# MOLTO: Machine Translation to Rely On?

Aarne Ranta

Nordic Seminar on Language Technology and Accessibility, SLTC,
Linköping, 27 October 2010


Multilingual Online Translation, FP7-ICT-247914

`www.molto-project.eu`

## Project summary

MOLTO's goal is to develop a set of tools for translating texts between multiple languages in real time with high quality. Languages are separate modules in the tool and can be varied; prototypes covering a majority of the EU's 23 official languages will be built.

...

## Who?

UGOT: University of Gothenburg, Sweden (coordinator)

UHEL: University of Helsinki, Finland

UPC: Universitat Politècnica de Catalunya, Barcelona, Spain

Ontotext: Ontotext AD, Sofia, Bulgaria

Mxw: Matrixware GmbH, Vienna, Austria

## How much?

Total: 3,000,000 EUR, EC contribution 2,375,000 EUR

90% for work (390 person months)

86% for RTD, 10% dissemination, 5% management

1 March 2010 – 28 February 2013

# The Call

ICT-2009.2.2: Language-Based Interaction

Majority of EU languages

Improvement by an order of magnitude

Use of existing linguistic resources

## What's new?

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

## Producer's quality

Cannot afford translating

- *prix 99 euros*

to

- *pris 99 kronor*

## Producer's quality

Cannot afford translating

- *I miss her*

to

- *je m'ennuie d'elle*

  ("I'm bored of her")

## The translation directions

Statistical methods (e.g. Google translate) work decently *to* English

- rigid word order
- simple morphology
- originates in projects funded by U.S. defence

