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

# D9.1A Appendix to MOLTO test criteria, methods and schedule

| | |
|---|---|
| **Contract No.:** | FP7-ICT-247914 |
| **Project full title:** | MOLTO - Multilingual Online Translation |
| **Deliverable:** | D9.1A Appendix to MOLTO test criteria, methods and schedule |
| **Security (distribution level):** | Public |
| **Contractual date of delivery:** | April 2012 |
| **Actual date of delivery:** | April 2012 |
| **Type:** | Report |
| **Status & version:** | Final |
| **Author(s):** | Lauri Carlson, Inari Listenmaa, Seppo Nyrkkö et al. (UHEL) |
| **Task responsible:** | UHEL |
| **Other contributors:** | |

ABSTRACT

During the review on March 20, 2012, an appendix was requested to better specify the methodology that MOLTO intends to adopt to carry evaluation of the work and results related to each workpackage. This document tries to clarify the goals and how they will be achieved in Workpackage 9.

## Requirements of the addendum D9.1A

Requirements of the addendum:

The first year review recommendation states:

1. Taking into account the numerous endeavors undertaken in the translation domain, both research and commercial, the market segment addressed by MOLTO should be identified with maximum precision.
2. The specific case studies should also be taken into account in this effort.

The second year review recommendation adds:

1. A concrete evaluation methodology is needed focusing on MOLTO major goals: how the consortium will prove that its objects were fully/partially met (target: producers, input: predictable, coverage:limited, quality: publishing).
2. This should also include an updated description of the test criteria and used methods for each of the use cases as they are progressing, so each of the use cases can be properly evaluated at the end of the project. This also holds for the new use cases.

# MOLTO use scenarios and the market segment

The scope of applicability for MOLTO translation is a function of the domain and language coverage. The locale and grammar coverage at the start of the project was fixed by the apported GF resource grammar library. One of the main tasks of the MOLTO project is to provide tools for extending domain coverage and the associated lexical coverage by MOLTO translation users themselves. The tools should make it feasible for user communities to extend MOLTO translation to new domains and vocabularies. The market segment that can be targeted by MOLTO tools by the end of the project is in turn a function of the availability and efficiency of these tools and thereby the potential coverage of MOLTO translation. We are aiming at making it feasible to build and use domain specific grammars with lexicons in the order of thousands of words (instead of hundreds).

The two properties: restricted coverage and predictable input, restrict the market segment to production (dissemination). The constrained language property means MOLTO will not offer a replacement for CAT, i.e. translation tools that help human translators with complex third party authored documents which they are not allowed to modify. But MOLTO translation can be added an additional facility in the CAT toolkit. Conversely, traditional TMS facilities may add value to the application and extension of MOLTO methods. These ideas are explored in WP 3.

MOLTO remains at the core a tool for constrained language multilingual generation. Its potential strengths are 1) multiple simultaneous target languages and 2) reliable enough quality for blind translation (translation from a known language to unknown languages). 2) can only be obtained if the quality is higher than human translation. In practice, some

level of human revision is probably going to be needed, but the need can be significantly less than in current workflows.

From this, we conclude that the most promising market segment for MOLTO translation is *constrained language content localization*. In current translation industry, there is a more or less clear split between interface localization, which involves translation of fixed short strings from a list of interface messages by professional or volunteer translators, and content translation, which is mostly done outside of the website using CAT tools.

MOLTO targets an as yet less explored and little exploited niche between them, viz. multilingual content localization of constrained language content. Typical use cases are a webstore inventory, a museum guide, rule generated correspondence, or formulaic parts of a more complex document type (say descriptions of chemical formulas in a patent). Here the content is already regulated and predictable.There are further such scenarios beyond those included in MOLTO use cases, typically involving some database generated information (e.g. product descriptions, user guides, chemical manufacturer's data sheets, job tickets, medical reports). In some such scenarios. real time blind translation to multiple languages would be a major selling point.

In the MOLTO translation scenario related to this market segment, there is a close interaction between some database/ontology and a human/ruleset that generates the text to translate, and the translation process itself. The content to translate can co-evolve with the grammar by which it is translatable. Such use cases will be tested in the MOLTO semantic wiki platform.

If or as the vision of Linked Data becomes reality, there is bound to be a growing demand for natural language verbalization of the web of linked data ontologies. The Web of Data is supposed to become an additional layer of the web that is tightly interwoven with the classic document Web and has many of the same properties:

- The Web of Data is generic and can contain any type of data.
- Anyone can publish data to the Web of Data.
- The Web of Data is open, meaning that applications do not have to be implemented against a fixed set of data sources, but can discover new data sources at run-time by following RDF links.

