/overview

3.3.3.1 List of work packages

WP No.	WP title	Activity	Leader	PMonths	Start	End
WP1	Management	MGT	1 UGOT	20	M01	M36
WP2	Grammar Developer's Tools	RTD	1 UGOT	48	M01	M24
WP3	Translator's Tools	RTD	2 UHEL	56	M07	M30
WP4	Knowledge Engineering	RTD	4 Ontotext	45	M01	M24

WP5	Statistical and Robust Translation	RTD	3 UPC	50	M07	M30
WP6	Case Study: Mathematics	RTD	3 UPC	36	M07	M30
WP7	Case Study: Patents	RTD	3 UPC	42	M10	M33
WP8	Case Study: Cultural Heritage	RTD	1 UGOT	29	M13	M30
WP9	User Requirements and Evaluation	RTD	2 UHEL	27	M01	M36
WP10	Dissemination and Exploitation	MGT	1 UGOT	37	M01	M36

3.3.3.2 Deliverables list

Del. No.	Deliverable title	WP	Nat.	Level	Date
D1.1	Work Plan for MOLTO	WP1	R	CO	M1
D10.1	Dissemination plan, with monitoring and assessment	WP10	R	CO	M3
D10.2	MOLTO web service, first version	WP10	P	PU	M03
D9.1	MOLTO test criteria, methods and schedule	WP9	R	PU	M06
D1.2	Periodic management report 1	WP1	R	CO	M07
D4.1	Knowledge Representation Infrastructure	WP4	RP	PU	M08
D2.1	GF Grammar Compiler API	WP2	P	PU	M12
D1.3	Periodic management report 2	WP1	R	CO	M13
D4.2	Data Models, Alignment Methodology, Tools and Doc.	WP4	RP	PU	M14
D2.2	Grammar IDE	WP2	P	PU	M18
D3.1	MOLTO translation tools API	WP3	P	PU	M18
D4.3	Grammar - Ontology	WP4	P,M	PU	M18

	Interoperability				
D5.1	Description of the final collection of corpora	WP5	RP	PU	M18
D6.1	Simple drill grammar library	WP6	P	PU	M18
D8.1	Ontology and corpus study of the cultural heritage domain	WP8	O	PU	M18
D1.4	Periodic management report 3	WP1	R	CO	M19
D7.1	Patent MT and Retrieval Prototype Beta	WP7	P	PU	M21
D3.2	MOLTO translation tools prototype	WP3	P	PU	M24
D6.2	Prototype of commanding CAS	WP6	P	PU	M24
D2.3	Grammar tool manual and best practices	WP2	RP,M	PU	M24
D5.2	Description and evaluation of the combination prototypes	WP5	RP	PU	M24
D8.2	Multilingual grammar for museum object descriptions	WP8	P	PU	M24
D1.5	Periodic management report 4	WP1	R	CO	M25
D7.2	Patent MT and Retrieval Prototype	WP7	P	PU	M27
D3.3	MOLTO translation tools workflow manual	WP3	RP,M	PU	M30
D5.3	WP5 final report: statistical and robust MT	WP5	RP,M	PU	M30
D6.3	Assistant for solving word problems	WP6	P,M	PU	M30
D8.3	Translation and retrieval system for museum object descriptions	WP8	P	PU	M30
D1.6	Periodic management	WP1	R	CO	M31

	report 5				
D7.3	Patent Case Study Final Report	WP7	RP,M	PP	M33
D9.2	MOLTO evaluation and assessment report	WP9	R,M	PU	M36
D10.3	MOLTO web service, final version	WP10	P	PU	M36
D10.4	MOLTO Dissemination and Exploitation Report	WP10	R,M	PU	M36
D1.7	Final management report	WP1	R	CO	M36
D.X	Reporting deliverables as well as public events/documents as detailed in Appendix X	WP 1, 7,8,10	see Appendix X	PU	see Appendix X

Main deliverables of each column marked as “M” in the “Nat.” column. Regular publications marked as “RP”, other reports as “R”, prototypes as “P”.

The Consortium will perform the tasks, deliver the outputs and take part in the events stipulated in Appendix X to this Description of Work.

3.3.4 Work package descriptions

WP No 1 **Leader** UGOT **Start** M1 **End** M36
WP Title Management
Activity type MGT

Beneficiary number	**1**	**2**	**3**	**4**	**5**	**total**
Beneficiary short name	UGOT	UHEL	UPC	Ontotext	Mxw	-
Person months	10	3	3	3	1	20

Objectives

The management WP has as its objective to keep the project running, guarantee the timely delivery of status reports, monitor the economical balance, and ensure communication between the partners and between the consortium and the Commission.

This work package is responsible for the overall coordination and financial management of the network. Among the duties are: directing the work to be done, monitoring the performance of the project partners, and communications with the Commission. The first deliverable is the Consortium Agreement which will define the terms of co-operation and the division of the ownership of IPRs.

Together with WP10 this work package is also responsible for setting up the infrastructure for communication and dissemination. This includes a web-based system which integrates a wiki, bug tracking and software development management, a portal with both a private and a public side, and a conferencing system for the project.

Description of work

The Coordinator takes care of communication with the Commission. Each partner has a Site Leader, who participates to reporting. A part-time Project Manager takes care of day-to-day administrative management. The Site Leaders and the Project Manager constitute a Steering Group. The Steering Group will convene in connection to the project meetings, and also at need to resolve conflicts and decide on any major changes in the project. Each Work Package has a Work Package Leader. The project has a kick-off meeting plus two project meetings every year. Each of the participants will be the organizer of at least one of the meetings.

See Appendix X for additional tasks, outputs and events.

Deliverables

Del. no	**Del. title**	**Date**
D 1.1	Work Plan for MOLTO	M1
D 1.2	Periodic management report 1	M7
D 1.3	Periodic management report 2	M13
D 1.4	Periodic management report 3	M19
D 1.5	Periodic management report 4	M25
D 1.6	Periodic management report 5	M31
D 1.7	Final management report	M36

WP No 2 **Leader** UGOT **Start** M1 **End** M24
WP Title Grammar Developer's Tools
Activity type RTD

Beneficiary number	**1**	**2**	**3**	**4**	**total**
Beneficiary short name	UGOT	UHEL	UPC	Ontotext	-
Person months	20	12	4	12	48

Objectives

The objective is to develop a tool for building domain-specific grammar-based multilingual translators. This tool will be accessible to users who have expertise in the domain of translation but only limited knowledge of the GF formalism or linguistics. The tool will integrate ontologies with GF grammars to help in building an abstract syntax. For the concrete syntax, the tool will enable simultaneous work on an unlimited number of languages and the addition of new languages to a system. It will also provide linguistic resources for at least 15 languages, among which at least 12 are official languages of the EU.

Description of work

The top-level user tool is an IDE (Integrated Development Environment) for the GF grammar compiler. This IDE provides a test bench and a project management system. It is built on top of three more general techniques: the GF Grammar Compiler API (Application Programmer's Interface), the GF-Ontology mapping (from WP4), and the GF Resource Grammar Library. The API is a set of functions used for compiling grammars from scratch and also for extending grammars on the fly. The Library is a set of wide-coverage grammars, which is maintained by an open source project outside MOLTO but will be via MOLTO efforts made accessible for programmers on lower levels of linguistic expertise. Thus we rely on the available GF resource grammar library and its documentation, available through digitalgrammars.com/gf/lib. The API is also used in WP3, as a tool for limited grammar extension, mostly with lexical information but also for example-based grammar writing.

UGOT designs APIs and the IDE, coordinates work on grammars of individual languages, and compiles the documentation. UHEL contributes to terminology management and work on individual languages. UPC contributes to work on individual languages. Ontotext works on the Ontology-Grammar interface and contributes to the ontology-related part of the IDE.

Deliverables

Del. no	**Del. title**	**Nature**	**Date**
D 2.1	GF Grammar Compiler API	P	M12
D 2.2	Grammar IDE	P	M18
D 2.3	Grammar tool manual and best practices	RP, Main	M24

WP No 3 **Leader** UHEL **Start** M7 **End** M30
WP Title Translator's Tools
Activity type RTD

Beneficiary number	**1**	**2**	**3**	**4**	**total**
Beneficiary short name	UGOT	UHEL	UPC	Ontotext	-
Person months	12	30	4	10	56

Objectives

The objectives are to (i) build an API for practical translation and production of multilingual documents; (ii) web-based front-end to the multilingual translators; allowing (iii) translation, example-based grammar authoring, syntax edition, context-sensitive word completion, and multilingual ontology-based lexicon building.

Description of work

The standard working method in current translation tools is to work on the source and translation as a bilingual text. Translation suggestions are sought from TM (Translation Memory) based on similarity, or generated by a MT system, are presented for the user to choose from and edit manually. The MOLTO translator tool extends this with two additional constrained-language authoring modes, a robust statistical machine translation (UPC) mode, plus vocabulary and grammar extension tools (UGOT), including: (i) mode for authoring source text while context-sensitive word completion is used to help in creating translatable content; (ii) mode for editing source text using a syntax editor, where structural changes to the document can be performed by manipulating abstract syntax trees; (iii) back-up by robust and statistical translation for out-of-grammar input, as developed in WP5; (iv) support of on-the-fly extension by the translator using multilingual ontology-based lexicon builder; and (v) example-based grammar writing based on the results of WP2.

The WP will build an API (D3.1, UHEL) and a Web-based translator tool (D3.2, by Ontotext and UGOT). The design will allow the usage of the API as a plug-in (UHEL) to professional translation memory tools such as SDL and WordFast. We will apply UHEL's ContentFactory for distributed repository system and a collaborative workflow for multilingual terminology.

