

# The First Year of MOLTO

Aarne Ranta

Second MOLTO Meeting, Gothenburg 9 March 2011



# Multilingual Online Translation

Non multa, sed multum not quantity but quality

ABOUT

NEWS

EVENTS

MOLTO's mission is to develop a set of tools for translating texts between *multiple languages* in *real time* with *high quality*. MOLTO will use multilingual grammars based on semantic interlinguas.

FP7-ICT-247914, Strep, [www.molto-project.eu](http://www.molto-project.eu)

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

March 2010 - February 2013

## What's new?

| <b>Tool</b> | <b>Google, Babelfish</b> | <b>MOLTO</b> |
|-------------|--------------------------|--------------|
| target      | consumers                | producers    |
| input       | unpredictable            | predictable  |
| coverage    | unlimited                | limited      |
| quality     | browsing                 | publishing   |

## Producer's quality

Cannot afford translating French

- *prix 99 euros*

to Swedish

- *pris 99 kronor*

Typical SMT error due to parallel corpus containing localized texts.  
(N.B. 99 kronor = 11 euros)

# Reliability

German to English

- *er bringt mich um -> he is killing me*

correct, but

- *er bringt meinen besten Freund um -> he brings my best friend for*

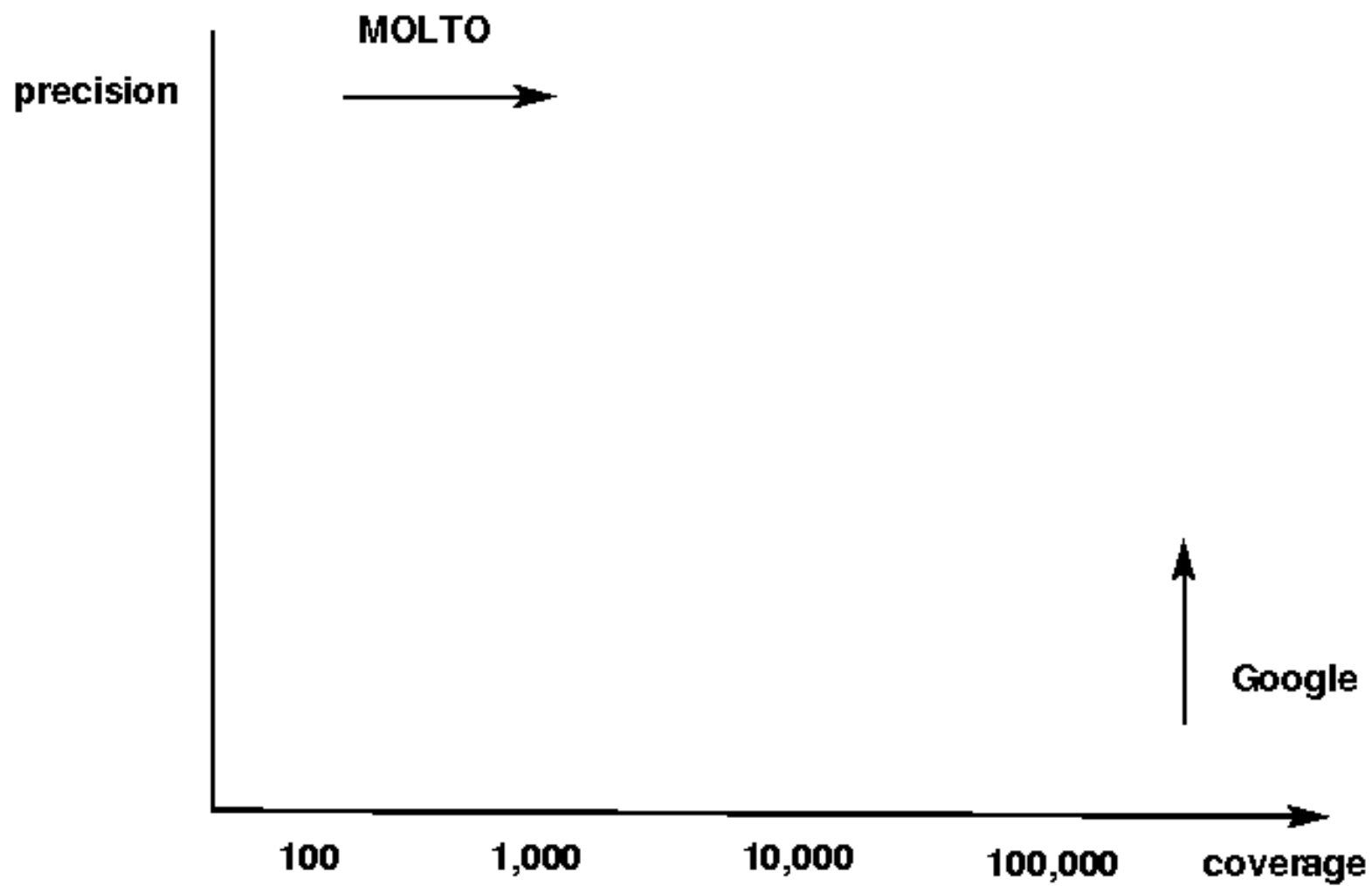
should be *he kills my best friend*. (Typical error due to **long distance dependencies**, causes **unpredictability**)

