MOLTO WPs 3 and 9
UHEL

Lauri Carlson

University of Helsinki
Department of Modern Languages
lauri.carlson@helsinki.fi
MOLTO WPs 3 and 9
Timetable

|   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  | 21  | 22  | 23  | 24  | 25  | 26  | 27  | 28  | 29  | 30  | 31  | 32  | 33  | 34  | 35  | 36  |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|   | WP1: Management |
|   | WP2: Grammar Developer’s Tools |
|   | WP3: Translator’s Tools |
|   | WP4: Knowledge Engineering |
|   | WP5: Statistical and Robust Translation |
|   | WP6: Case Study: Mathematics |
|   | WP7: Case Study: Patents |
|   | WP8: Case Study: Cultural Heritage |
|   | WP9: User Requirements and Evaluation |
|   | WP10: Dissemination and Exploitation |
MOLTO WPs 3 and 9

Tasks and roles

- MOLTO translation editor: Krasimir Angelov UGOT
- MOLTO vocabulary editor: Junyou Shen UHEL
- GF runtime in C: Lauri Alanko UHEL
- TermFactory multilingual term mgmt (Lauri Carlson)
- Term harvesting from Web: Inari Listenmaa UHEL; Adam Slaski UGOT
- Ontology / Vocabulary extraction: Seppo Nyrkkö UHEL
- Human evaluation: Maarit Koponen UHEL / Langnet
- SMT evaluation: Cristina España UPC
MOLTO WP 3
Translation tools

Lauri Carlson

University of Helsinki
Department of Modern Languages
lauri.carlson@helsinki.fi
WP 3 deliverables

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  RP  M30
WP 3 Translation scenarios

• Multilingual website maintenance
  – Google TT, WikiBhasha, DeskShare: Site Translator
• Museum and Math cases
• Patent case

Same: constrained language, professional translation

• Different:
  – size of the domain
  – authority over the source
WP 3 Translation scenario

- Author/translator/s work/s on web content online
- Have or create source content
- Validate source against domain ontology / GF grammar
- Fill gaps in ontology / grammar coverage
- Translate the text with GF
- Fill gaps in target language coverage
- Get translations revised (loop back to previous step)
- Publish translations on website
- Publish ontology / lingware in community
MOLTO WP 3

WP 3 Translation tools API

- API
  - runtime translation engine/s
  - translation editing tools
  - lingware maintenance tools

- Web API
  - mashups
  - plugins

- reuse existing communities
- support collaboration
WP 3 TT API

- Parts needed:
  - MOLTO translation editor
    - edit and translate constrained language online (UGOT)
  - MOLTO translation editor term tool
    - add vocabulary while editing (UHEL)
  - GF runtime engine/s
    - multilingual translation (UGOT, UHEL)
  - GF on-the-fly lingware compiler
    - use vocabulary added in editing (UGOT)
  - Ontology-to-GF bridge (UGOT, ONTOTEXT)
WP 3 TT API

- TermFactory
  - Manage terms in Semantic Web
    - Query
    - Edit
    - Index
      - store content in repositories
  - Collaborative terminology work
    - Discuss
    - Browse
    - Edit
  - Offline tools for ontology work
WP 3 work done

- Tools API
- MOLTO translation editor (Krasimir Angelov)
- MOLTO vocabulary editor (Junyou Shen)
- TermFactory back end (Lauri Carlson)

- Content production
- Term harvesting from Web of Data (Inari Listenmaa)
- Ontology / vocabulary learning (Seppo Nyrkkö)
- WordNet en-fi
WP 3 TT API

- **C language GF runtime**
    - Minibar generation demo
    - This returns the abstract parse tree in JSON format. Other operations are also supported.
WP 3 work to do

- Tools API (complete)
- Integrate vocabulary editor to MOLTO translation editor
- Integrate vocabulary editor to TermFactory
- Integrate TermFactory with KRI

- Content production
- WordNet en-fi to KRI
- Populate TF from KRI
- Set up MOLTO TF community → evaluation WP
MOLTO WP 9
User reqs and evaluation

Lauri Carlson

University of Helsinki
Department of Modern Languages
lauri.carlson@helsinki.fi
Deliverables

- D 9.1 MOLTO test criteria, methods and schedule
- D 9.2 MOLTO evaluation and assessment report

- D 9.1 due date was too early (given the delays)
  Need to do another round with case studies data
WP 9 work done

- Human evaluation (Maarit Koponen / Langnet)
- PhD thesis project
  - Methods for evaluating fidelity (<>fluency)
  - MT semantic error analysis
  - Translator evaluation of output
- Machine evaluation (Cristina España-Bonet)
  - Asiya
WP 9 work to do

- Get results to evaluate from use cases
- Use MOLTO web API to evaluate use cases
  - Math WP
  - Museum WP
- SMT workpackage does its own internal evaluation
- USE MOLTO TF community to quality manage
  - ontologies
  - vocabularies
Web of words

- There is a lot of words on the web -
- But are they any good for MOLTO?
  - Special domain vs. general language
  - Domain vs. language skills
  - Consistency
  - Quality
Wikipedia and Wiktionary

- Collaborative encyclopedia / dictionary work
- Quality managed by peers
- Human to human

- Strengths and weaknesses as in Wikipedia

Web of Data

- Machine processed web content
  - Databases
  - SPARQL endpoints
  - Ontologies

  http://wordnetweb.princeton.edu/perl/webwn?s=entity
  http://api.talis.com/stores/wordnet/items?query=entity&max=50
  http://factforge.net/search?q=entity&search=
  http://www.mpi-inf.mpg.de/yago-naga/yago/
Harvest the Web

• Web of Data
  – Inari Listenmaa: term queries from FactForge

• Web of Text
  – Adam Slaski: term extraction from Wikipedia

• Match
  – Seppo Nyrkkö: text based ontology matching
TermFactory

TermFactory aim:

- bridge between
  - professional multilingual terminology work
  - collaborative content creation
  - web of data
- using Semantic Web tools
- reusing existing collab platforms
TermFactory

- Traditional term work
  - Source collection
  - Preterm collection
  - Concept selection
  - Concept analysis
  - Term description
  - Compilation
  - Publishing
  
  Terminologist + expert committees

- TF tools & workflow
  - Community
  - Term statistics
  - Voting
  - Discuss / Edit
  - Term ontologies
  - Query language
  - Transformations

  TF + expert community
GF, TF and KRI

GF

<--

KRI

<--

TF

<-->

- GF gets quality approved content directly from KRI. The content is compiled into GF grammars. (This is invisible to end user.)

- GF term editor gets term suggestions from TF. TF can use KRI as a TF repository. (Both quality controlled and working content can be served here.)

- Raw content is harvested to TF from various sources. Suggestions are cached in TF format and quality checked in TF. TF can use KRI as a TF repository.

- TF quality approved content is saved in KRI to be used by GF directly. (Loop closes).
TermFactory ...

Publications

FinnWordNet - a Finnish WordNet database
Linden, K., Carlson, L. & Niemi, J. A. 12.2010

FinnWordNet - WordNet på finska via översättning

Utility Evaluation of Tools for Collaborative Development and Maintenance of Ontologies
Publications

TermFactory: A Platform for Collaborative Ontology-based Terminology Work

Proceedings of the XIV Euralex International Congress (Leeuwarden, 6–10 July 2010)

TermFactory: Collaborative Editing of Term Ontologies


Dublin 2010
MOLTO WPs 3 and 9

Publications

Assessing Machine Translation Quality with Error Analysis
Maarit Koponen
MikaEL : Electronic proceedings of the KäTu symposium on translation and interpreting studies
Volume 4 2010