Deliverables

Del. no	**Del. title**	**Nature**	**Date**
D 3.1	MOLTO translation tools API	P	M18
D 3.2	MOLTO translation tools prototype	P	M24
D 3.3	MOLTO translation tools / workflow manual	RP, Main	M30

WP No 4 **Leader** Ontotext **Start** M1 **End** M24
WP Title Knowledge Engineering
Activity type RTD

Beneficiary number	**1**	**2**	**4**	**total**
Beneficiary short name	UGOT	UHEL	Ontotext	-
Person months	3	12	30	45

Objectives

The objectives of WP4 are (i) research and development of two-way grammar-ontology interoperability bridging the gap between natural language and formal knowledge; (ii) infrastructure for knowledge modeling, semantic indexing and retrieval; (iii) modelling and alignment of structured data sources; (iv) alignment of ontologies with the grammar derived models.

Description of work

We will provide knowledge representation infrastructure (D4.1, by Ontotext); aligned semantic models and instance bases (D4.2, by Ontotext and UHEL); two-way grammar-ontology and NL (Natural Language) to ontology interoperability (D4.3, by Ontotext and UGOT). The knowledge engineering infrastructure of MOLTO is based on pre-existing products based on open standards to ensure a mature basis. The infrastructure will provide for the storage and retrieval of both knowledge and content covering all modalities of the data. We will adapt and deliver the knowledge representation infrastructure accompanied with documentation of the technology building blocks, overall architecture, standards used, query languages and inference rules.

Having the knowledge engineering infrastructure in place, the partners will focus on building the conceptual models and knowledge bases needed for grammar development (WP2) and the use cases of MOLTO (WP6-8) - one base set and three specialized knowledge sets for the use cases. The base will be a set based on the PROTON ontology, extended with a large coverage knowledge base focused on named entities and a thesaurus. The specialized sets will include the necessary domain specific models and instances, e.g. multi-lingual patent classification taxonomies, museum ontology and instance base, etc. To ensure reuse we will use a semantic alignment methodology paired with a set of data source transformation tools for each of the structured data sources.

The WP will deliver an engine for dual way grammar to ontology interoperability. The engine will allow semi-automatic creation of abstract grammars from ontologies; deriving ontologies from grammars, and instance level knowledge from NL. In terms of retrieval, NL queries will be transformed to semantic queries and the resulting knowledge, expressed back in NL.

Deliverables

Del. no	**Del. title**	**Nature**	**Date**
D 4.1	Knowledge Representation Infrastructure	RP	M8
D 4.2	Data Models, Alignment Methodology, Tools and Documentation	RP	M14
D 4.3	Grammar-Ontology Interoperability	P,Main	M18

WP No 5 **Leader** UPC **Start** M07 **End** M30
WP Title Statistical and Robust Translation

Beneficiary number	**1**	**2**	**3**	**5**	**total**
Beneficiary short name	UGOT	UHEL	UPC	Mxw	-
Person months	9	3	38	0	50

Objectives

The goal is to develop translation methods that complete the grammar-based methods of WP3 to extend their coverage and quality in unconstrained text translation. The focus will be placed on techniques for combining GF-based and statistical machine translation. The WP7 case study on translating Patents text is the natural scenario to test the techniques developed in this package. Existing corpora for the WP7 will be used to adapt SMT and grammar-based systems to the Patents domain. This research will be conducted on a variety of languages of the project (at least three).

Description of work

The work in this package is organized in three main lines:

1. Extend the GF domain grammar for the Patents domain developed in WP7 by introducing probabilistic predictions.
2. Adapt a state-of-the-art SMT system to the Patents domain, by using in-domain multilingual corpora provided by WP7 and synthetic aligned corpora generated in a controlled environment by the GF grammar from (1). All corpora used for domain adaptation will have to be pre-processed with linguistic analyzers.
3. Develop combination approaches to integrate grammar-based and statistical MT models in a hybrid MT system. At least four variants will be studied (i) (*baseline*) cascade of independent MT systems; (ii) (*hard integration*) GF partial output is fixed in a regular SMT decoding (Moses to be used); (iii) (*soft integration I*) GF partial output, in the form of phrase pairs, is integrated as a discriminative probability feature model in a phrase-based SMT system (Moses to be used); (iv) (*soft integration II*) GF partial output, in the form of tree fragment pairs, is integrated as a discriminative probability model in a syntax-based SMT system to be used).

The contribution by partners will be as follows: UGOT will work on the domain GF grammar probabilities and the generation of synthetic corpora for SMT adaptation. UPC will lead the Package, provide the SMT technology (phrase and syntax-based), coordinate the corpora compilation/alignment, and develop the combined MT models. The corpus will be provided by EPO for training and adapting the SMT systems. UHEL will work on the usability aspects of the combined system, which are preparatory for WP3.

Deliverables

Del. no	**Del. title**	**Nature**	**Date**
D 5.1	Description of the final collection of corpora	RP	M18
D 5.2	Description and evaluation of the combination prototypes	RP	M24
D 5.3	WP5 final report: statistical and robust MT	RP,Main	M30

WP No 6 **Leader** UPC **Start** M7 **End** M30
WP Title Case Study: Mathematics
Activity type RTD

Beneficiary number	**1**	**2**	**3**	**4**	**total**
Beneficiary short name	UGOT	UHEL	UPC	Ontotext	-
Person months	3	3	24	6	36

Objectives

The ultimate goal of this package is to have a multilingual dialog system able to help the math student in solving word problems.

Description of work

The UPC team, being a main actor in the past development of GF mathematical grammars and having ample experience in mathematics teaching, will be in charge of the tasks in this work package with help from UGot and UHEL on technical aspects of GF and translator’s tools, along with Ontotext on ontology representation and handling.

We will start by compiling examples of word problems. In parallel, we will take the mathematical multilingual GF library which was developed in the framework of the WebALT project and organize the existing code into modules, remove redundancies and format them in a way acceptable for enhancement by way of the grammar developer’s and translator’s tools of work packages 2 and 3 (D6.1). The next step will be writing a GF grammar for commanding a generic computer algebra system (CAS) by natural language imperative sentences and integrating it into a component (D6.2) to transform the commands issued to the CAS (Maybe as a browser plugin).

For the final deliverable (D6.3), we will use the outcome of work package 4 to add small ontologies describing the word problem: We will end with a multilingual system able to engage the student into a dialog about the progress being made in solving the problem. It will also help in performing the necessary computations.

Deliverables

Del. no	**Del. title**	**Nature**	**Date**
D 6.1	Simple drill grammar library	P	M15
D 6.2	Prototype of commanding CAS	P	M23
D 6.3	Assistant for solving word problems	P,Main	M30

WP No 7 **Leader** UPC **Start** M10 **End** M33
WP Title Case Study: Patents
Activity type RTD

Beneficiary number	**1**	**3**	**4**	**5**	**total**
Beneficiary short name	UGOT	UPC	Ontotext	Mxw	-
Person months	12	15	15	0	42

Objectives

The objectives are to (i) create a commercially viable prototype of a system for MT and retrieval of patents in the bio-medical and pharmaceutical domains, (ii) allowing translation of patent abstracts and claims in at least 3 languages, and (iii) exposing several cross-language retrieval paradigms on top of them.

Description of work

The work will start with the provision of user requirements (WP9) and the preparation of a parallel patent corpus (EPO) to fuel the training of statistical MT (UPC). In parallel UGOT will work on grammars covering the domain and subsequently, together with UPC, apply the hybrid (WP2, WP5) MT on abstracts and claims.

Ontotext will provide semantic infrastructure with loaded existing structured data sets (WP4) from the patent domain (IPC, patent ontology, bio-medical and pharmaceutical knowledge bases, e.g. LLD). Based on the use case requirements, Ontotext will build a prototype (D7.1, D7.2) exposing multiple cross-lingual retrieval paradigms and MT of patent sections.

The accuracy will be regularly evaluated through both automatic (e.g. BLEU scoring or more generally a combination of lexical, semantic and syntactic metrics as defined within the IQmt package) and human based (e.g. TAUS) means (WP9). Ultimately, Ontotext will examine the feasibility of the prototype as a part of a commercial patent retrieval system (D7.3 and WP10).

Deliverables

Del. no	**Del. title**	**Nature**	**Date**
D 7.1	Patent MT and Retrieval Prototype Beta	P	M21
D 7.2	Patent MT and Retrieval Prototype	P	M27
D 7.3	Patent Case Study Final Report	RP, Main	M33

WP No 8 **Leader** UGOT **Start** M13 **End** M30
WP Title Case Study: Cultural Heritage
Activity type RTD

Beneficiary number	**1**	**2**	**3**	**4**	**total**
Beneficiary short name	UGOT	UHEL	UPC	Ontotext	-
Person months	12	6	3	8	29

Objectives

The objective is to build an ontology-based multilingual grammar for museum information starting from a CRM ontology for artefacts at Gothenburg City Museum[35], using tools from WP4 and WP2. The grammar will enable descriptions of museum objects and answering to queries over them, covering 15 languages for baseline functionality and 5 languages with a more complete coverage. We will moreover build a prototype of a cross-language retrieval and representation system to be tested with objects in the museum, and automatically generate Wikipedia articles for museum artefacts in the 5 languages with extensive coverage.