## **Aspects of reliability**

Separation of levels (syntax, semantics, pragmatics, localization)

Predictability (generalization for similar constructs, and over time)

Programmability / debugging and fixing bugs (vs. holism)



# 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

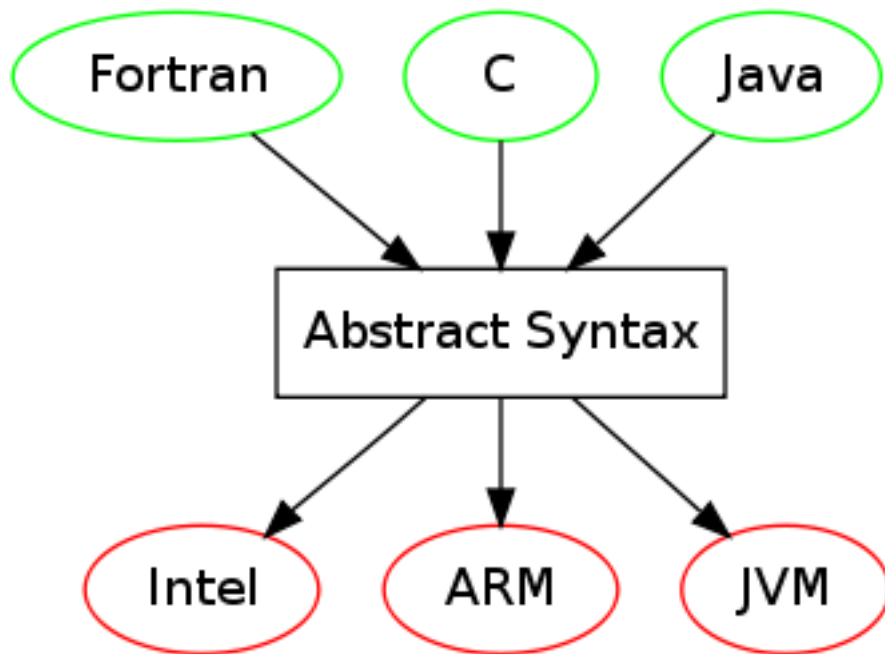
GF, [grammaticalframework.org](http://grammaticalframework.org)