In particular,

- Data publishers are not constrained in choice of vocabularies with which to represent data.
- Data is self-describing. If an application consuming Linked Data encounters data described with an unfamiliar vocabulary, the application can dereference the URIs that identify vocabulary terms in order to find

their definition.

The growing linked data cloud can create a growing market segment for a matching linked cloud of multilingual MOLTO ontology verbalizers. Ontotext's GF based natural language query interface into Ontotext linked data is a first application of MOLTO resources in this direction. As the review points out, a generalization of the ad hoc ontology/GF mappings the KRI and museum cases gets a high priority here.

## A. The multilingual semantic wiki scenario

### A. The multilingual semantic wiki scenario

The new workpackage 11 aims to use GF to extend AceWiki to a multilingual constrained language semantic wiki. Like the original AceWiki, it allows users to express in natural language logical constraints that are subject to automated reasoning. AceWiki already has facilities for extending the lexicon. A subset of the constraints expressible in ACE are interatranslatable with the OWL ontology language. In the scenario envisaged here, the multilingual semantic wiki works as a tool for extending a special domain ontology through natural language verbalization. This platform supports the scenario where a special domain ontology and its verbalizations are extended simultaneously.

In one natural scenario, a special domain expert expresses the constraints in unconstrained natural language as comments in the wiki. One or more ontology experts refine the description into a set of simpler statements in a constrained subset that maps to OWL, using already existing ontologies as base and creating the missing ontology resources and their verbalizations in a common natural language using the lexicon editor. The domain experts can test the conceptualization by asking questions of the ontology. The questions are answered in natural language using the wiki's reasoners. When the coverage of the ontology and its verbalization in the chosen language/s is sufficient, the lexicon is extended for the remaining languages, using existing term ontologies as a base, by target language experts.

## B. The MOLTO CAT scenario

### B. The MOLTO CAT scenario

More traditional translation projects can also contain parts which can be handled with constrained language translation. The MOLTO patents case has shown that certain sections of patent text, in particular complex chemical compound descriptions, are not well covered by SMT. The MOLTO translator tools workpackage looks into ways of embedding MOLTO constrained language translation as one tool in the toolkit of a more traditional CAT platform. In this use case, we also test the ability of a translation community (company) to collaboratively extend coverage of the fragment handled with MOLTO tools. This sort of a hybrid SMT+MOLTO+CAT workflow is tested with the patents use case in the MOLTO Translators tools platform as described in D 3.1. Note that the two scenarios are not exclusive.

In an overarching scenario, a domain translation is developed in the first scenario and it is applied in production translation in the MOLTO CAT scenario. Actors in some of the supporting roles of the MOLTO CAT scenario may use the wiki tool in their work.

The CAT scenario is described in more detail below under WP 3.

# Relating the scenarios to the MOLTO use cases

# Relating the scenarios to the MOLTO use cases

The following details the MOLTO use cases relating them to the scenarios above. Each section lists the evaluation criteria, measures and methods applied in the use case.

## WP2 - Grammar developers tools

WP2: GRAMMAR DEVELOPERS TOOLS

The grammar developers tools promise to enable quick development of a new domain and language. This promise is best tested directly by measuring the time and expertise taken

1. to create or extend an ontology to a new domain using MOLTO tools
2. to generate an abstract grammar for a domain from an ontology for it
3. to create or extend the concrete grammar for a language to a new domain
4. to extend the vocabulary for a language to cover a new or extended domain

The measures are taken for a system with a coverage in the order of a) 100 concepts b) 1000 concepts. The platforms used in carrying out the tests include the multilingual semantic wiki (tasks 1 and 2), the TermFactory? platform (tasks 1 and 4) and the grammar editing tools (tasks 2,3,4). To test these claims, we need to fix one or more domains to create/extend. We haven't got a great many domains to choose from yet. We would do well to extend in the direction of known 'good' ontologies.

1. phrasebook < travel ontologies, e.g. http://sites.google.com/site/ontotravelguides/Home/ontologies
2. museum/painting < music ontology, http://musicontology.com/
3. patents/pharmaceutics < chemical safety data sheets + chemical ontologies e.g. CheBI? , ChemINF?

Baseline evaluation figures prior to the use of MOLTO tools for a domain of smaller size were obtained in the phrasebook exercise reported in Ranta et al.2010 [9]. For comparability, the same criteria and measures are to be applied in subsequent evaluations.