Grammar-based methods work equally well for different languages

- Finnish cases
- German word order

# Main technologies

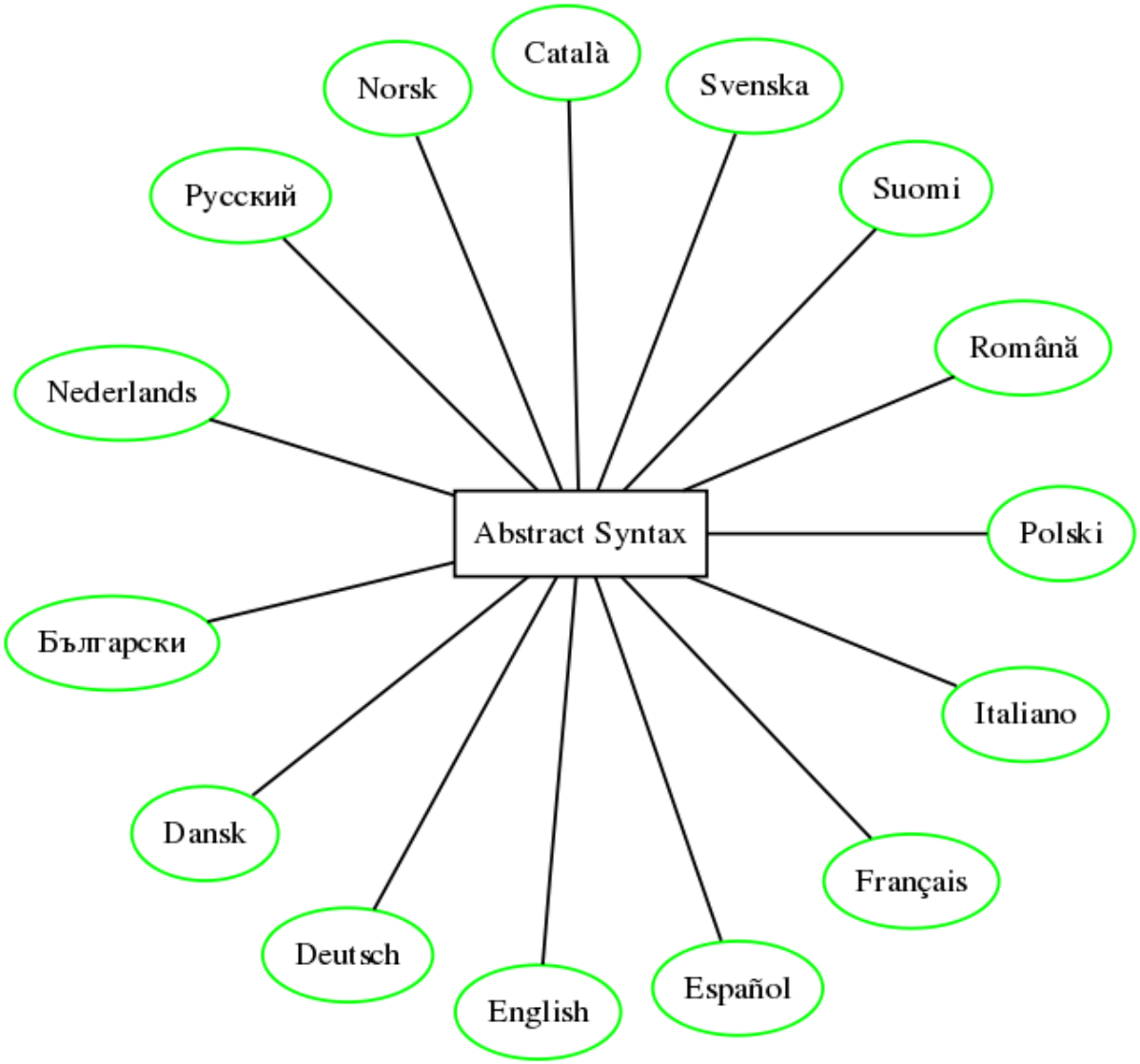
GF, grammaticalframework.org

- Domain-specific interlingua + concrete syntaxes
- GF Resource Grammar Library
- Incremental parsing
- Syntax editing

OWL Ontologies

Statistical Machine Translation

# 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

*The technology is there - MOLTO will apply it and scale it up.*

## Domain-specific interlinguas

The abstract syntax must be formally specified, well-understood

- semantic model for translation
- fixed word senses
- proper idioms

For instance: a mathematical theory, an ontology

# Example: social network

Abstract syntax:

```
cat Message ; Person ; Item ;
fun Like : Person -> Item -> Message ;
```

Concrete syntax (first approximation):

```
lin Like x y = x ++ "likes" ++ y     -- Eng
lin Like x y = x ++ "tycker om" ++ y -- Swe
lin Like x y = y ++ "piace a" ++ x   -- Ita
```

## Complexity of concrete syntax

Italian: agreement, rection, clitics (*il vino piace a Maria* vs. *il vino mi piace* ; *tu mi piaci*)

```
lin Like x y = y.s ! nominative ++ case x.isPron of {
  True  => x.s ! dative ++ piacere_V ! y.agr ;
  False => piacere_V ! y.agr ++ "a" ++ x.s ! accusative
  }
oper piacere_V = verbForms "piaccio" "piaci" "piace" ...
```

Moreover: contractions (*tu piaci ai bambini*), tenses, mood, ...

## Two things we do better than before

**No** universal interlingua:

- *The Rosetta stone is not a monolith, but a boulder field.*

**Yes** universal concrete syntax:

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

# The GF Resource Grammar Library

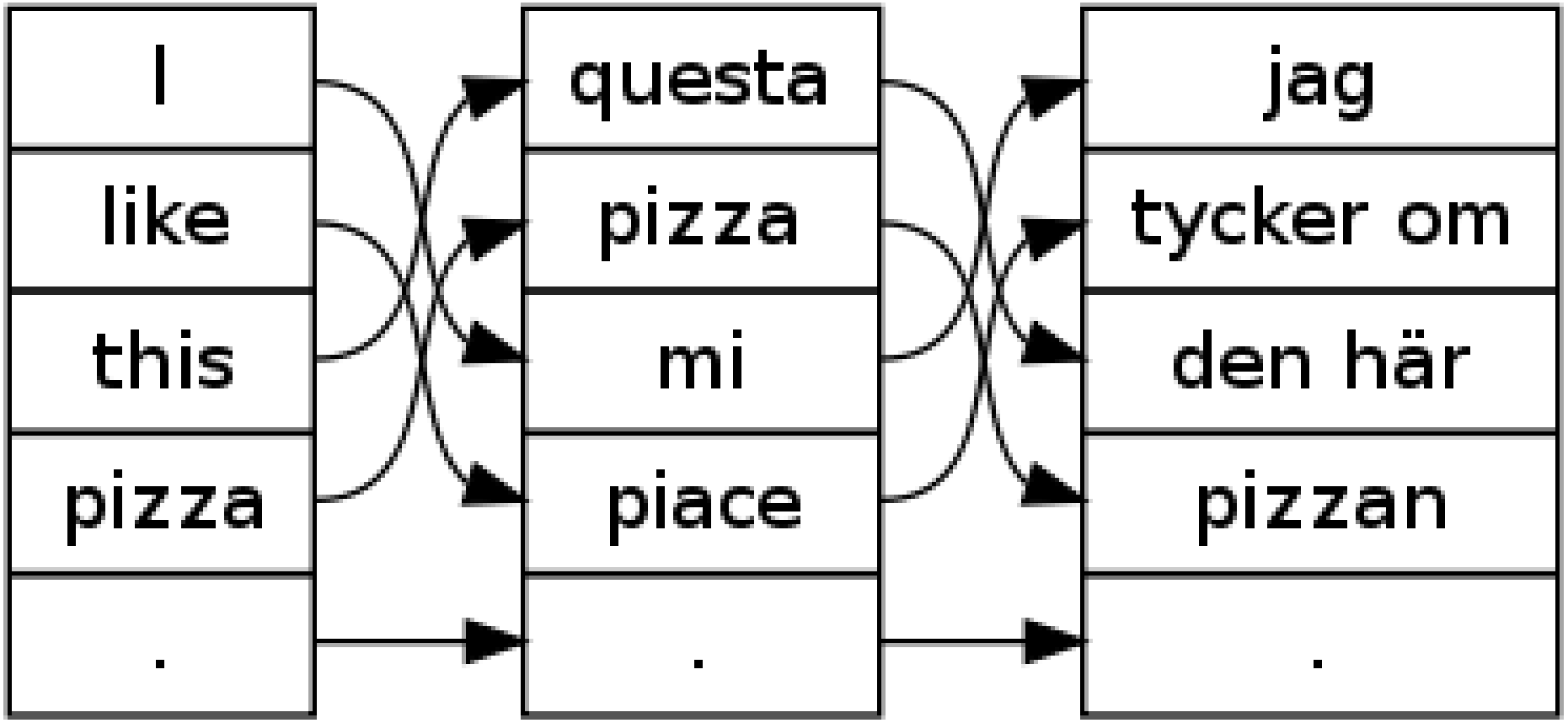
Currently for 16 languages; 3-6 months for a new language.