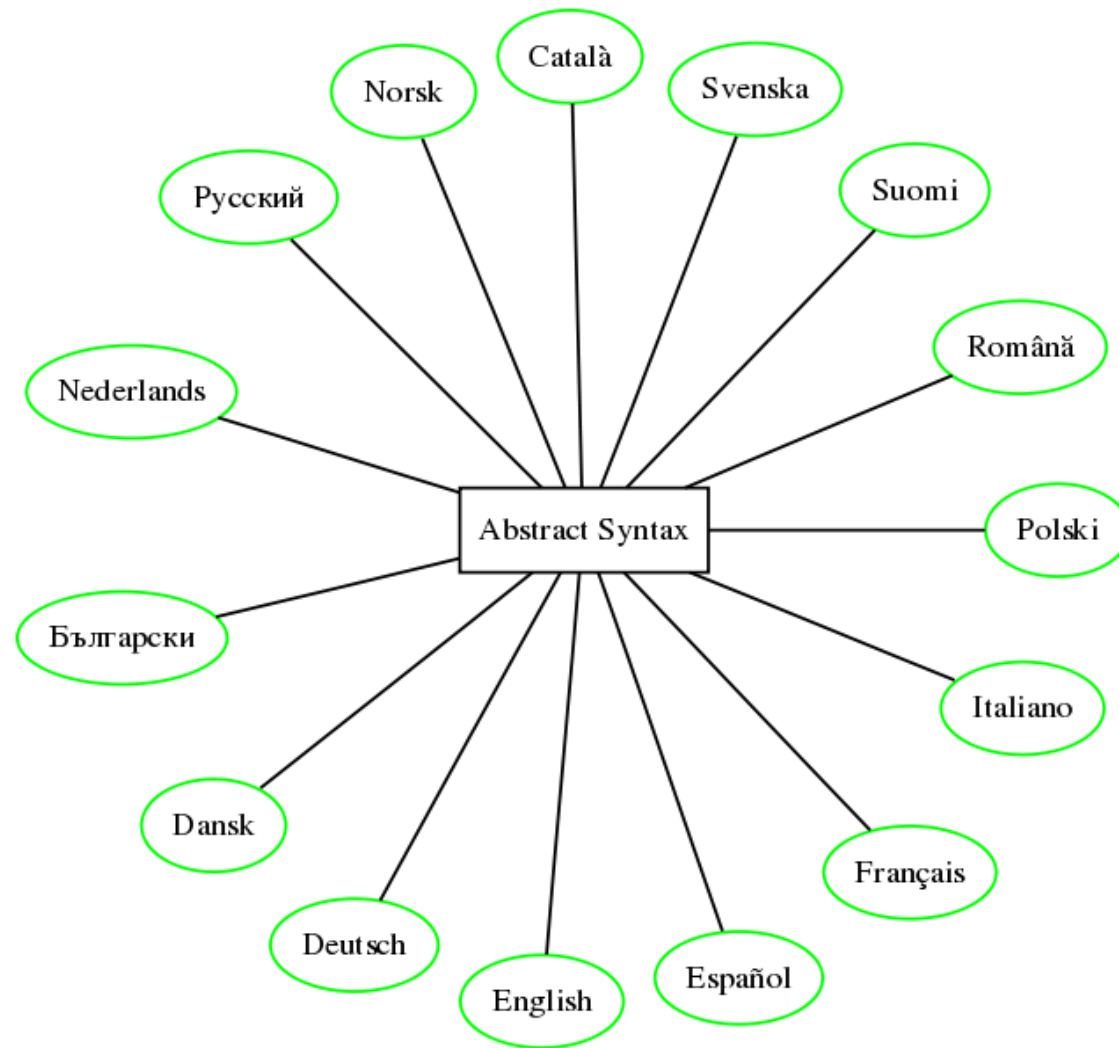
- Domain-specific interlingua + concrete syntaxes
- GF Resource Grammar Library
- Incremental parsing
- Syntax editing

OWL Ontologies

Statistical Machine Translation

## 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

*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 - anything that is definable in **type theory**