## WP3 - MOLTO CAT tools

WP3: MOLTO CAT TOOLS

The MOLTO CAT scenario is designed to serve a translation community that carries out translation projects using MOLTO tools as an additional CAT tool. The translation community members are assigned different roles. What they may do depends on the role. Roles are assigned in the translation management system. In the MOLTO demonstration system, the TMS is Globalsight. The TMS manages the resources of a project. The resources include

- documents
- grammars

- translation memories
- term collections

A MOLTO CAT translation project is composed by a collection of resources and a community of actors playing different roles in the project. One actor can bear more than one role.

The roles include

- project manager (rights to manage the resources and the workflow)
- editor (source competence in domain, domain expert, authority to edit the source)
- translator (bilingual competence in domain, not necessarily domain expert)
- revisor (target language competence in domain, domain expert)
- ontologist (competence to extend the domain ontology)
- terminologist (bilingual or target competence in the domain)
- grammarian (competence to extend domain GF grammar)

The TMS manages the project workflow, that is, routes documents through different steps between the actors. The actions include

- project manager:
    - create users
    - assign roles to users
    - create a translation project
    - prepare resources for a translation project
    - plan the workflow
    - assign actors to actions

- editor
    - split source to constrained/unconstrained sections
    - indicate allowed/authorize new deviations from constrained language source

- translator:
    - translate unconstrained sections using CAT tools (including SMT proposals from translation memory)
    - translate constrained language sections using MOLTO
    - propose term for lexical gap

        - create grammar extension request

- ontologist
    - find or create missing concept
    - create grammar extension request
    - create terminology extension request

- terminologist
    - find or define equivalents to a new concept

- grammarian a revisor
    - carry out grammar extension

The typical envisaged workflow is this. A translator in a multilingual translation project works on a structured multipart document, some of whose parts are marked as amenable to translation with the MOLTO editor. The rest is translated with traditional CAT tools. A subsection appropriate for MOLTO translation is opened in the MOLTO translation editor. The appropriate GF grammar and terminology are specified in the project resources. If the section is properly within the fragment covered by the grammar, the section should parse and translate correctly without translator intervention. This is the default if the MOLTO marked section has been created in scenario A. However, until the domain grammar has been fully tested for blind translation in all target languages, a target language translator or revisor must check that the target text is correct.

If the grammar coverage is not complete, the translation editor shows some parts of the section marked as untranslatable.

In the easy case, the coverage problem can be fixed by a conservative paraphrase or, if the translator's brief permits pre-editing, by a more creative rewrite of the section source to bring it under the coverage of the MOLTO grammar. The original source and its paraphrase get stored in the translation memory as an instance of source rewrite, and will be available for other translators as a model solution of the coverage problem. If a rewrite is not possible, the next move depends on the workflow.

1. If the translator's brief is just to produce a complete translation to the target language in a bilingual project, the translator just translates the part not covered by MOLTO using traditional CAT tools. The out-of-coverage segment gets marked as a manually translated MOLTO section segment in the translation memory. Such segments can be collected and sent off as non-coverage tickets to the project's terminology and grammar management.

1. The task may be to extend MOLTO translation to a language whose coverage in the given domain is not complete.
    a. In the case of a simple out-of-vocabulary term or concept belonging to a category known to the grammar, the MOLTO equivalents editor can be used to extend the concrete and/or abstract vocabulary of the grammar. If a concept with a matching GF category and verbalizations is found in an existing MOLTO term ontology, the missing term can be added into the translation project's GF grammar extension

module so as to become immediately available to further MOLTO translation in the project and subsequently included in the project ontology.
   b. If a candidate term is found using some non-authoritative lexical source, the candidate term gets added as a term candidate to the relevant domain for community approval. That is, the translation unit containing the proposed candidate concept/term in its abstract/concrete grammar context is saved in translation memory and sent to the terminology management platform for terminology checking and approval.

1. The task may be to develop a master text or pilot translation, in preparation for a subsequent multilingual translation project (pre-editing). A gap in the MOLTO coverage can arise when the special domain section subject to MOLTO translation has not been authored in the semantic wiki, but for instance generated from a database or merged from text from more than one subdomain. In this case, more effort is worth spending to extend the coverage of the MOLTO grammar to the source before proceeding to multilingual translation.
   a. In the case of out of vocabulary terms or concepts, the grammar can be extended through the translation editor as above.
   b. In more complex cases needing grammar extension, the translator just creates a model translation and submits it back to the ontology/grammar editing workflow. The model translation is saved in translation memory and can be used in regression testing against the edited grammar.