Description of work

The work is started by a study of the existing categorizations and metadata schemas adopted by the museum, as well as a corpus of texts in the current documentation which describe these objects (D8.1, UGOT and Ontotext). We will transform the CRM model into an ontology aligning it with the upper-level one in the base knowledge set (WP4) and modeling the museum object metadata as a domain specific knowledge base. Through the interoperability engine from WP4 and the IDE from WP2, we will semi-automatically create the translation grammar and further extend it (D8.2, UGOT, UHEL, UPC, Ontotext). The final result will be an online system enabling museum (virtual) visitors to use their language of preference to search for artefacts through semantic (structured) and natural language queries and examine information about them. We will also automatically generate a set of articles in the Wikipedia format describing museum artefacts in the 5 languages with extensive grammar coverage (D8.3, UGOT, Ontotext).

Deliverables

Del. no	**Del. title**	**Nature**	**Date**
D 8.1	Ontology and corpus study of the cultural heritage domain	O	M18
D 8.2	Multilingual grammar for museum object descriptions	P	M24
D 8.3	Translation and retrieval system for museum object descriptions	P,Main	M30

[35] www.stadsmuseum.goteborg.se/

WP No 9 **Leader** UHEL **Start** M1 **End** M36
WP Title User Requirements and Evaluation
Activity type RTD

Beneficiary number	**1**	**2**	**3**	**4**	**5**	**total**
Beneficiary short name	UGOT	UHEL	UPC	Ontotext	Mxw	-
Person months	3	10	8	6	0	27

Objectives

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.

Description of work

The work will start with collecting user requirements for the grammar development IDE (WP2), translation tools (WP3), and the use cases (WP6-8). We will define the evaluation criteria and schedule in synchrony with the WP plans (D9.1). We will define and collect corpora including diagnostic and evaluation sets, the former, to improve translation quality on the way, and the latter to evaluate final results.

To measure the quality of MOLTO translations, we compare them to (i) statistical and symbolic machine translation (Google, SYSTRAN); and (ii) human professional translation. We will use both automatic metrics (IQmt and BLEU; see section 1.2.8 for details) and TAUS quality criteria (Translation Automation Users Society[36]) As MOLTO is focused on information-faithful grammatically correct translation in special domains, TAUS results will probably be more important. Given MOLTO’s symbolic, grammar-based interlingual approach, scalability, portability and usability are important quality criteria. These criteria are quantified in (D9.1) and reported in the final evaluation (D9.2). In addition to the WP deliverables, there will be continuous evaluation and monitoring with internal status reports according to the schedule defined in D9.1.

Deliverables

Del. no	**Del. title**	**Nature**	**Date**
D 9.1	MOLTO test criteria, methods and schedule	R	M6
D 9.2	MOLTO evaluation and assessment report	R,M	M36

[36] http://www.translationautomation.com/best-practices/quality-evaluation-and-ta.html

WP No 10 **Leader** UGOT **Start** M1 **End** M36
WP Title Dissemination and Exploitation
Activity type MGT

Beneficiary number	**1**	**2**	**3**	**4**	**5**	**total**
Beneficiary short name	UGOT	UHEL	UPC	Ontotext	Mxw	-
Person months	23	3	3	8	0	37

Objectives

The objectives of this WP are to (i) create a MOLTO community of researchers and commercial partners; (ii) make the technology popular and easy to understand through light-weight online demos; (iii) apply the results commercially and ensure their sustainability over time through synergetic partnerships with the industry.

Description of work

Early in the project we will start by delivering a Web site uniting research, industry and users facing information about MOLTO's technology and potential (D10.2, UGOT and Ontotext). There we will feature our pre-existing work with light-weight demos, regularly updated as our work progresses, and ultimately including the use case systems. Some of these demos will be easy to integrate in third party applications like Wikis or social networks, to face larger audiences. The web site will also include a blog section with frequent informal posts on internal progress and plans and encouraging community contributions. Dissemination on conferences, symposiums and workshops will be in the areas of language technology and translation, semantic technologies, and information retrieval and will include papers, posters, exhibition booths and sponsorships (by Ontotext at web and semantic technology conferences like ISWC, WWW, SemTech), and academic/professional events such as the Information Retrieval Facility Symposium. We will also organize a set of MOLTO workshops for the expert audience, featuring invited speakers and potential users from academy and industry. Ontotext will examine the possibility of integrating MOLTO translation and retrieval technology in their intellectual property information retrieval systems. Ontotext will make the multi-lingual NL retrieval and presentation interfaces to structured knowledge as a standard feature in their semantic search products.

Deliverables

Del. no	**Del. title**	**Nature**	**Date**
D 10.1	Dissemination plan, with monitoring and assessment	R	M3
D 10.2	MOLTO Web Services, first version	P	M3
D 10.3	MOLTO Web Services, final version	P	M36
D 10.4	MOLTO Dissemination and Exploitation Report	R,Main	M36

3.3.5 Efforts for the full duration of the project

Ben.	Short name	WP1	WP2	WP3	WP4	WP5	WP6	WP7	WP8	WP9	WP10	Tot. PM
1 (CO)	UGOT	10	20	12	3	9	3	12	12	3	23	107
2	UHEL	3	12	30	12	3	3	0	6	10	3	82
3	UPC	3	4	4	0	38	24	15	3	8	3	102
4	Ontotext	3	12	10	30	0	6	15	8	6	8	98
5	Mxw	1	0	0	0	0	0	0	0	0	0	1
Total	-	20	48	56	45	50	36	42	29	27	37	390

3.3.6 List of milestones and planning of reviews

M. No.	Milestone name	WPs involved	Date	Means of verification
2*MS1	15 Languages in the Library	WP2, WP10	M6	Code and documentation available, web demonstration functional
2*MS2	Knowledge representation infrastructure	WP4	M6	Retrieval access provided to the consortium
2*MS3	Web-based translation tool available	WP3,WP10	M12	Tool accessible on MOLTO website
5*MS4	Grammar-ontology interoperability	WP4	M18	Dual way interoperability between GF grammar and ontologies in the semantic repository: grammars generated from ontology and ontology from grammars, described in D4.3.
3*MS5	First prototypes of the cascade-based combination models	WP5	M18	Translation combining grammars and and statistics is working and evaluated on a specific test set.
MS6	Grammar tool complete	WP2	M24	IDE and documentation complete.
2*MS7	First prototypes of hybrid combination models	WP5	M24	The methods are implemented and evaluated on a

				specific test set. Reported in D5.4
MS8	Translation tool complete	WP5,WP3	M30	Integrated grammar and STM available.
MS9	Case studies complete	WP6,WP7,WP8	M33	Case translators available.

Reviews will be held annually, after months 12, 24, and 36 (final review).

4 Implementation

4.1 Management structure and procedures

The management structure will provide the mechanism for the MOLTO team to reach their full synergistic potential of achievement through integrative activities, collaboration and shared expertise. It will also provide leadership and direction in science and will establish and nourish collaborative work enabling each research group to perform its tasks. With a strong, simple, and flexible management, MOLTO will be greater than the sum of its parts. The management's structure is shown in Figure 4.

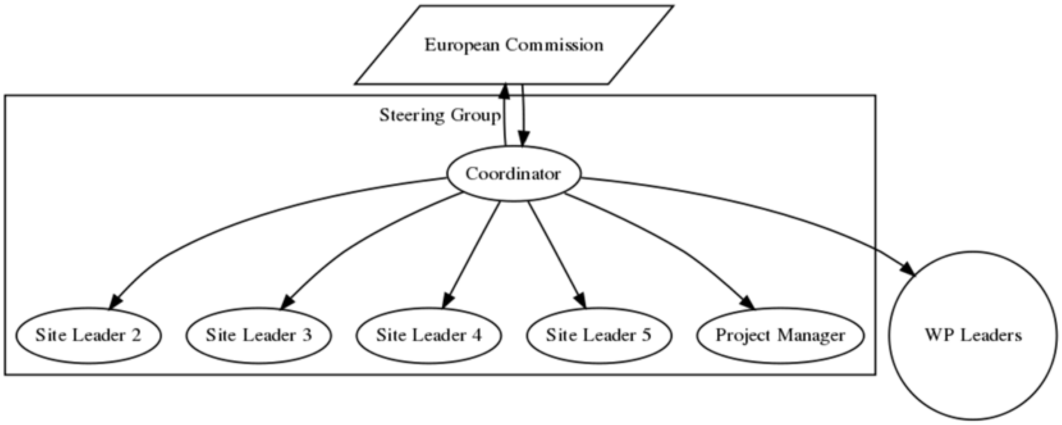


Figure 4: The management structure of MOLTO. The Coordinator is the same as Site Leader 1.

4.1.1 The Coordinator

The project leads by the Coordinator and five Site Leaders. The Coordinator is also the Site Leader of the coordinating site (Aarne Ranta, Professor of Computer Science, UGOT). The Coordinator is responsible for the functioning of the project, in particular financial, legal and administrative affairs. The Coordinator will ensure communications flow between the participants and organise together with the responsible partner (see WP1) the meetings for the project participants. The Coordinator maintains also all communications with the European Commission and is responsible for submitting all scientific and financial reports in due time. The Coordinator is thus the representative for the project towards the Commission.

The coordinator will be Prof. Aarne Ranta (UGOT), email aarne@chalmers.se. His deputy is the Project Manager, Dr. Olga Caprotti (UGOT).

The project will start on 1 March 2010.

An earlier start would be inconvenient because the end-of-year reporting period makes the administration heavily loaded in most of the participant sites, and also because of the Coordinator's teaching commitments in the beginning of 2010.

4.1.2 The Administrative Management

The coordinator will employ a part-time Project Manager, who will be in charge of the day-to-day administrative management of the project. The Project Manager will, in close connection with the Coordinator, prepare the scientific and financial reports, lead the dissemination activities, keep the Site Leaders informed about the state of the project, and take care of the public announcement of positions available within the project.