Complete morphology, comprehensive syntax, lexicon of irregular words.

Common syntax API:

```
lin Like x y = mkCl x (mkV2 (mkV "like")) y         -- Eng
lin Like x y = mkCl x (mkV2 (mkV "tycker") "om") y -- Swe
lin Like x y = mkCl y (mkV2 piacere_V dative) x    -- Ita
```

## Word/phrase alignments via abstract syntax



## Domains for case studies

Mathematical exercises (<- WebALT)

Patents in biomedical and pharmaceutical domain

Museum object descriptions

## Other potential uses

Wikipedia articles

E-commerce sites

Medical treatment recommendations

Tourist phrasebooks

Social media

SMS

## Challenge: grammar tools

Scale up production of domain interpreters

- from 100's to 1000's of words
- from GF experts to domain experts and translators
- from months to days
- writing a grammar $\approx$ translating a set of examples

# Example-based grammar writing

| | | |
|---|---|---|
| Abstract syntax | `Like She He` | first grammarian |
| English example | *she likes him* | first grammarian |
| German translation | *er gefällt ihr* | human translator |
| resource tree | `mkCl he_Pron gefallen_V2 she_Pron` | GF parser |
| concrete syntax rule | `Like x y = mkCl y gefallen_V2 x` | variables renamed |

## Challenge: translator's tools

Transparent use:

- text input + prediction
- syntax editor for modification
- disambiguation
- on the fly extension
- normal workflows: API for plug-ins in standard tools, web, mobile phones...

## Demos

MOLTO phrasebook on the web

MOLTO phrasebook on Android phones

Editor in `grammaticalframework.org:41296/editor`

## Innovation: OWL interoperability

Transform web ontologies to interlinguas

Pages equipped with ontologies... will soon be equipped by translation systems

Natural language search and inference

## Scientific challenge: robustness and statistics

1. Statistical Machine Translation (SMT) as fall-back

2. Hybrid systems

3. Learning of GF grammars by statistics

4. Improving SMT by grammars

# Learning GF grammars by statistics

| | | |
|---|---|---|
| Abstract syntax | `Like She He` | first grammarian |
| English example | *she likes him* | first grammarian |
| German translation | *er gefällt ihr* | **SMT system** |
| resource tree | `mkCl he_Pron gefallen_V2 she_Pron` | GF parser |
| concrete syntax rule | `Like x y = mkCl y gefallen_V2 x` | variables renamed |

Rationale: SMT is *good* for sentences that are *short* and *frequent*

## Improving SMT by grammars

Rationale: SMT is *bad* for sentences that are *long* and involve *word order variations*

*if you like me, I like you*

`If (Like You I) (Like I You)`

*wenn ich dir gefalle, gefällst du mir*

## Availability of MOLTO tools

Open source, LGPL (*except* parts of the patent case study)

Web demos

Mobile applications

Events:

- MOLTO meeting Gothenburg 9-11 March 2011
- GF/MOLTO Summer School Barcelona 15-26 August 2011

# Conclusion

You shouldn't expect

- general-purpose translation ("Google competitor")

You can expect

- high quality multilingual translation
- portability to limited domains (up to 1000's of words)
- productivity (days, weeks, months)