**The MOLTO translation editor**

As indicated in the MOLTO CAT system design, the MOLTO translation editor is integrated as a plugin to the translation management system alongside more traditional CAT editors. The MOLTO CAT scenario sets the following requirements on the editor and its integration to the TMS.

- editor
    - the MOLTO translation editor parser can out from the source parts it can translate and indicate what it lacks for parts that do not translate.
    - the GF back end is able to include proposed extensions into the grammar.

The development of the translation editor to satisfy these requirements is taken over by UGOT, as it is closely coupled to the ongoing development of the GF robust parsing and grammar extension services.

- integration
    - the TMS environment is able to extract from structured source text parts which are subjected to MOLTO translation.
    - the editor has access to a term/ontology manager to propose terms/concepts to fill the indicated gaps and submit new

proposals for approval

These requirements remain the responsibility of UHEL.

### TERM ONTOLOGY MANAGEMENT WITH TERMFACTORY?

The TermFactory? term management specification and query/editing API is a Tomcat Axis2 webservice API for querying, editing, and storing small RDF/OWL ontologies representing concepts and multilingual expressions/terms associated with the concepts. TermFactory? contains a term ontology schema that follows professional terminology standards, but the tools can also be used to edit any RDF/OWL ontologies through an XHTML representation RDF. The XHTML representation is extremely configurable. It can be parametrized for the presentation layout (concept oriented, lemma oriented), filtered for content, and even localized with another TF term ontology so that names of properties and classes shown to the user are chosen from the localization ontology. The term ontology editor is a pluggable javascript editor that is offered as a standalone Tomcat servlet as well as a MediaWiki? extension. A simpler tabular editor exists for the common task of adding different language equivalents to an existing ontology term.

TermFactory? is to be integrated with the MOLTO KRI over the JMS transport interface provided in the KRI. Besides the Ontotext repositories, TermFactory? also talks to Jena RDB and triple set repositories. TermFactory? user management is planned to happen through the GlobalSight? API.

### WP3 EVALUATION

The GlobalSight? translation management system forms a platform to test the MOLTO TT scenario that combines traditional CAT tools with the MOLTO translation editor. The best dataset for testing the full MOLTO CAT scenario should be the patents, since it already uses hybrid methods and generates a translation of less than 100% coverage. To have a complete use case of the mixed scenario, a pure GF grammar for chemical compounds could be applied to translate chemical compound definitions in the patent text.

The MOLTO CAT review workflow will be used manage translation quality evaluation of the multilingual translations produced in the other use cases. This exercise in itself also serves to test the usability of MOLTO scenario B.

## WP4 - Ontology-grammar interoperability

### WP4: ONTOLOGY-GRAMMAR INTEROPERABILITY

The second year review considered Deliverable 4.2 and Deliverable 4.3 insufficient and they were not approved by the reviewers in their current status. The objectives of WP4 are, as stated in the DoW? :

(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) modeling and alignment of structured data sources; (iv)

semantic indexing and retrieval, (iii) modeling and alignment of structured data sources, (iv) alignment of ontologies with the grammar derived models.

D4.2 should contain a report on the Data Models, Alignment Methodology, Tools and Documentation. More specifically, it should contain information about the aligned semantic models and instance bases. While D4.2. contains information about Reason-able views and the key principles constituting these views are stated in the document, it does not state how these key principles have been implemented in the MOLTO-project. D4.2 does not comply with the key principle stating "Clean up, post-process and enrich the datasets if necessary, and do this in a clearly documented and automated manner." D4.2 should contain exactly all details about the automation process of multiple ontologies. so that this knowledge and technique can be re-used to integrate new ontologies with the existing ones.

D4.3. should clear out the issue of the two-way interoperability between ontologies and GF grammars. This is still unclear, although objective (i) of WP4 is clear that this is a research-intensive part of MOLTO. Based on the WP4 presentation given in the review, this process requires the manual writing of mapping rules (NL Query -> GF, GF-> SPARQL query), which means limited potential for further re-use. The partners must clear the degree of automation that can be performed. What is required for porting this to a new application? Concrete steps should be provided making clear what can be automated and what cannot with the provided infrastructure. Details about mapping rule induction etc. should be provided.