The Coordinator and the Project Manager will work in close collaboration with the Research Support Office of the University of Gothenburg (UGOT). The Office has a long track record of supporting research projects within the European Frameworks. The office will support the coordinator and the project manager from the negotiation phase to the end of the project. Its staff will prepare the consortium agreement and takes care of all other contractual matters during the duration of the project as well as the financial reports to the Commission.

The consortium agreement (CA) will be delivered at Month 2. The intellectual property rights will be defined in the CA. Staff from the Research Service Office will also inform all participants at the Kick-off meeting concerning the general conditions in an EC-project and the special conditions of this one. They will also support all the participants during the duration of the project in contractual and financial matters which need to be discussed.

4.1.3 Steering Group

The five Site Leaders together form the Steering Group, which is chaired by the Coordinator. The Project Manager will act as a secretary of the Steering Group. The Steering Group will convene in connection to the Project Meetings, which are held at six-month intervals. The first meeting will be held at Month 1 (Kick-off meeting). The Site Leaders and other key persons, as well as their deputies, will be nominated in the first Project Meeting The Steering group will also convene at need, which includes conflicts between participants, and needs for major changes in the work plan. It is expected that most decisions will be reached in consensus, but if a formal vote is required, every Site Leader has one vote in the Steering Group.

4.1.4 Work Package Leaders

Each Work Package has a Work Package Leader. The Leader is possibly, but not necessarily, the same person as one of the five Site Leaders. The Coordinator appoints the Work Package Leaders. The Work Package Leaders will also organize Work Package Meetings in connection with the project meetings.

4.1.5 Management of Gender Aspects

The lack of gender equality is prominent among the partnership, as only two of the 15 key persons are women. This lack of balance is a known problem in the research area of the MOLTO project. In order to make a change the Site Leaders will actively promote gifted female researchers in their groups to work in the MOLTO project. This is the current most important step to ensure gender equality at all levels.

4.1.6 Advisory Board

To perform independent quality assurance, the project will have an advisory board. The board consists of two eminent scientists from the areas of language technology and translation, and web technology. They are independent of MOLTO, i.e. come from outside the partner organizations and do not work in joint projects with MOLTO staff.

These scientists will participate in the annual meetings (i.e. the meetings at the end of each year) and deliver assessment reports to the Commission. In these reports, they will assess the quality of research performed in MOLTO and its relevance and usability for the community, both in the academia and outside. We will come up with a short list of names for the advisory board before the beginning of the project and name the persons before the first management report (M7). Two names already under discussion are Prof. Stephen Pulman (Oxford) and Prof. Fernando Pereira (UPenn).

4.2 Beneficiaries

4.2.1 UGOT, Goeteborgs universitet

The University of Gothenburg has approximately 50,000 students (25,000 full-time students) and 5,000 employees. It is one of the largest universities in Europe. With its eight faculties and approximately sixty departments, the University of Gothenburg is also the most wide-ranging and versatile university in Sweden. The distinctive characteristic of university education at the University of Gothenburg is the close interaction between teaching and research. Students are kept informed of the latest developments in the field they are studying and researchers gain inspiration from their students' expectations and needs. In an international perspective too, the University of Gothenburg is unusually comprehensive, with cutting-edge research in a number of dynamic research areas. Cooperation with Chalmers University of Technology, Sahlgrenska University Hospital, society at large and trade and industry has been consistently strengthened and intensified over recent years, as have international contacts and collaborative projects with partners abroad.

The MOLTO project involves three departments of the university: Computer Science and Engineering (shared with Chalmers University of Technology), Swedish Language, and Department of Philosophy, Linguistics, and Theory of Science. Groups of researchers from these departments together form the CLT (Centre for Language Technology), which is one of the eight focus areas of research of the University. The UGOT key persons of MOLTO are members of the CLT.

Aarne Ranta

Dr. Aarne Ranta is Professor of Computer Science in the Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, since 2005. His earlier positions are Associate Professor in the same Department during 1999-2005, Visiting Professor at Xerox Research Centre Europe, Grenoble, during 1997-1999, and researcher at the Academy of Finland, 1988-1998. In his Department, Ranta is in charge of the Language Technology group, which has ten members. Ranta's main research topic, since 1998, is the Grammatical Framework, GF, of which he is the main designer. His other interests are type theory, functional programming, and compiler construction. Ranta has supervised 5 PhD theses and has currently 2 PhD students. In 2008–2009, Ranta is acting as Head of Division, with budget responsibility for the Division of Computing Science with 25 employees.

Robin Cooper

Dr. Robin Cooper is Professor of Computational Linguistics at the Department of Philosophy, Linguistics, and the Theory of Science, University of Gothenburg. He is the head of GSLT (Graduate School of Language Technology[37]), a national graduate school with 50 past and present PhD students. Cooper has a B.A. (hons), Modern Languages, 1969 and M.A. (awarded 1974) from Corpus Christi College, Cambridge, and Ph.D., Linguistics, 1975, from the University of Massachusetts at Amherst. In Gothenburg since 1995, Cooper has previously worked as Assistant Professor at the University of Texas, Austin, as Associate Professor at the University of Wisconsin, Madison, and as a Lecturer at the University of Edinburgh. Cooper's main research interest is in the semantics of natural language, from both a theoretical and computational perspective.

Lars Borin

Dr. Lars Borin is professor of natural language processing in the Department of Swedish Language, University of Gothenburg, and director of Språkbanken (the Swedish Language Bank [38]), a national language resource infrastructure institution. His educational background is in languages (Slavic and Finno-Ugric linguistics), Political Science and Computer Science, followed by a PhD in Computational Linguistics. He worked at the universities in Uppsala and Stockholm before taking up his position in Gothenburg in 2002. Among his research interests are linguistic resources, in particular corpus and lexicon resources, language technology-based eScience, language technology for low-density languages and intelligent computer-assisted language learning. He has been an organizer of several conferences and workshops, the most recently of LaTeCH-SHELT&R 2009 (Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education) at EACL 2009.

Olga Caprotti

Olga Caprotti was the Network Manager of the Joining Educational Mathematics, JEM, thematic network and has been project manager of the WebALT, EDC-22253 eContent project. She has been working in technologies for the electronic communication of mathematics since joining the European OpenMath Esprit Project (1997-2000) after graduation in symbolic computation at RISC-Linz. She has substantially contributed to the latest version of the OpenMath language and is one of the editors of the standard. Her interests and competences range from semantic

[37] http://gslt.hum.gu.se

[38] http://spraakbanken.gu.se

markup and metadata for mathematical documents and interactive e-learning materials, to mathematical web services and interfaces to symbolic computation software. She is currently the secretary of the OpenMath Society and member of the W3C-Math WG and of the Committee for Electronic Information and Communication of the IMU. Previously at UHEL, Caprotti is a member of the UGOT staff in MOLTO, working as project manager and dissemination officer.

Krasimir Angelov

Krasimir Angelov is a PhD student at the Department of Computer Science, working on GF-based parsing, authoring, and translation. His thesis is planned for year 2011. He will work in connection to MOLTO with own contribution, and after the defence as a postdoc funded by MOLTO.

PhD student Y

At the beginning of MOLTO, a PhD student will be hired to work on UGOT’s tasks. We are looking for a person with a MSc in Computer Science or related subject, and experience in natural language processing, as well as functional programming and compiler construction.

PhD student or intern Z

At the mid-point of MOLTO, a PhD student or an intern will be hired to work on the Cultural Heritage task, and possibly with other similar case studies, as well as evaluation. We are looking for a person with a MSc in Computer Science or related subject, and experience in natural language processing.

4.2.2 UHEL, Helsingin yliopisto

The University of Helsinki[39], established in 1640, is the largest and most versatile university in Finland. It includes eleven faculties: Agriculture and Forestry, Arts, Behavioural Sciences, Biosciences, Law, Medicine, Pharmacy, Science, Social Sciences, Theology, and Veterinary Medicine. The university has around 38,000 students working on degrees and 7,000 employees. The Academy of Finland, which is an expert organisation in research funding and science policy, has designated 11 units of the University of Helsinki as National Centres of Excellence in Research for 2002-2007, 13 units for 2006-2011 and 12 units for 2008-2013[40]. The University of Helsinki is a member of the League of the European Research Universities (LERU).

In 2009, the University of Helsinki is pooling together the departments of Translation Studies and General Linguistics (including Language Technology). The Department of Linguistics is well known for its pioneering work in computational morphology and finite state parsing. The Department of Translation Studies has long standing experience in multilingual terminology, both in practical terminology work and in terminology management technology development projects.

Lauri Carlson

Lauri Carlson is professor of linguistic theory and translation at the Department of Translation Studies of the University of Helsinki since 1993. Carlson has coordinated or participated in a number of national and EU research projects in language technology. In 2000-2006 Carlson was invited to work as professor of language technology in the Department of General Linguistics of the University of Helsinki. Recent projects include the EU eContent project WebALT and the national LT projects Interact, 4M, CoGKS, and FinnOnto. Currently, responsible leader of the national LT project ContentFactory. Carlson’s research interests include logical semantics, dialogue, unification based parsing and machine (aided) translation. He has published two monographs and articles on semantics, dialogue and discourse analysis, and machine (aided) translation. He is the author of a constrained language parser/generator CPARSE.

Krister Linden, PhD

[39] http://www.helsinki.fi/university

[40] http://www.aka.fi

Dr. Krister Linden is an Adjunct Professor at the Department of Linguistics and a Research Project Leader for Helsinki Finite-State Technology.

Seppo Nyrkkö, PhD Student

Seppo Nyrkkö will work full-time in MOLTO as a PhD student.

