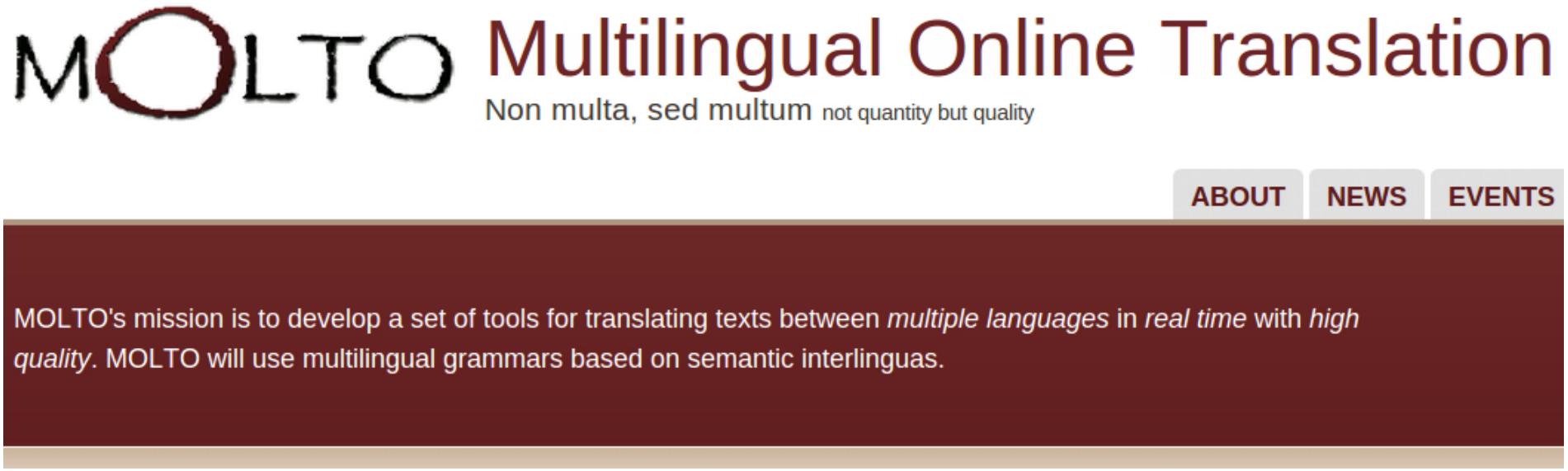
MOLTO Second Year Review

Aarne Ranta

Barcelona, 20 March 2012



FP7-ICT-247914, Strep, `www.molto-project.eu`

U Gothenburg, U Helsinki, UPC Barcelona, Ontotext (Sofia),

(since January 2012) U Zurich, Be Informed (Apeldoorn)

March 2010 - May 2013

EC contribution 2,975,000 EUR

What's different?

Tool	**Google, Bing, Babelfish**	**MOLTO**
target	consumers	producers
input	unpredictable	predictable
coverage	unlimited	limited
quality	browsing	publishing