As for the ontology/grammar mappings, here is what we have concretely got so far:

- Ontotext has defined one instance of single ontology triple to GF translation in WP 4.3.
- Aarne et al. have defined a more complex property tuple to text translation for the Museum case.

The examples show that the owl to GF mapping need not be difficult in any given case. What seems open is how to generalize these examples for the general case of generating a mapping for a new domain. In particular, we want a solution that allows the reuse of ontology to GF mappings to create more complex grammars from existing parts. The modularity of both OWL and GF suggest ways of approaching this goal.

One approach to a more general solution is to use the term ontologies developed in TermFactory? to also store parts of mappings needed for GF verbalization. In a TermFactory? term ontology, a term is a pair of a general language expression and a special language concept. In this approach, an ontology concept would map to an abstract grammar term. Individual language expressions and terms associated with the concept map to concrete grammar terms. A term or expression would inherit GF grammar properties from classes to which it belongs (say, exp:Noun). Grammatical properties common to all uses of a given general language expression would be stored as properties of the expression. GF terms or grammatical properties that are specific to a domain GF grammar would stored as properties of a domain specific term.

Instead of having to define a new grammar and create concept to grammar associations from scratch, a grammar would be compiled from appropriate choices of resource from the term

ontology plus a language and/or domain specific syntactic base. To extend a vocabulary, we add a new term (expression, concept) instance, typed in the appropriate categories, and add to it any further GF properties that are relevant to its correct linearization. The concrete expression associated to a compositional abstract grammar term need not be specified in the ontology, if it can be compositionally derived from the GF abstract syntax associated to the concept and other resources in the ontology. The above does not claim to do more than propose a way to decompose the ontology to grammar mapping into reusable parts.

If the approach seems useful, UHEL is prepared to invest effort to building a test case using the museum case as a starting point.

## WP5 - Statistical Machine Translation

WP5: SMT

The research goal was to develop translation methods that complement the grammar-based methods of WP3 to extend their coverage in unconstrained text translation. Specifically, WP 5 promised to 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.

### WP5 evaluation

WP5 is has its own internal evaluation complementing that of WP9. Since statistical methods need fast and frequent evaluations, most of the evaluation within the package is automatic. The WP7 case study on translating Patents text is the use scenario to test the techniques developed in this package. Ultimately, Ontotext will examine the feasibility of the prototype as a part of a commercial patent retrieval system (D7.3).

### Corpora

Statistical methods are linked to patents data. This is the quasiopen domain where the hybridization is going to be tested. The languages of the corpus are English, German, and French, the official languages of the European Patent Office (EPO).

Besides the large training corpus, we need at least two smaller data sets, one for development purposes and another one for testing. The order of magnitude of these sets is usually around 1,000 aligned segments or sentences. For this, we have used a subset of MAREC patents (http://www.ir-facility.org/prototypes/marec), and a collection of 66 patents provided by the EPO. The concrete figures are explained in WP5 and summarised in the table below.

Seg DE-EN Seg FR-EN Seg FR-DE dev MAREC 993 993 993 test MAREC 1,008 1,008 1,008 test EPO 847 858 831

### Metrics

BLEU [3] is the de facto metric used in most machine translation evaluation. We plan to use it together with other lexical metrics such as WER or NIST in the development process of the statistical and hybrid systems. Lexical metrics have the advantage of being language-independent, since most of them are based on n-gram matching. However, they are not able to catch all the aspects of a language and they have been shown not to always correlate well with human judgments. So, whenever it is possible, it is a good practice to include syntactic and/or semantic metrics as well. The Asiya package provides tools for (S)MT translation quality evaluation. For a few languages, it provides metrics to do this deep analysis. At the moment, the package supports English and Spanish, but other languages are planed to be included soon. We will use Asiya for our evaluation on the supported language pairs.

**Manual evaluation**

Final translations will be also manually evaluated. This is the most realiable way to quantify the quality of a translation since automatic metrics cannot capture all the aspects that a human evaluator takes into account as said in the previous section.

We now propose to follow the ranking for evaluation that is used in patent offices such as EPO. It can be applied to sentences but also to full patents. So, automatic metrics will also be adpated to deal with full patent evaluation and see how they correlate. This way we will be able to perform a deep study.

Quality level: Ranking for human evaluation

- Accurate + consistence IPC vocabulary

The translation is understandable and actionable, with all critical information accurately transferred. Most of the text is well written using a language consistent with patent literature.

- Fluent - consistence IPC vocabulary

The translation is understandable and actionable, with all most critical information accurately transferred. Some text is well written using a language consistent with patent literature.

