Molto project meeting, 7th March 2012, Zurich

Over the past year, the MOLTO project has continued to make progress on many of the work-packages. The consortium has successfully recovered from the unexpected loss of an industry partner during the first year and has a new industry partner with a strong use-case for MOLTO technology (more details below).

Some highlights related to the education impact of GF, the fomalism behind the MOLTO project include: 1) the GF book has been published in 2011, 2) the 2011 GF summer school was held in August, and 3) continued development on the grammar building tools.

WP 2 Grammar developers' tools:
Continued development on the cloud-based GF editing environment is progressing, although at present it seems there is a small user base and consequently little feedback on its utility. These tools appear to be best for helping developers and maintainers of GF grammars for systems which are already in use. The motivation for a cloud-based solution may need some strengthening, but certainly lowers the bar for entry for new developers of GF grammars. We would suggest that measures are taken to increase its use and utility, perhaps by harnessing effort outside the project, for example by encouraging students to use it.
The Eclipse plugin provides another mechanism to help GF grammar developers. While the workflow is very "software-engineering" focused, there are some additional tools provided that allow for evaluating a grammar under development on some gold-standard trees. This appears to be focused on the large-scale grammar developer - something that will be necessary for MOLTO if it's to be adopted by industrial partners.

WP3 - Translators' tools, lexicon extraction
There has been quite a bit of work on the translator tools as part of WP3, however, the integration of these tools into the grammar development tools is incomplete. A few of the highlights here include the integration of translation memories and TermBank into the framework (it's not clear if this is underway or to be done). Integrating these tools into the Molto framework is critical, but only a small part of the work that is going on. There is continued work on
automatically extracting concepts and terms from raw text using distributional techniques (e.g., bio-chemical compound names from the patent corpus). This work is one of the many approaches to increase the coverage of the MOLTO approach for real-world applications. As this matures, the project will need to be able to track progress (e.g., precision and recall of systems that extract concepts and terms) With any such approach there is an issue over how to validate the results for consistency and usefulness, which the team are aware of but for which there is as yet no concrete proposal for an evaluation set and an evaluation metric.

WP4 (and WP7) - Interface to Knowledge-base/ontology
Continued development on the interface between controlled natural language and information retrieval using GF is being done at OntoText. As we saw last year, this work continues to be very promising. In the previous year we saw natural language being automatically transformed into SPARQL queries in order to retrieve information from a knowledge-base. This year, this

has been extended to include the transformation from RDF triples (the format for raw data in the knowledge-base) into natural language. Now, both the input and output of the knowledge-base are transformed to and from natural language - this is great progress. The one shortcoming here is that the coverage for input queries is very small and the quality of the output is poor (there appears to be an impoverished concept inventory created from the knowledge-base). In order to both evaluate and direct further research and development in this area, it is important that a standard evaluation set is created as well as an evaluation metric. Up to now, progress has been at a very high level (adding components). From here on out, the actual quality of these components will require a more fine-grained evaluation. Another problem, given the lack of coverage, is conveying a model of that coverage to a user. There is perhaps already one partial solution to this latter problem via the Attempto Controlled English input device brought to the project by Zurich University, one of the new partners.

WP5 - Integration of SMT and RBMT
MOLTO translation is based on hand-built grammars which often lack coverage and is somewhat brittle to extend beyond its primary use-case (e.g., it captures a controlled natural language). The work on comparing and integrating SMT and GF-based translation has continued nicely. It's clear that the coverage of the GF is low and the team can now quantify exactly how low that is on the patent claims dataset. On the integration side, they have experimented with multiple approaches to integrate GF into an SMT framework. They have shown that they can achieve small increases in accuracy (under a variety of automatic metrics) using a model where the GF translation contributes to the SMT predictions. There is a clear path for continued research in this area. In the domain of patents it is worth pointing out that Google now have an online patent translation system with the collaboration of the European Patent Office, who were originally intended to be a partner in the Molto project. This should give the WP5 team another point of reference.

For WP6 (Mathematics) we saw a GF based natural language interface to SAGE, a computer algebra system. Given that there is a use case for a controlled natural language interface for mathematical expressions, the work in this area is making progress. There is now CNL input and output for the expressions embedded in SAGE. The requirements of this component and the research questions that this work addresses are unclear. It would be good to reevaluate this component and determine whether continued development on this WP is necessary. There is again here the issue of finding a suitable evaluation measure and conveying to a user what can and cannot be asked. A natural comparison to be made here is with Wolfram Alpha, which for English at least, overlaps in its functionality with the Molto system.

WP8 is another case study on Cultural Heritage, specifically access to museum collections. Last year, we heard the general pitch for how MOLTO will be used to create a common cataloging interface for current museum databases. Up to this point, they have achieve a mapping between the different schema used by various museums - resulting in a common concept inventory and the mapping from those concepts to the database realizations. It appears that the one difficulty in making this multilingual is that it requires that the "description" field be

translated. Unfortunately, this is not controlled natural language and will require full-scale translation. It might be a good time to revisit the goals of this WP and make sure that there is a clear path to progress including a description of expected accomplishments for the next year.

WP 11 and 12 are based around the contributions brought by Zurich, a new partner, which are quite mature pieces of software already. We would have found it helpful to have seen a clear plan for how these contributions will relate to the existing activities, since there is scope for many enhancements, but also the danger of some redundancy. For example, The University of Zurich has been developing their own transformation from controlled natural language to a logical formalism. In order to provide multilingual support, they are investigating the transformation from their formalism to GF. (e.g., creating an abstract grammar that will directly generate their logical expressions). It's very nice to see the integration of complementary techniques. If this is successful, this would be a great example of how MOLTO can provide a framework for content producers.

beInformed
We also heard a presentation from the other new partner,`Be Informed', which contained a clear and interesting wish list of what they would hope to gain from participation in the Molto project. These included components corresponding to most of the aspects of restricted domain natural language processing that the GF approach should excel at, and we thought that, as such, the addition of this partner was particularly apt, and could provide a convincing demonstration of the practical utility of some of the technology being developed in the Molto project. We would encourage the project to focus their efforts to ensure that this testing and - if successful - transfer of technology is given adequate resources within the remainder of the timetable. Thought also needs to be given to support for industrial users of GF/MOLTO after the end of the project.