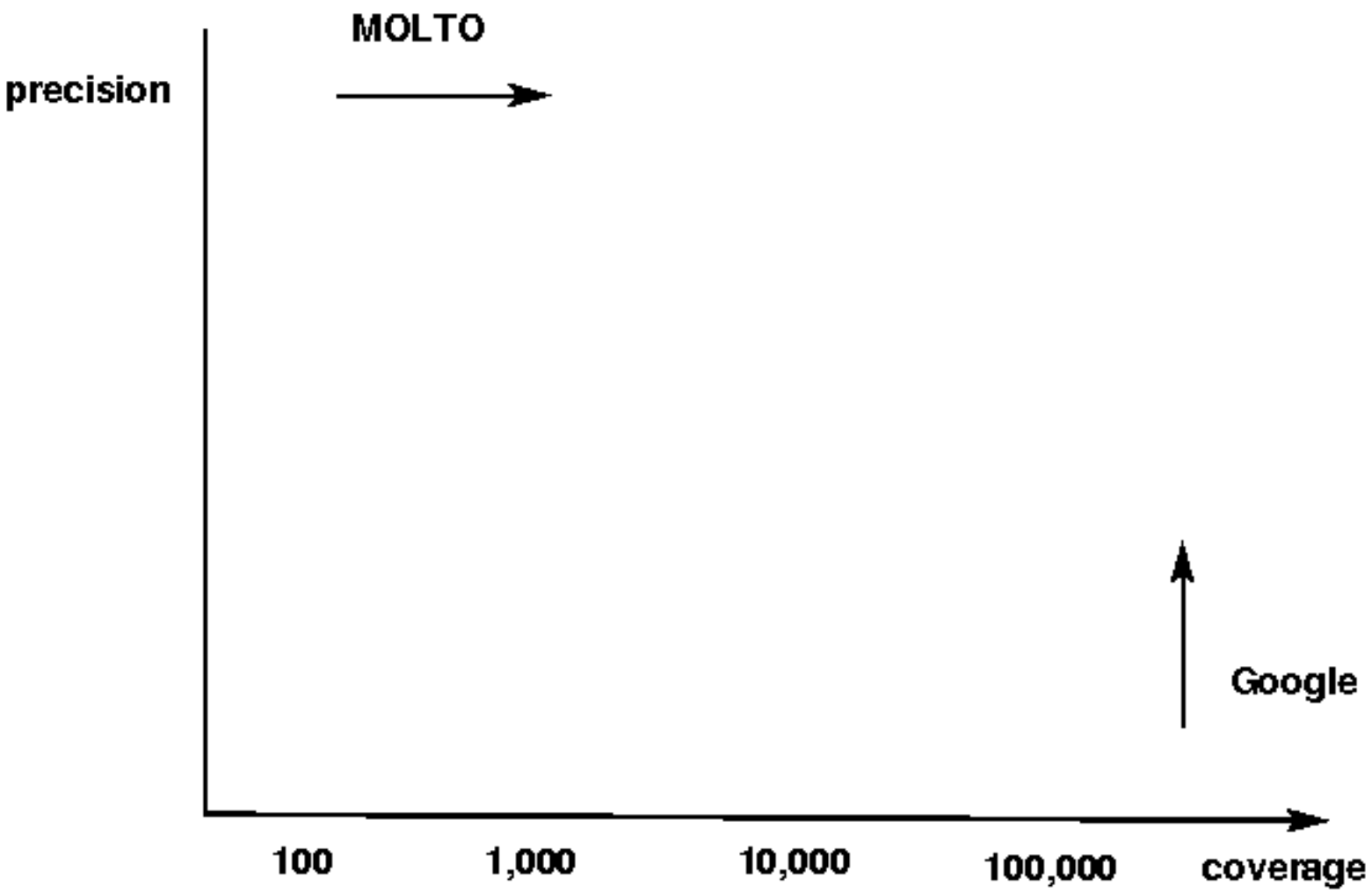
Aspects of reliability

Linguistic knowledge

Predictability (vs. randomness)

Programmability (vs. holism)

Coverage/precision trade-off: we cannot deal with millions of concepts


MOLTO
precision
Google
100
1,000
10,000
100,000
coverage

Main technologies

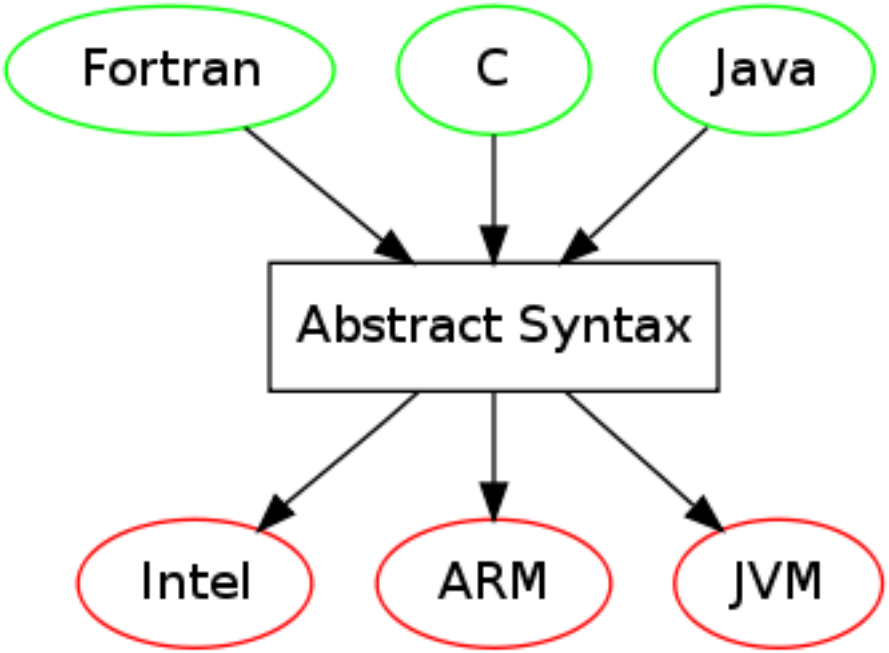
GF = Grammatical Framework: multilingual grammars