4.2.3 UPC, Universitat Politecnica de Catalunya

The Technology University of Catalonia (UPC) imparts, among others, engineering degrees (Civil, Industrial, Electrical, Computer Science) and Mathematics and Statistics degrees. It is involved in research and technology transfer, with many quality doctoral programs, including: Artificial Intelligence, Applied Mathematics and Statistics. Starting September 2006, it started offering four new international masters degrees: Artificial Intelligence (with an intensification in Natural Language Processing), Applied Mathematics, Mathematical Engineering, and Statistics and Operations Research. For selected students, many double degrees in Mathematics and Engineering are possible with an additional year of studies. UPC also excels in the participation in research projects within the EC Framework programs, with extensive collaboration with other universities and private companies.

Two departments are involved: Applied Mathematics will contribute with personnel who were active in the developing of exercises and GF grammars in the scope of the WebALT project. The Natural Language Processing Research Group (GPLN) belongs to the Software Department in UPC and was founded in 1986. Ever since its creation, GPLN has worked on technologies and applications of automatic natural language processing. In the context of this project, GPLN investigates machine learning techniques for structure processing, syntactic–semantic linguistic analyzers for Statistical Machine Translation, and discriminative learning models for phrase selection in SMT. GPLN also has extensive experience in the evaluation of machine translation and has released the IQmt suite which provides a suite of MT metrics at several linguistic levels. Research activity in GPLN involves Spanish, Catalan, English, Arabic and Chinese languages. GPLN has participated in the following European projects on SMT: LC-STAR, FAME, TC-STAR-P and TC-STAR, and has taken part in the following international evaluations on MT systems: NIST (2008), IWSLT (2005-2008), WMT (2006-2009), TC-STAR (2006-2008), and MT metric evaluations: WMT (2007-2009), NIST MetricsMATR (2008).

Jordi Saludes

Dr. Jordi Saludes received his Ph.D. degree in Science (Mathematics) from the Universitat Autònoma de Barcelona (UAB) in 1991. He is associated professor of one of the departments of Applied Mathematics in UPC, lecturing in the Escola Tècnica Superior d’Enginyeries Industrial i Aeronàutica de Terrassa and the Facultat de Matemàtiques i Estadística. He has been working in Computer Vision for industrial problems in the Centre de Visió per Computador (CVC, CIRIT-CIDEM-UAB) and medical image analysis. He is currently interested in formalization and representation of mathematical content in docent applications.

Sebastià Xambó

Sebastià Xambó Descamps is Full Professor of Information and Coding Theory (since 1993) in the Department of Applied Mathematics of the Technical University of Catalonia (UPC), with teaching in the Facultad de Matemáticas y Estadística (FME). Formerly Full Professor of Algebra and Algebraic Geometry (1989-1993) at the Department of Algebra of the Universidad Complutense of Madrid. Before, he was Associate Prefessor at the Department of Algebra and Geometry of the Universidad de Barcelona (1982-1989). His research interests bearing to the MOLTO proposal are systems of mathematical computation (effective algorithms), including Web-accessible systems, and their applications, including the teaching and learning of mathematics. Expert in internet platforms for doing, teaching, learning and assessing mathematics, coordinated the development of Wiris (http://www.wiris.com/) and developed on-line, mathematically interactive materials, such as http://www.wiris.com/cc/. This experience will help in writing good interactive exercises and in assessing the validity and usability of the tools. He has been President of the Catalan Mathematical Society (1995-2002), Vicerector of Information and Documentation Systems (1998-2002), Dean of the FME (April 2003 to March 2009), and President of the Spanish Conference of Deans of Mathematics (February 2004 to November 2006).

Lluís Màrquez

Associate Professor at UPC since 2000. PhD. in Computer Science (UPC 1999; owning the UPC prize for doctoral dissertations in CS). His research focuses on Machine Learning methods for Natural Language structure prediction problems, including syntactic and semantic parsing, and statistical machine translation. He has published over 75 papers in NLP and Machine Learning journals and conferences. Usual Program Committee member of the major conferences in the area and Program Chair of CoNLL-2006, SemEval-2007, several CoNLL shared tasks, EAMT-2009, and SEW-2009. Guest editor of special issues in Computational Linguistics and Language Resources and Evaluation. Currently, he acts as president of the ACL SIG on Natural Language Learning (SIGNLL). He has participated in 4 EU funded projects and 10 Spanish government funded projects, acting as local coordinator in several of them.

Horacio Rodríguez

Dr. Horacio Rodríguez received a PhD degree in Computer Science, UPC, 1989. He is Graduate in Sciences (Physics), UB, 1977 and Industrial Engineer, UPC, 1970. He has a full time permanent position as Associate Professor at UPC (Software, LSI, department), since 1989. Previously he spent 15 years working in several Spanish companies and part time at the university. His teaching activity includes both undergraduate, at UPC, and postgraduate (at UPC, U. Alicante, U. Barcelona, U. País Vasco, U. Sevilla, IPN -México- and U. San Marcos -Perú-) studies. H. Rodriguez has lead several Catalan, Spanish, European and USA funded projects, as EuroWordNet (1996-1999), ITEM (1996-1999), CatalanWordNet (1997-1999), Aliado(2002-2005), Arabic WordNet (2005-2007) and participated in many others, as ACQUILEX (1989-1992), ACQUILEX II (1993-1995), NAMIC (1999-2001), HERMES (2001-2003), FAME(2001-2004), Text/Mess(2006-2009) among others (see http://www.lsi.upc.es/ nlp/ for details). He has advised 10 PhD theses in the area of NLP. He has a large number of publications in journals (Machine Learning, Artificial Intelligence, Terminology, Machine Translation, etc.) and international conferences (ACL, Coling, RANLP, etc.). His research interests are Natural Language Processing (both resources and tools) and Artificial Intelligence methods and tools.

Lluís Padró (Associate Professor)

Associate Professor (TU) at LSI-UPC. He is an expert on resources and software architectures for linguistic analyzers. He has also worked extensively on MT projects, especially on rule-based MT systems. His contribution will be very valuable to the corpora compilation and annotation (WP5 and WP8) and as a bridge between pure statistical and interlingua-based MT technologies.

Cristina España-Bonet

Post-doc researcher at LSI-UPC. She is a specialist on Statistical Machine Translation, specifically on the usage of Machine Learning techniques to enrich pure phrase-based SMT systems. She will program a significant part of the SMT modules in the hybrid systems and coordinate several tasks in WP5.

Xavier Carreras

Research professor (Ramon y Cajal position) at LSI-UPC. He is an specialist on machine learning techniques for natural language processing. He has developed a syntax-based statistical machine translation system, which will be the basis for advanced hybridization experiments at WP5.

David Farwell

Professor David Farwell contributes to WP5 as a member of ICREA, a third party under UPC.

4.2.4 Ontotext, Ontotext AD

Ontotext AD is a Sirma Group company focused on research and development of core technologies for knowledge representation, information extraction and retrieval and a developer of several outstanding products and major contributor to open-source platforms including KIM semantic annotation platform; wsmo4j semantic web services API and the WSMO Studio service development environment; OWLIM - the fastest and most scalable OWL engine;

GATE language engineering platform; Sesame semantic repository. The company’s competence covers ontology design, management, and alignment; knowledge representation, reasoning; information extraction (IE), applications in information retrieval (IR); Upper-level ontologies and lexical semantics; NLP and language engineering: POS-tagging, gazetteers, co-reference resolution, etc; Machine Learning: HMM, NN, CRF; Semantic Web Services. The company is a participant in a number of EC-funded projects, and as a member of W3C, involved in the development of the vision and the standards powering the development of the web.

At present Ontotext has over 35 employees and a number of scientific affiliates. Its researchers have more than 50 publications in refereed journals and international events.

Borislav Popov, head of semantic annotation and search group

Borislav Popov studied CS and specialized in Artificial Intelligence at the Sofia University, Bulgaria. His research interests include KR, ontologies, information extraction and information retrieval and have resulted in more than a dozen scientific papers. He leads multiple commercial and several European projects based on semantic technologies, and also leads the development of several products, among which the semantic annotation and search platform KIM (http://ontotext.com/kim/) and is CTO of Namerimi, developing a semantic search engine for the Bulgarian market.

Atanas Kiryakov, CEO

Atanas Kiryakov obtained his M.Sc. degree in CS from the Sofia University, Bulgaria in 1995 with a thesis on knowledge representation (KR). His research interests include KR, ontologies, lexical semantics, reasoning, information extraction, information retrieval. He is an organizer and a member of programme committees of a number of international forums; author of more than 20 publications. Kiryakov lectured courses in KR at the Sofia University, as well as at international forums. He heads the company as CEO and is leading the knowledge representation and reasoning group.

Milena Yankova, Head of NLP

Milena Yankova has a M.Sc. degree in CS (artificial intelligence program) at Sofia University, with a thesis on Information Extraction. Currently she proceeds with her PhD in Computer Science at University of Sheffield. Her research interests include identity resolution, knowledge representation and information extraction. Yankova is co-author of a number of scientific publications and member of the program committees of international scientific forums. She led the development of products in the areas of data acquisition, information extraction and identity resolution.

Mihail Konstantinov

Mihail Konstantinov will work in MOLTO as knowledge engineer.

Marin Nozhchev

Marin Nozhchev will work in MOLTO as knowledge engineer.

Georgi Georgiev

Georgi Georgiev will work in MOLTO as natural language engineer.

Boyan Kukushev

Boyan Kukushev will work in MOLTO as natural language engineer.