## Example: social network

Abstract syntax:

```
fun Like : Person -> Item -> Fact
```

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

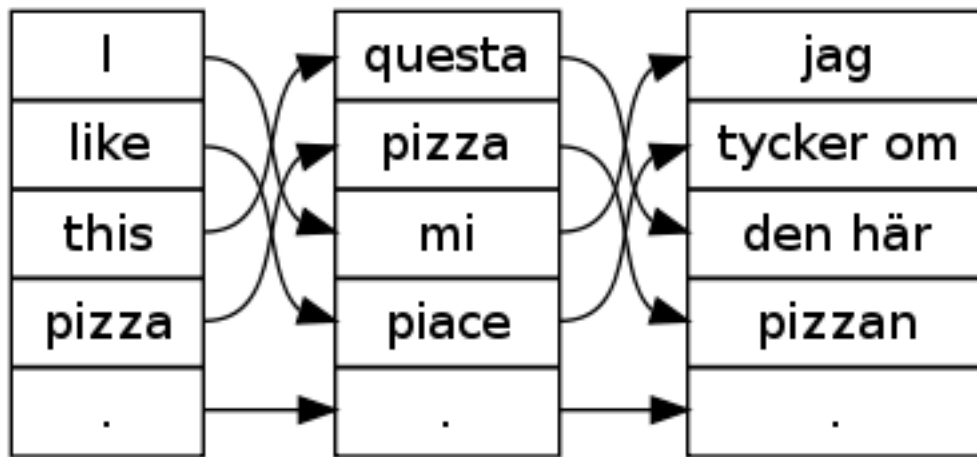
Currently for 16 languages; 3-6 months for a new language.

Complete morphology, comprehensive syntax, lexicon of irregular words.