OWL Ontologies

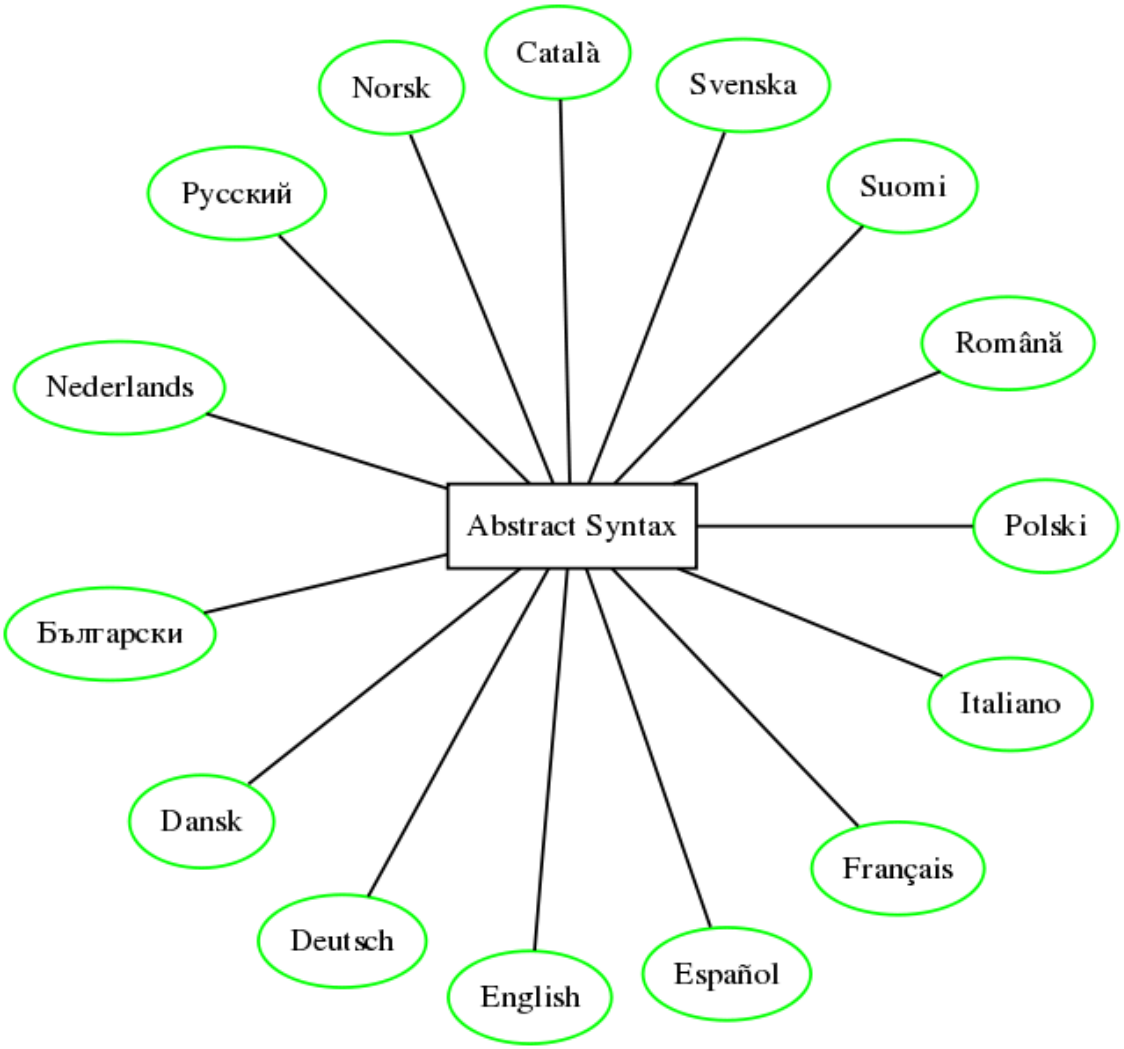
Statistical Machine Translation

Controlled language technology

The GF model: multi-source multi-target compilers



MOLTO languages

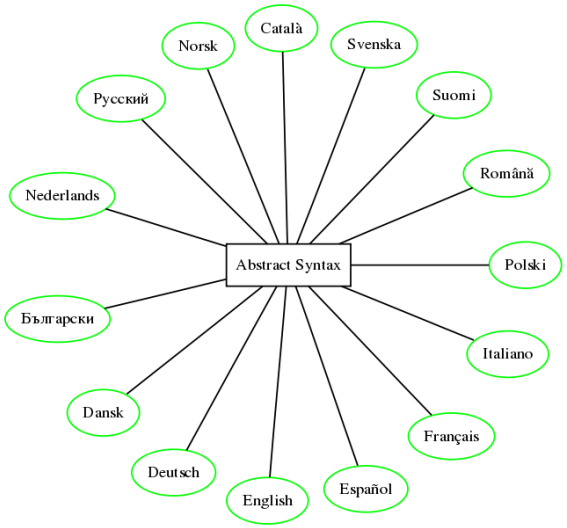


The multilingual document

Master document: semantic representation (abstract syntax)