- Actionable

The translation is not entirely understandable and actionable, with some critical information accurately transferred. The text is of the text is well written using a language consistent with patent literature.

- May be actionable

Possibly understandable and actionable (given enough context and/or time to work it out), with some information stylistically or grammatically odd, but the language may still reflect a sound content to a patent professional. Most of the text written using a language consistent with patent literature.

- Not useful

Absolutely not comprehensible and/or little or no information is transferred accurately.

# WP6 - Math

WP6: MATH

The math use case remains as it was, except that the use case may assume that premises requiring encyclopedic knowledge needed to frame word problems are given. Assuming that the math scenario will be embedded in the semantic wiki, the background premises may be given by the author of the problem in the facts database where the problems are formulated.

The mathematics use cases involve a problem author, a student and a teacher. The usability of the scenario is tested with realistic subjects playing each of these roles and the evaluation collected with a questionnaire and/or a journal. In addition, we should try estimate the savings from the system when scaled up to a larger use base and variety of languages, since these are the novelties in the MOLTO solution.

### Evaluation for WP6

#### Diagnostic and progress evaluation for translation quality

WP 6 has developed a treebank based method for doing regression testing on the translations produced by the math grammar. A treebank entry consists of:

- An abstract tree for the gf grammar
- For each language (encoded as ISO 3 letter code), one or more Changesets.

A Changeset has:

- source: The person submitting it;
- revision: An integer equal to the svn revison in which this item is commited
- concrete: The proposed linearization
- and optionally a comment.

A defect is a difference between the actual linearization of an entry and the sample in the last changeset.

The procdure is as follows.

1. Using the gr command, create a list of abstract trees
2. Refine this list by removing or modifying unnatural productions (too deep, too long, too meaningless);
3. Add linearizations for all targeted languages: This makes the initial changeset;
4. Send the pairs (abstract tree, L linearization) to a fluent speaker of language L and ask for corrections;

5. Add the corrections to the treebank as new changesets.
6. Generate a list of defects and tackle them
7. Generate new linearizations, and go to step 4. Cycle until satisfied or out of resources.

See http://www.molto-project.eu/wiki/living-deliverables/d61-simple-drill-grammar-library/5-testing for further discussion.

#### Use case and usability evaluation

## WP7 - Patents

WP7: PATENTS

The first year review recommended that WP7 work should focus on the major issues examined in MOLTO, especially in relation to the grammar-ontology interoperability rather than chemical compound splitting. Specific scenarios are needed for the exploitation of MOLTO tools in this case study. It was recommended to include such scenarios in a new version of deliverable D9.1.

In response, two use case scenarios were described: UC-71 and UC-72.

- UC-71 focuses on grammar-ontology interoperability. User queries, written in CNL (controlled natural language) are used to query the information retrieval system.
- UC-72 focuses on high-quality machine translation of patent documents. It uses an SMT baseline system to translate a big dataset and fill up the retrieval databases. In order to study the impact of hybrid systems in translation quality, a smaller dataset will be translated using the hybrid system developed in WP5.

### Evaluation related to WP7

WP7 corresponds to the Patents Case Study. Its objective is to build a multilingual patents retrieval prototype. The prototype consists of three main modules: the multilingual retrieval system, the patents translation and the user interface. This document proposes a methodology to evaluate these modules within the MOLTO framework.

#### Translation system

The automatic translations included in the retrieval database have been produced by the machine translation systems developed within the WP5. Hence, the evaluation related to this module is the same as the one described for the WP5 systems.

#### Retrieval system

Nowadays, the IR-facility organizes the TREC Chemical IR Evaluation campaign (http://www.ir-facility.org/trec-chem-2011-cfp) The evaluation campaign has three different

tracks. One of them is very related to our objective in this WP. - Technology Survey - Given an information need (from the bio-chemistry domain) expressed in natural language, retrieve all patents and scientific articles which may satisfy this need.

Following the guidelines described in the TREC campaign, the methodology proposed to evaluate the patents retrieval system is as follows.

1. Select a set of topics (between 5-10) and create a natural language queries for each topic (preferably, they must be manually created by experts). Each query must express an information needed based on the data described in a patent. The priority is to be as similar as possible to a genuine information need of an expert searcher.
2. The system will have to return a set of documents that answer this information need as best as possible. For any of the runs, it may return a maximum of 100 relevant documents (our database will contain ~8000 documents), preferably using the standard trec_eval format: Topic_number query_number document_id rank score run_name.
3. Manually annotate the retrieved documents as match/mismatch.
4. Calculate the AP (Average Precision, [1]) and NDCG (Normalized Discounted Cumulative Gain [2]), which are common metrics for these kind of systems [4].