4.2.5 Mxw, Matrixware GmbH[41]

Matrixware Information Services offers superior solutions and services for professional Information Retrieval. These solutions and services help organizations to face the information economy and, thereby, provide them with a distinct business advantage. Matrixware's capabilities are fuelled by the findings of leading global scientists through extensive links with industrial partners and academia, building strong, trusting relationships through cutting-edge, open science, open source and open business concepts.

Neil Tipper

Mr. Tipper is a research project manager in the Science Division at Matrixware Information Service Gmbh. He has previously worked in research at the Oesterreichisches Forschungs Institut fuer Artificial Intelligence (OeFAI); Motorola Australia Research Centre; and the Information Technology Research Institute at the University of Brighton; and has published in the area of Natural Language Generation.

Dominique Maret

Dr. Maret is the Vice President and Chief Scientific Officer of the Science Division at Matrixware and has extensive experience in the management of business, technical and research aspects of technologies such as text mining, information retrieval, machine translation and natural language processing. He gained his doctorate in Applied Mathematics and Computer Science in 1987.

Andreas Tuerk

Dr. Türk is a Computational Linguist in the Science Division at Matrixware and is an expert in the area of Speech Processing and Machine Translation. He gained a PhD degree from Cambridge University on the subject of Speech Recognition. He has worked in Speech Recognition at Philips Speech Processing Vienna, Cambridge University Engineering Department, Canon Research Europe, Sail labs Technology AG and at the Telecommunications Research Center Vienna. He has published in the areas of Signal Processing, Speech Recognition and Pattern Recognition.

Robert Loibl

Robert Loibl is a Team Leader in Matrixware's Data Services division. He will be involved in data acquisition and data preparation.

Veronika Zenz

Veronika Zenz is a Researcher, Dipl. Ing (Masters) in Software Engineering & Internet Computing and research experience in the information retrieval domain. She will be involved in the evaluation task and have some involvement in the data preparation task.

4.3 Consortium as a whole

The MOLTO Consortium has four partners, of which three are academic and one industrial. The consortium was built with a great care to match the vision of MOLTO and provide the competences needed without too much overlap. The result of the process is a consortium that also covers a representative set of five different countries, diverging both geographically as in terms of language families: Fenno-Ugric, Germanic, Romance, and Slavic.

An essential question in a multilingual project like MOLTO is to find a sufficient basis of developers and testers for the different languages. Here, the consortium itself comes a long way towards the goal: its key persons alone have

[41] Matrixware left the Consortium in April 2010.

proficiency in at least ten languages. More languages are available in the immediate vicinity: in the Department of Computer Science and Engineering of the coordinating site UGOT alone, 30 nationalities are represented.

The main competences and responsibilities of each partner can be summarized as follows:

UGOT, University of Gothenburg, Coordinator. UGOT has a leading competence in multilingual grammar formalisms and grammar resources, and the group coordinates the collaborative open-source development of GF. In MOLTO, UGOT is responsible for the design and implementation of grammar development tools (WP2) and the availability of linguistic resources. UGOT also provides technical help in integrating GF with the translation tools (WP3), the Knowledge Engineering (WP4), and statistical methods (WP5). Moreover, UGOT has the leading role in the Cultural Heritage case study, where it builds on its previous competence on the domain, as well as collaboration with Gothenburg City Museum. In WP7, UGOT will develop the grammars needed in the hybrid model. As Coordinator of MOLTO, UGOT has the main responsibility for management (WP1) and dissemination (WP10).

UHEL, University of Helsinki. UHEL has competence in human translator training and translation tools, as well as in grammar development and ontologies. UHEL is therefore the main responsible partner for translator's tools (WP3) and requirements and evaluation (WP9). The group has both research and practical experience with CAT and MT tools (taught Trados tools since 1995), including involvement in national R&D projects where CAT tools have been or are developed (MLIS Lingmachine, Masterin TM/MT system, Multilingual Workbench). The group was also involved in the development of mathematical GF grammars for the WebALT project.

UPC, Universitat Politecnica de Catalunya. From UPC, two groups are involved: applied mathematics, the main party responsible in WP6, and computational linguistics, the main party responsible in WP5 and WP7. In WP5, UPC will provide the SMT technology needed for the research in this package, coordinate the corpora compilation/alignment, and develop the grammar/statistical-based combined MT models Wide experience in the construction and evaluation of Statistical Machine Translation systems, machine learning of statistical natural language parsers, and a combination of different sources of linguistic information in the construction of SMT systems. In WP6, UPC is the main responsible partner, developing grammars for natural language generation and parsing, collecting exercise samples and validation, and implementing automated mathematical reasoning. In WP7, UPC will test the hybrid model developed in WP5 on the patent corpus provided by EPO. The group has ample teaching experience in mathematics at university level in several curricula and was the main designer of mathematical GF grammars for the WebALT project. ICREA (Institucio Catalana de Recerca i Estudis Avancats Fundacio Privada) is a research organization functioning as a third party under UPC.

Ontotext, Ontotext AD. Ontotext will make the spectrum of its semantic technology and competence available to the MOLTO consortium by leading WP4 (Knowledge Engineering). They will deliver research and development of two-way grammar - ontology interoperability, infrastructure for knowledge modeling, semantic indexing and retrieval and ontology modeling and alignment of structured data sources. Ontotext will contribute to the retrieval, navigation and visualization of knowledge and ontology-grammar interoperability in the grammar development IDE (WP2) and the use case systems (WP6, WP7, WP8), developing the prototypes for two of the use cases (museums and patents). The company will heavily participate in the dissemination and exploitation (WP10) activities, on the forums for semantic technology it usually sponsors and maintain the MOLTO Web portal with live demos. The retrieval and MT outcomes of MOLTO will be integrated in the products of the company.

Mxw, Matrixware GmbH.[42] An information services company, Mxw is the leader of the patent MT and retrieval case study in WP7. Mxw specializes in analysis and retrieval of technical content in the intellectual property domain (patents, prior-art). It invests heavily in research and development in the areas of information retrieval, information extraction, large data set visualization, and recently semantic annotation and search. It is the aggregator and developer of the Alexandria patent repository for patents and patent metadata. Alexandria provides an extensible global storage facility for high-quality scientific, technical and business information, which includes a substantial collection of international patents (at the time of writing this the patent collection numbers over 70 million documents). Mxw will define the use case requirements (WP9), provide parallel corpora (WP5) and participate in the evaluation (WP9) and feasibility studies (WP10). Mxw will disseminate through their academic and industrial partnerships and appropriate events.

Here is a list of the recent previous experiences of the key persons of MOLTO in European projects:

- UGOT: TYPES (FP6-2002-IST-C), CLARIN (FP7-RI-2122230), TALK (IST-507802), DHomme (IST-2000-26280), TRINDI (LE4-8314)

[42] Mxw left the Consortium in April 2010. The tasks of Mxw have been reallocated among the remaining partners.

- UHEL: WebALT (eContent 22253)
- UPC: ACQUILEX (Esprit BRA 3030), ACQUILEX–II (Esprit BRA 7315), EuroWordNet (LE-4003), NAMIC (IST-1999-12392), MEANING (IST-2001-34460), FAME (IST-2001-28323), LC-STAR (IST-2001-32216), CHIL (IST 506909), HOPS (IST? 507967), WebALT (eContent 22253).
- Ontotext: TAO (IST-2004-026460), TripCom (IST-4-027324-STP), RASCALLI (IST-27596-2004), SEKT (IST-2003-506826)
- Mxw: ePatent, eMage

4.4 Resources to be committed

The effort table appears as a separate Annex.

Human resources

The total size of MOLTO is 390 person months, of which

- 20 (5.2%) for management
- 37 (9.6%) for dissemination and exploitation
- 348 (85.2%) for research and technology development

Partners 1–4 are roughly equal in terms of person months: UGOT 107 (97 without management costs related to coordination), UHEL 82, UPC 102, Ontotext 98. Mxw is smaller, originally 36 months but they left after 1 month. In addition to the personnel paid by MOLTO, research staff from each partner will participate in meetings, supervision, etc.

Other costs

In addition to labour, we have allocated money for travels. There are five kinds of travels:

- MOLTO’s consortium meetings
- other internal travel in MOLTO
- invited speakers to MOLTO workshops
- advisory board travels to annual reviews
- dissemination, mostly conferences where MOLTO technology is shown

The total travel budget is ca. 200 kEUR, that is, 512 EUR/PM.

Of this, the budget for consortium meetings is 100.8 kEUR (7 meetings, 12 travelling persons in each, 1200 EUR per person). 25 kEUR is for inter-site visits and coordination-related travel. 7.2 kEUR is for invited speakers to workshops, and the same amount for the trips of the advisory board (3*2 trips in each, i.e. 3 meetings à 2 persons) We allocate 60 kEUR for dissemination-related travelling, since we find this part to be essential for the MOLTO’s goal to make its technology widely known and available.

Apart from travel, some money is allocated for demonstration hardware serving the use case prototypes and live demos of MOLTO. These costs amount to ca. 30 kEUR.

The only subcontracted costs of MOLTO are the auditing costs.

5 Potential impact

5.1 Strategic impact

MOLTO is addressing the task of high-precision translation of restricted language, which in the past has not belonged to the main stream of machine translation, but which is becoming increasingly relevant due to the advent of the Semantic Web. We expect the technology created in MOLTO to help greatly in the multilingual distribution of web content and also in its usage for information access and retrieval.