Common syntax API:

```
lin Like x y = mkC1 x (mkV2 (mkV "like")) y          -- Eng
lin Like x y = mkC1 x (mkV2 (mkV "tycker") "om") y   -- Swe
lin Like x y = mkC1 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

Demo: a tourist phrasebook (web and Android phones)

## **Other potential uses**

Wikipedia articles

E-commerce sites

Medical treatment recommendations

Social media

SMS

Contracts

## 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      | <i>she likes him</i>              | first grammarian  |
| German translation   | <i>er gefällt ihr</i>             | human translator  |
| resource tree        | mkC1 he_Pron gefallen_V2 she_Pron | GF parser         |
| concrete syntax rule | Like x y = mkC1 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...

## **Innovation: OWL interoperability**

Transform web ontologies to interlinguas

Pages equipped with ontologies... may 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      | <i>she likes him</i>              | first grammarian  |
| German translation   | <i>er gefällt ihr</i>             | <b>SMT system</b> |
| 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*

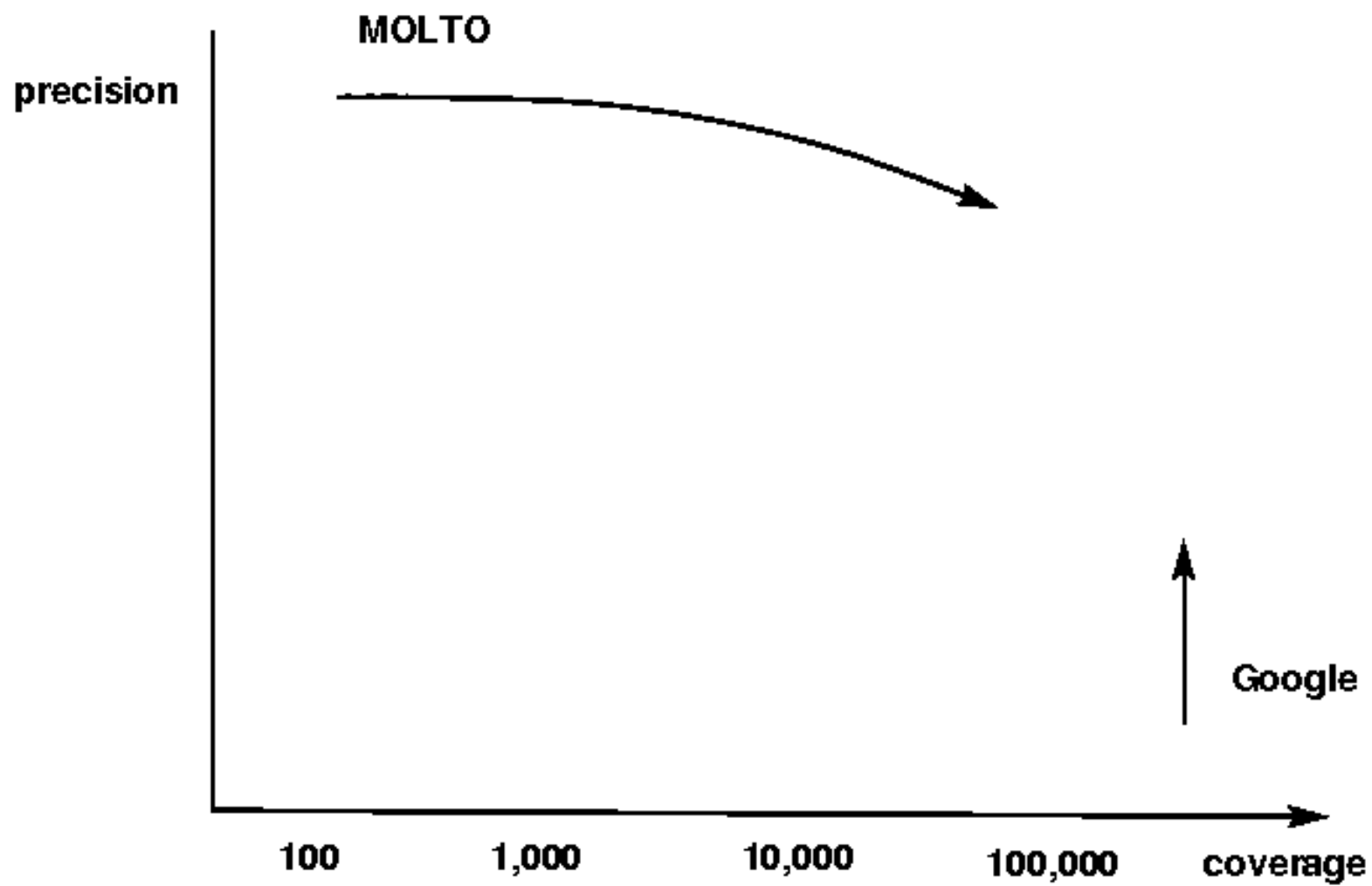
## Improving SMT by grammars

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, gefälltst du mir*



## **Availability of MOLTO tools**

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

Web demos

Mobile applications (Android)

## Highlights of the first year

WP2: Grammar Development tools

- web-based grammar development environment
- Term Factory
- multilingual resource grammar API

## WP3: Translator's tools

- web-based translation interface
- Android on-board translator
- Java, C, and Python ports of GF

## WP4: Knowledge engineering

- GF-OWL interoperability
- the MOLTO KRI

## WP5: Statistical and robust parsing

- phrase alignments and probabilities in GF
- hybrid GF/SMT decoding

## WP6: Mathematics case study

- OpenMath exercise grammar library in 10 language

## WP7: Patents case study

- good domain-specific SMT system for medical patents



## WP9: Evaluation

- syntax and semantics based evaluation methods

## WP10: Dissemination

- MOLTO phrasebook
- GF tutorials: LREC-2010, CNL-2010, CADE-2011
- publications
- GF Summer School 2011

# Distinguished speakers / MOLTO Advisory Board

Prof. Stephen Pulman, University of Oxford

- language-based human-computer interaction
- grammar-based machine translation (Core Language Engine)

Dr. Keith Hall, Google Zürich

- structured models in automatic language processing
- statistical machine translation

# Conclusion

You shouldn't expect

- general-purpose translation ("Google competitor")

You should expect

- high quality multilingual translation
- portability to new domains (up to 1000's of words)
- productivity (days, weeks, months)
- ease of use (no training for authoring, a few days for grammarians)