**The user interface**

User interfaces are usually evaluated by means of their Usability. According to the ISO 9241-11, usability must measure the "Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.".

Hence, to get a complete picture of the usability, we need to measure the user satisfaction (users reaction to the interface), effectiveness (can people complete their tasks?) and efficiency (how long do people take?).

The three measures of usability are effectiveness, efficiency and satisfaction. They are independent and it must be measured all three to get a rounded measure of usability.

1. Effectiveness. This can be automatically by logging the user interactions with the system, and manually analysing the system responses. The measure can be also contrasted with a specific question in the satisfaction questionnaire.
2. Efficiency. This measure can be automatically obtained by logging the user interaction. To do so, the experiment requires to implement the needed mechanisms to a) determine the start and end of the experiment (for each scenario and/or for the complete experiment), b) relate the previous record with a specific user and the other two measures (effectiveness and satisfaction). We could also request the users to time themselves, but this measurements will be less reliable.
3. Satisfaction. This measure can be obtained through requesting the users to answer a questionnaire. Commonly used questionnaires for this tasks are the IBM CSUQ [5] or the SUS Questionnaire [6]. Another novel method is the cloud of words, in which users have to select a subset of words describing the system among a predefined set of

adjectives. An general description of this method can be found in [7].

The experiment setting may consist of two scenarios: a closed one (i.e., specifying the information that must be obtained) and an open one (i.e., let the user search any type of information). The users are requested to complete both scenarios, and the order in which they are done must be balanced (i.e., Half of them will do the open scenario first). They must answer the questionnaire twice, just after each scenario.

The potential users might be of two types: MOLTO participants and related people (internal) and external users. The internal users can be used as the control test. External participants can be engaged from tools like the Mechanical Turk Requester [8].

## WP8 - Museum

WP8: MUSEUM

D8.2 (AR, DD, RE 2012) -->

The museum grammar creates multilingual descriptions from a museum ontology using GF grammar for the verbalization. The GF grammar provides a direct verbalization of the triples and different types of complex discourse patterns: a text generated by the grammar has necessary elements painting, painting type and painter, and as optional information year, museum, colour, size and material. For a detailed description, see D8.2 (Ranta et al. 2012).

An abstract syntax for the direct verbalization grammar can be generated automatically from the ontology. The discourse patterns have been human-generated, and they can be reused for different language versions and for more objects. For example, the type of a complete painting is described in an abstract syntax as following:

```
cat CompletePainting Painting PaintingType Painter OptYear
OptMuseum OptColour OptSize OptMaterial ;
```

CompletePainting is a type constructor that takes type parameters to construct a type for a painting. A painting from Gothenburg City Museum has a following type:

```
data GSM940042ObjPainting : CompletePainting GSM940042Obj
MiniaturePortrait JKFViertel (MkYear (YInt 1814)) (MkMuseum
GoteborgsCityMuseum) (MkColour Grey) (MkSize (SIntInt 349 776))
(MkMaterial Wood) ;
```

In the concrete syntax all this complexity is hidden. Porting the grammar to a new language requires only writing the concrete syntax. However, the underlying ontology makes sure that the grammar generates only valid descriptions and not random combinations of paintings, painters and other properties.

As of March 2012, the translation of the museum objects and the additional lexicon (painting materials, colours) needs to be done manually. The future plan is to combine tools developed

in WP3 to make the lexicon extension automatic, by using multilingual lexicon harvesting from term ontologies or other reliable sources (DBPedia, TermFactory? ).

**Evaluation for WP8**

D8.2 has promised to increase the coverage from 5 languages to 10 languages, and extend the grammar and the lexicon for at least 5 languages. The GF grammar can be tested continuously, while developing, with the treebank method described earlier in this document. A grammar developer should be fluent in the language she is developing the concrete syntax, and the treebank testing should be thorough. If the testing is done properly in the grammar development phase, there shouldn't be need to have specific translation quality evaluation experiments. The best way to spot problems is through real usage, so UHEL is offering a bug tracking platform, where users can report all kinds of issues, including language errors.