MOLTO translation will be highly interoperable with Semantic Web standards (such as OWL) and adaptive to standard tools (web browsers and translators’ tools). The interoperability with Semantic Web standards will open existing ontologies and entity knowledge bases for the needs of MT tools. In turn grammar-based translation will strongly impact the way humans access structured knowledge, by providing NL query rendering to ontologies. The semantic retrieval results will also be rendered to grammatically flawless textual representations and presented to the end users as a high usability alternative to traditional table and graph based visualizations. Additionally, the grammar/ontology interoperability will empower knowledge extraction directly from text - a powerful metadata acquisition technique strongly desired by the Semantic Web, as a metadata layer struggling to capture the semantics of existing Web content.

Translators are easy to build for new domains and to extend to new languages. They can even learn to translate better "on the fly", by the use of example-based grammar writing, lexicon extension with minimal human intervention, and new statistical/grammar-based hybrid methods.

A typical MOLTO translation system will work on a well-defined domain equipped with an ontology. The MOLTO developer's tools will permit a domain expert, even without training in linguistics, to efficiently build a system that translates between an ontology and natural language. What is needed is a domain-specific lexicon and a set of example sentences describing the key properties of objects in the domain. This is made possible by the GF Resource Grammar Library (RGL) and the technique of example-based grammar writing. Porting the system into a new language is even easier, since the main relations between ontology and natural language tend to be similar in different languages; yet this similarity need not be followed, but can be overridden by transfer rules, most of which can be applied at compile time.

Once a translation system is there and integrated in a web page, a wiki, or a translator's tool, its usage is as easy as using a text editor. The predictive parser, generic for all multilingual GF grammars, helps the author in a way that is similar to a T9 system, but it gives a guarantee of grammaticality and semantic well-formedness and not only of spelling. The syntax editor makes it easier than with text editors to maintain the consistency of documents: every change is propagated to all those places that have to be changed in consequence (e.g. due to agreement).

In MOLTO, prototype systems will be built to cover 15 languages, which include at least 12 of the 23 official languages of the European Union. However, the technique is readily usable for the addition of more languages. The RGL is being developed in a collaborative project independently of MOLTO, and will in the near future cover the 23 European languages plus a number of other languages. Thus the technology will enable enriched information flows not only within the EU, but also throughout the rest of the world, opening Europe's culture and its values for the good of all.

In contrast to many other technologies within natural language processing, MOLTO is open-source and free software. It will build on open standards and enhance the interoperability between standards and components. We expect our demos and practically oriented documents to make MOLTO an attractive choice for a large population of potential users of the technology.

Expected impacts listed in the work programme

- *Automated translation that is more interoperable, more adaptive, better capable of self-learning and more user-friendly*. The project produces translation technology that is fully *automated* for its domains of application, *interoperable* with current standards and tools, *adapted* to new domains, languages, and workflows, *capable of self-learning* from minimal information given by users, and *user-friendly* in its low demands for both translation system developers and authors of new translatable content. It should be noted that these goals are achieved without compromising the quality of translation, as regards information content, grammaticality, and idiomaticity.
- *Gaps in language coverage removed, and speed and quality of translation increased*. The *language coverage* exceeds 50% of the official EU languages and is designed for painless growth. The *speed* for creating multilingual content is unforeseen, due to the full automation of translations and their updates in existing domains and languages, and to the easy adaptation to new domains and languages. The *quality* of translation is the main criterion of all MOLTO translation, which aims at reaching publishing quality in most case studies, and at least "useful" on the TAUS scale in the most experimental cases involving non-restricted language.

5.2 Plan for the use and dissemination of foreground

Dissemination is a central part of MOLTO, not the least because we see the project itself as a starting phase of something that has the potential of growing to much larger dimensions: to cover hundreds of languages, thousands of applications, and millions of users.

Early in the project we will start by delivering a Web site uniting research, industry and user facing information about MOLTO’s technology and potential. There we will feature our pre-existing work with light-weight demos, regularly updated as our work progresses, and ultimately including the use case systems. Some of these demos will be easy to integrate in third party applications like Wikis or social networks, to face larger audiences. The web site will also include a vibrant blog section with frequent informal posts on internal progress and plans and encouraging community contributions.

Dissemination on conferences, symposiums and workshops will be in the areas of language technology and translation, semantic technologies, and information retrieval and will include papers, posters, exhibition booths and sponsorships (by Ontotext at web and semantic technology conferences like ISWC, WWW, SemTech), and academic/professional events such as the Information Retrieval Facility Symposium. We will also organize a set of MOLTO workshops for the expert audience, featuring invited speakers and potential users from academy and industry.

Ontotext will make the multi-lingual NL retrieval and presentation interfaces to structured knowledge as a standard feature in their semantic search products. A major target group for dissemination is Patent Searchers/Researchers. The ability to do inline translation of segments of patents will enhance their productivity. Moreover, Ontotext, in cooperation with EPO, will endeavour to disseminate the results of the MOLTO project at academic, industrial or semi-industrial events such as, for instance, the Information Retrieval Facility Symposium, an event which brings together scientists in the field of information retrieval and intellectual property searchers/researchers, as well as the PAIR workshop. Due to the evolving nature of the industry actual appropriate events will be decided during the course of the project

Dissemination through related networks. MOLTO dissemination efforts will benefit from close cooperation with expected activities of T4ME (Technologies for the Multilingual European Information Society) Network of Excellence. T4ME language resource infrastructure will be used as a primary channel for distribution of open source tools and resources developed by MOLTO.

MOLTO will use opportunities to organize joint events, presentations, online and printed publications and other activities that will be possible in the T4ME NoE framework and within the META-Net Initiatives such as META-Share.

5.2.1 Intellectual property

MOLTO software will be released as open-source software under GNU LGPL[43].

The data sets provided by the EPO will remain under the licenses imposed by the EPO.

Ontotext contributes a stack of semantic technology, that has been developed in a period of over 8 years, involving heavy investment. The intellectual property rights of previously developed software will remain as they are (TRREE being proprietary; OWLIM, proprietary, but with a fully functional free version; ORDI[44] - open source, LGPL; SAR - open-source, LGPL). All software developed in MOLTO will be shared with the community as open-source under LGPL license.

5.2.2 References

Alshawi, H. (1992). *The Core Language Engine*. Cambridge, Ma: MIT Press.

Angelov, K. (2008). Type-Theoretical Bulgarian Grammar. In B. Nordström and A. Ranta (Eds.), *Advances in Natural Language Processing (GoTAL 2008)*, Volume 5221 of *LNCS/LNAI*, pp. 52–64. URL http://www.springerlink.com/content/978-3-540-85286-5/.

Angelov, K. (2009). Incremental Parsing with Parallel Multiple Context-Free Grammars. In *Proceedings of EACL’09, Athens*.

Bar-Hillel, Y. (1964). *Language and Information*. Reading, MA: Addison-Wesley.

Beckert, B., R. Hähnle, and P. H. Schmitt (Eds.) (2007). *Verification of Object-Oriented Software: The KeY Approach*. LNCS 4334. Springer-Verlag.

Bender, E. M. and D. Flickinger (2005). Rapid prototyping of scalable grammars: Towards modularity in extensions to a language-independent core. In *Proceedings of the 2nd International Joint Conference on Natural*

[43] http://www.gnu.org/licenses/lgpl.html

[44] http://www.ontotext.com/ordi/

Language Processing IJCNLP-05 (Posters/Demos), Jeju Island, Korea. URL http://faculty.washington.edu/ebender/papers/modules05.pdf.

Bresnan, J. (Ed.) (1982). *The Mental Representation of Grammatical Relations*. MIT Press.

Bringert, B., K. Angelov, and A. Ranta (2009). Grammatical Framework Web Service. In *Proceedings of EACL'09, Athens*.

Bringert, B., R. Cooper, P. Ljunglöf, and A. Ranta (2005, June). Multimodal dialogue system grammars. In *Proceedings of DIALOR'05, Ninth Workshop on the Semantics and Pragmatics of Dialogue*, pp. 53–60.

Brown, P. F., J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin (1990). A statistical approach to machine translation. *Computational Linguistics 16*(2), 76–85.

Burke, D. A. and K. Johannisson (2005). Translating Formal Software Specifications to Natural Language / A Grammar-Based Approach. In P. Blache and E. Stabler and J. Busquets and R. Moot (Ed.), *Logical Aspects of Computational Linguistics (LACL 2005)*, Volume 3492 of *LNCS/LNAI*, pp. 51–66. Springer. URL http://www.springerlink.com/content/?k=LNCS+3492.

Butt, M., H. Dyvik, T. H. King, H. Masuichi, and C. Rohrer (2002). The Parallel Grammar Project. In *COLING 2002, Workshop on Grammar Engineering and Evaluation*, pp. 1–7. URL http://www2.parc.com/isl/groups/nltt/pargram/buttetal-coling02.pdf.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, and J. Schroeder (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation*, pp. 136–158.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, and J. Schroeder (2008). Further (Meta-) Evaluation of Machine Translation. In *Proceedings of the ACL-2008 Workshop on Statistical Machine Translation*.

Callison-Burch, C., M. Osborne, and P. Koehn (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of EACL*.

Caprotti, O. (2006). WebALT! Deliver Mathematics Everywhere. In *Proceedings of SITE 2006. Orlando March 20-24*. URL http://webalt.math.helsinki.fi/content/e16/e301/e512/PosterDemoWebALT_e%ng.pdf.

Carlson, L., J. Saludes, and A. Strotmann (2005). Study of the state of the art in multilingual and multicultural creation of digital mathematical content. Project Deliverable D1.2, the WebALT Consortium.

Carreras, X. and M. Collins (2009, August). Non-projective parsing for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 200–209. Association for Computational Linguistics.

