MOLTO

D1.3A Advisory Board Report

http://www.molto-project.eu

Contract No.:	FP7-ICT-247914
Project full title:	MOLTO - Multilingual Online Translation
Deliverable:	D1.3A. Advisory Board Report
Security (distribution level):	Confidential
Contractual date of delivery:	M13
Actual date of delivery:	22 March 2011
Type:	Report
Status & version:	Final
Author(s):	Keith Hall and Stephen Pullman
Task responsible:	UGOT
Other contributors:	All

ABSTRACT

Advisory board report for the first year of the MOLTO project lifetime, 1 Mar 2010 - 28 Feb 2011.

The goals of the MOLTO project can be summarized as an attempt to make small scale, rule-based, grammatical analysis usable for some practical purposes, with the aim that this will provide greater accuracy and precision than current wide-coverage, usually statistically driven, approaches. The domains chosen for case studies are: web based mathematical exercises (in a sublanguage specifically aimed at mathematics); search and translation of patents in the biomedical and pharmaceutical domains; descriptions of exhibits in museums, and a tourist phrasebook running on smart phones.

Progress for the first year of the MOLTO project has been very strong, despite an unforeseeable dropping out of a consortium member. Each of the partner institutions has made considerable progress on their respective work-packages.

We would like to highlight the most notable aspects of the progress. In particular, the level of outreach and dissemination for MOLTO and GF is very good. There have been a number of demos at top-tier conferences of the MOLTO phrase-book and the GF summer school planned for summer 2011 brings both awareness of the project and creates practitioners who are ready to work with GF, something that the MOLTO project is predicated upon. This is both beneficial to the members of the MOLTO consortium, but also to the EU funding body; the more awareness there is of the project, the more obvious the impact is on the community.

With regards to the individual projects for the work-packages, the loss of Matrixware and the choice to continue the Patent-Search component of the project seems to be somewhat unresolved. We understand that the delay was the result of the suggestion of the EU to incorporate the EPO as a replacement partner: while absorbing a lot of time, these negotiations were eventually unsuccessful. Given that there is still a component of the project which focuses on using GF to allow for multilingual patent-search, there needs to be more said about exactly what will be done. Ontotext has agreed to take over this work package: the proposed solution to use GF to allow you to create more concrete queries for patent search seems reasonable. We would strongly encourage the development of some convincing use cases here: there are already many ways of searching patents that would be difficult to improve on directly, but there may be a type of search or a type of user for whom this method solves a problem. We would make a similar suggestion for the components of other work packages that involve interaction between NL queries and a large ontology: again, some convincing use cases for how this could be actually deployed in practice need to be developed. Perhaps one direction to explore might be help for developers who have to maintain or change an existing ontology. Linking suggested changes in the ontology to examples involving the relevant concepts and relations in the associated text might be a useful tool to check that the any changes or additions to the ontology are justified by the way they are visualized by people familiar with the domain.

Adding to the general outreach noted above, the progress on a new interface for GF grammar developers is impressive. The current web-based demo helps one understand how a concrete grammar (language dependent grammar) can be constructed from other grammars, allowing the user to alter the grammar rather than writing it from scratch. Again, this is really important for the entire MOLTO project as each of the use cases assumes the development of resource grammars for the abstract grammar's concepts.

The demos of the current systems are very helpful for understanding the progress at the end of the first year (phrasebook, latex-translation). The ontology interface (KRI) is very impressive. It's unfortunate that there is not a resource language, which could generate SPARQL queries. It

would help to explain why this is infeasible. The work on SMT using GF for alignments is also very interesting. Since the black art of mixing alignment techniques for phrase-extraction, etc. is one of the more questionable components of standard SMT systems, it might be nice to understand how a low-coverage solution could help so much. What happens when the domain of translation is neither the domain captured by the GF syntax, nor the domain of the parallel training data? The current work is a great first step in understanding how we can improve MT with deep syntax.

It will be important to focus on evaluation techniques that do not bias against the principles of the MOLTO project. Currently popular benchmarks for technologies like parsing or machine translation assume that all systems are aiming at wide coverage, whereas this is not the case for MOLTO, which aims rather to maximize precision over recall. Perhaps the museum collection application would provide a convincing demonstration along these lines. If it proves possible to write descriptions of exhibits in one language, and then have the other 14 or 15 language translations come out immediately with much higher quality than could be obtained by current machine translation methods, that would be a truly impressive achievement,

In summary, the MOLTO project seems to be right on-track, fulfilling the promises and deliverables of the project and doing an excellent job of outreach.