Updates: from any language that has a concrete syntax

Rendering: to all languages that have a concrete syntax



Two things we do better than before

No universal interlingua:

- a framework for domain-specific interlinguas: **type theory**

Yes universal concrete syntax:

- a general-purpose **Resource Grammar Library**
- no hand-crafted *ad hoc* grammars

Controlled language

Almost what MOLTO is, except that we

- generalize this to **multilingual controlled language systems**
- support ambiguous language (and **disambiguation**)

Prime example: Attempto Controlled English (U Zurich)

- generalized to 5 languages in GF (CNL 2009)
- extended to 15 in MOLTO

Summary of work packages

WP1: management (UGOT)

Two new members: UZH, Be Informed

WP2: grammar tools (UGOT)

IDE's: Eclipse (John Camilleri) and cloud-based (Thomas Hallgren)

Support for on-the-fly extension

Resource grammars: Hindi, Latvian, Nepali, Persian, Punjabi, Sindhi, Thai (Shafqat Virk & al., Normunds Gruzitis)

Issues

Move WP end to Month 27.

WP3: translator's tools (UHEL)

Terminology tools (Lauri Carlson, Inari Listenmaa, Seppo Nyrkkö)

Translator user interface (Lauri Carlson, Inari Listenmaa)

Fast large-scale parsing: a C runtime for GF (Lauri Alanko, Krasimir Angelov)

Issues

Integration of tools

WP4: knowledge engineering (Ontotext)

Natural language queries (Milen Chechev, Borislav Popov)

OWL ontology verbalization (Milen Chechev)

WP5: statistical and robust translation (UPC)

Hybrid architecture with soft/hard integration (Cristina España, Lluís Màrquez, Ramona Enache, Aarne Ranta)

Robust parsing in GF (Krasimir Angelov)

Issues

Access to data (solved)

WP6: case study: mathematics (UPC)

Grammar and lexicon for the OpenMath standard, 12 languages (Jordi Saludes et al.)

- great interest in CADE community (automated theorem proving)

A dialogue system for computer algebra Sage (Jordi Saludes)

Issues

Evaluation; use cases; the role of Ontotext.

WP7: case study: patents (UPC)

SMT baseline + GF improvements for English to French (Cristina España, Lluís Màrquez, Ramona Enache, Adam Slaski)

Natural-language information retrieval from patents (Meritxell Gonzalez, Milen Chechev)

Issues

Access to data (solved); grammar for chemical compounds; expert evaluation.

WP8: case study: cultural heritage (UGOT)

Data collection: CRM ontology, Gothenburg City Museum database (Dana Dannélls, Milen Chechev)

Prototype with natural language generation for English, Finnish, French, Italian, Swedish (Dana Dannélls, Ramona Enache, Aarne Ranta)

Issues

Human resources for this WP: we suggest extending it to Month 36.

WP9: user requirements and evaluation (UHEL)

Hybrid evaluation with Asiya (Cristina España, Lluís Màrquez)

Software testing methods (QuickCheck) applied to grammars (Ramona Enache, Koen Claessen)

Issues

Proper metrics for each of the case studies.

WP10: dissemination and exploitation (UGOT)

GF book, CSLI publications, 2011 (Aarne Ranta)

GF Tutorial at CADE, Wrocław, August 2011

2nd GF Summer School, Barcelona, August 2011

FreeRBMT12 in Gothenburg, 13-15 June (submission deadline 7 April)

GF Tutorial at ICFP, Copenhagen, September 2012

WP11: multilingual semantic wiki (UZH)

The ultimate user interface

- multilingual collaborative document editing (the MOLTO vision)
- combine translation and grammar extension
- reasoning based on abstract syntax

ACE-Wiki ported to GF (Kaarel Kaljurand, Tobias Kuhn, Norbert Fuchs)

WP12: interactive knowledge-based systems (BI)

Multilingua questionnaires and decision making

- user input + reasoning
- explanations generated in the users' languages

A new category of grammarians: software engineers with minimal GF training

Conclusion

Increased weight on controlled language scenarios.

But also: patent case with large coverage, increasing precision.

Work plan changes approved in WP2, WP7.

Work plan changes suggested in WP6, WP8.