Chandioux, J. (1976). MÉTÉO: un système opérationnel pour la traduction automatique des bulletins météreologiques destinés au grand public. *META 21*, 127–133.

Chen, Y., A. Eisele, C. Federmann, E. Hasler, M. Jellinghaus, and S. Theison (2007, June). Multi-engine machine translation with an open-source SMT decoder. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 193–196. Association for Computational Linguistics.

Curry, H. B. (1963). Some logical aspects of grammatical structure. In R. Jakobson (Ed.), *Structure of Language and its Mathematical Aspects: Proceedings of the Twelfth Symposium in Applied Mathematics*, pp. 56–68. American Mathematical Society.

Dada, A. E. and A. Ranta (2007). Implementing an Open Source Arabic Resource Grammar in GF. In M. Mughazy (Ed.), *Perspectives on Arabic Linguistics XX*, pp. 209–232. John Benjamin's.

Damljanovic, D. and K. Bontcheva (2008). Enhanced semantic access to software artefacts. In *Workshop on Semantic Web Enabled Software Engineering (SWESE) held in conjunction with ISWC'08, Karlsruhe, Germany*.

Dean, M. and G. Schreiber (2004). OWL Web Ontology Language Reference. URL http://www.w3.org/TR/owl-ref/.

Dymetman, M., V. Lux, and A. Ranta (2000). XML and multilingual document authoring: Convergent trends. In *COLING, Saarbrücken, Germany*, pp. 243–249. URL http://www.cs.chalmers.se/~aarne/articles/coling2000.ps.gz.

Fuchs, N. E., K. Kaljurand, and T. Kuhn (2008). Attempto Controlled English for Knowledge Representation. In C. Baroglio, P. A. Bonatti, J. MaŃuszyński, M. Marchiori, A. Polleres, and S. Schaffert (Eds.), *Reasoning Web, Fourth International Summer School 2008*, Number 5224 in Lecture Notes in Computer Science, pp. 104–124. Springer.

García, M., J. Giménez, and L. Màrquez (2009). Enriching statistical translation models using a domain-independent multilingual lexical knowledge base. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing-2009*, Ciudad de Mexico, Mexico, pp. –.

Giménez, J. (2008). *Empirical Machine Translation and its Evaluation*. Ph. D. thesis, Universitat Politècnica de Catalunya.

Giménez, J. and E. Amigó (2006). IQMT: A framework for automatic machine translation evaluation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

Giménez, J. and L. Màrquez (2008). A Smorgasbord of Features for Automatic MT Evaluation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pp. 195–198.

Giménez, J. and L. Màrquez (2006). Low-cost Enrichment of Spanish WordNet with Automatically Translated Glosses: Combining General and Specialized Models. In *Proceedings of COLING-ACL*.

Hallgren, T. and A. Ranta (2000). An extensible proof text editor. In M. Parigot and A. Voronkov (Eds.), *LPAR-2000*, Volume 1955 of *LNCS/LNAI*, pp. 70–84. Springer. URL http://www.cs.chalmers.se/~aarne/articles/lpar2000.ps.gz.

Harper, R., F. Honsell, and G. Plotkin (1993). A Framework for Defining Logics. *JACM 40*(1), 143–184.

Huang, F. and K. Papineni (2007, June). Hierarchical system combination for machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, pp. 277–286. Association for Computational Linguistics.

Hähnle, R., K. Johannisson, and A. Ranta (2002). An Authoring Tool for Informal and Formal Requirements Specifications. In R.-D. Kutsche and H. Weber (Eds.), *Fundamental Approaches to Software Engineering*, Volume 2306 of *LNCS*, pp. 233–248. Springer.

Jonson, R. (2006). Generating statistical language models from interpretation grammars in dialogue system. In *Proceedings of EACL'06, Trento, Italy*.

Joshi, A. (1985). Tree-adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions. In D. Dowty, L. Karttunen, and A. Zwicky (Eds.), *Natural Language Parsing*, pp. 206–250. Cambridge University Press.

Karakos, D., J. Eisner, S. Khudanpur, and M. Dreyer (2008, June). Machine translation system combination using itg-based alignments. In *Proceedings of ACL-08: HLT, Short Papers*, Columbus, Ohio, pp. 81–84. Association for Computational Linguistics.

Khegai, J. (2006). GF Parallel Resource Grammars and Russian. In *Coling/ACL 2006*, pp. 475–482.

Khegai, J., B. Nordström, and A. Ranta (2003). Multilingual Syntax Editing in GF. In A. Gelbukh (Ed.), *Intelligent Text Processing and Computational Linguistics (CICLing-2003), Mexico City, February 2003*, Volume 2588 of *LNCS*, pp. 453–464. Springer-Verlag. URL http://www.cs.chalmers.se/~aarne/articles/mexico.ps.gz.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst (2007). Moses: Open Source Toolkit for Statistical Machine Translation. Technical report, (ACL 2007) demonstration session.

Koehn, P., F. J. Och, and D. Marcu (2003). Statistical Phrase-Based Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Larsson, S. and P. Ljunglöf (2008). A grammar formalism for specifying ISU-based dialogue systems. In B. Nordström and A. Ranta (Eds.), *Advances in Natural Language Processing (GoTAL 2008)*, Volume 5221 of *LNCS/LNAI*, pp. 303–314. URL http://www.springerlink.com/content/978-3-540-85286-5/.

Lemon, O. and X. Liu (2006). DUDE: a Dialogue and Understanding Development Environment, mapping Business Process Models to Information State Update dialogue systems. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics*.

Ljunglöf, P., G. Amores, R. Cooper, D. Hjelm, O. Lemon, P. Manchón, G. Pérez, and A. Ranta (2006). Multimodal Grammar Library. TALK. Talk and Look: Tools for Ambient Linguistic Knowledge. IST-507802. Deliverable 1.2b. URL http://www.talk-project.org/fileadmin/talk/publications_public/delivera% bles_public/TK_D1-2-2.pdf.

Macherey, W. and F. J. Och (2007, June). An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, pp. 986–995. Association for Computational Linguistics.

Martin-Löf, P. (1984). *Intuitionistic Type Theory*. Napoli: Bibliopolis.

Matusov, E., N. Ueffing, and H. Ney (2006). Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In *Proceedings of EACL*, Trento, Italy, pp. 33–40.

Mellebeek, B., A. Khasin, K. Owczarzak, J. V. Genabith, and A. Way (2006). Improving online machine translation systems. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, pp. 290–297.

Meza Moreno, M. S. and B. Bringert (2008). Interactive Multilingual Web Applications with Grammarical Framework. In B. Nordström and A. Ranta (Eds.), *Advances in Natural Language Processing (GoTAL 2008)*, Volume 5221 of *LNCS/LNAI*, pp. 336–347. URL http://www.springerlink.com/content/978-3-540-85286-5/.

Montague, R. (1974). *Formal Philosophy*. New Haven: Yale University Press. Collected papers edited by Richmond Thomason.

Och, F. J. and H. Ney (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th ACL*, pp. 295–302.

Och, F. J. and H. Ney (2004). The alignment template approach to statistical machine translation. *Computational Linguistics 30*(4), 417–449.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318.

Perera, N. and A. Ranta (2007). Dialogue System Localization with the GF Resource Grammar Library. In *SPEECHGRAM 2007: ACL Workshop on Grammar-Based Approaches to Spoken Language Processing, June 29, 2007, Prague*. URL http://www.cs.chalmers.se/~aarne/articles/perera-ranta.pdf.

Pollard, C. and I. Sag (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Ranta, A. (1994). *Type Theoretical Grammar*. Oxford University Press.

Ranta, A. (2004). Grammatical Framework: A Type-Theoretical Grammar Formalism. *The Journal of Functional Programming 14(2)*, 145–189. URL http://www.cs.chalmers.se/~aarne/articles/gf-jfp.ps.gz.

Ranta, A. (2007). Modular Grammar Engineering in GF. *Research on Language and Computation 5*, 133–158. URL http://www.cs.chalmers.se/~aarne/articles/multieng3.pdf.

Ranta, A. (2008). How predictable is Finnish morphology? an experiment on lexicon construction. In J. Nivre and M. Dahllöf and B. Megyesi (Ed.), *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, pp. 130–148. University of Uppsala. URL http://publications.uu.se/abstract.xsql?dbid=8933.

Ranta, A. (2009). Grammars as Software Libraries. In Y. Bertot, G. Huet, J.-J. Lévy, and G. Plotkin (Eds.), *From Semantics to Computer Science*. Cambridge University Press.
URL http://www.cs.chalmers.se/~aarne/articles/libraries-kahn.pdf.

Ranta, A. and K. Angelov (2009). Implementing Controlled Languages in GF. In *Proceedings of CNL-2009, Athens*, LNCS. to appear.

Ranta, A. and R. Cooper (2004). Dialogue Systems as Proof Editors. *Journal of Logic, Language and Information*.

Rayner, M., B. A. Hockey, and P. Bouillon (2006). *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. CSLI Publications.

Rosetta, M. T. (1994). *Compositional Translation*. Dordrecht: Kluwer.

Rosti, A.-V., S. Matsoukas, and R. Schwartz (2007, June). Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 312–319. Association for Computational Linguistics.

Shieber, S. M. and Y. Schabes (1990). Synchronous tree-adjoining grammars. In *COLING*, pp. 253–258.

Simard, M., N. Ueffing, P. Isabelle, and R. Kuhn (2007, June). Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 203–206. Association for Computational Linguistics.

Terumasa, E. (2007). Rule based machine translation combined with statistical post editor for Japanese to English patent translation. In *MT Summit XI Workshop on patent translation*, Copenhagen, Denmark, pp. 13–18.