The idea is not to translate existing texts, but to generate descriptions in response to user queries. As described in D8.2,

D8.2: *The grammar presented here allows to generate well-formed multilingual natural language descriptions about museum artefacts with the aim of empowering users who wish to access cultural heritage information through different computing devices.*

Other question is to evaluate the use of the queries. Currently the grammar has one discourse pattern with optional elements; the variety comes from adding or leaving out some information. One possibility discussed in D8.2 is to include more variety in the generated text. A qualitative evaluation study with non-expert human subjects would serve this purpose. The aspects to test in this experiment would be the ease of querying and whether the results answer the query. However, as long as this plan is not certain, we are not designing any concrete test methods.

A third question is the ease of the grammar writing and the reusability of the grammar -- is it possible for other museums to use the grammar, if they have their own standards? Currently a prerequisite for the museum grammar is an ontology that follows Cidoc-CRM standard. This is an important aspect, if we are to make MOLTO tools used outside the test cases within the project. The step from a specified format to verbalizations are well defined, now it should be given more thought how to cover the first step of the process: whichever type of museum database to a CRM format. We could, as a part of evaluation, interview some domain specialists and survey the needs and interests for this kind of system, and whether the first step is a big enough threshold to prevent them to use the system.

## WP11, WP12 - Multilingual semantic wiki and beInformed

WP 11 MULTILINGUAL SEMANTIC WIKI

The main goal of the proposed work-package is to build an engine for a multilingual semantic wiki, where the involved languages are precisely defined (controlled) subsets of the 15 languages that are studied in the MOLTO project.

The wiki engine would allow the input and presentation of the wiki content in all the languages, and perform formal logic based reasoning on the content in order to enable e.g. natural language based question answering. The users of the wiki can contribute to the wiki in any of the supported languages by adding statements to the wiki, as well as extending its concept lexicon. The wiki would integrate a "predictive editor" that helps the user cope with the restricted syntax of the input languages, so that explicit learning of the syntactic restrictions is not required. Ideally, the wiki would also integrate semantics-support, e.g. a paraphraser and a consistency-checker that could be used to enhance the quality of the wiki articles. The wiki engine is going to be implemented by combining the resources and technologies developed in the MOLTO project (GF grammar library, tools for translation and smart text input) with the resources and technologies developed in the Attempto project (Attempto Controlled English, AceWiki? ).

The task of WP11 will be to combine the technologies developed in the MOLTO project with ACE and AceWiki? , concretely:

- porting the ACE grammar from English to the 15 MOLTO languages. The work in this task will be supported by the other MOLTO work-packages who are involved in developing GF-based grammars;
- extending AceWiki? to allow input in multiple different languages, i.e. develop AceWiki? into a multilingual controlled language wiki. This task includes work on modularizing AceWiki? and integrating existing GF tools for translation and smart text input;
- using existing ACE application domains and test cases to evaluate the new multilingual wiki-system.

**WP 11 evaluation**

In this document, the list of application domains to evaluate multilingual semantic wiki becomes longer, since we envisage using the multilingual wiki as a common testbed for those MOLTO use cases where an ontology and its verbalization are developed in parallel. This can include some or all of the following cases:

- Creation/extension of a domain and its verbalization (WP 2)
- Developing mathematics exercises (WP 6)
- The museum guide browser (WP 3)
- beInformed scenario

WP12: BEINFORMED

It is too early to describe evaluation of this case in detai pending a description of the use case itself. But we can suggest that the beInformed use case could be framed and tested as an instance of the multilingual semantic wiki scenario, if the business logic reasoning rules can be expressed in the semantic wiki database.

# References

# References

[1] AP. E. M. Voorhees and D. K. Harman, editors. TREC: Experiment and Evaluation in Information Retrieval. MIT Press, 2005.

[2] NDCG. K. Kärvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst., 20(4):422--446, 2002.

[3] BLEU. Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation" in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311--318

[4] IR Metrics. http://en.wikipedia.org/wiki/Information_retrieval#Mean_average_precision

[5] IBM CSUQ. http://hcibib.org/perlman/question.cgi?form=CSUQ

[6] SUS. http://www.usabilitynet.org/trump/methods/satisfaction.htm

[7] Word Cloud. Usability. http://www.userfocus.co.uk/articles/satisfaction.html

[8] Mechanical Turk Requester. https://requester.mturk.com/

[9] Ranta, Aarne, Enache Ramona, and Détrez Grégoire, Controlled Language for Everyday Use: the MOLTO Phrasebook. Controlled Natural Languages Workshop (CNL 2010) http://www.molto-project.eu/sites/default/files/everyday.pdf

---