MOLTO: Multilingual On-Line Translation

Aarne Ranta

Xerox Research Centre Europe, Grenoble, 19 January 2012

MOLTO Multilingual Online Translation

Non multa, sed multum *not quantity but quality*

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`

U Gothenburg, U Helsinki, UPC Barcelona, Ontotext (Sofia), U Zurich, Be Informed (Apeldoorn)

March 2010 - May 2013

EC contribution 2,975,000 EUR

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 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**)

Linguistic knowledge

(From Google Translate 1 September 2011)

Finnish: *yö, yön, yötä, yönä, yöksi, yössä, yöstä, yöhön, yöllä, yöltä, yölle, yöttä, öineen, öin, yöt, öitä, öiden, öinä, öiksi, öissä, öistä, öihin, öillä, öiltä, öille, öittä, öin*

English: *Night, night, night, night, night, night, night, night, night, night, nights, yöttä, öineen, night, night, nights, nights, nights States by quotas, domestic insurance companies, nights, nights, öillä, against loss, States, öittä, night*
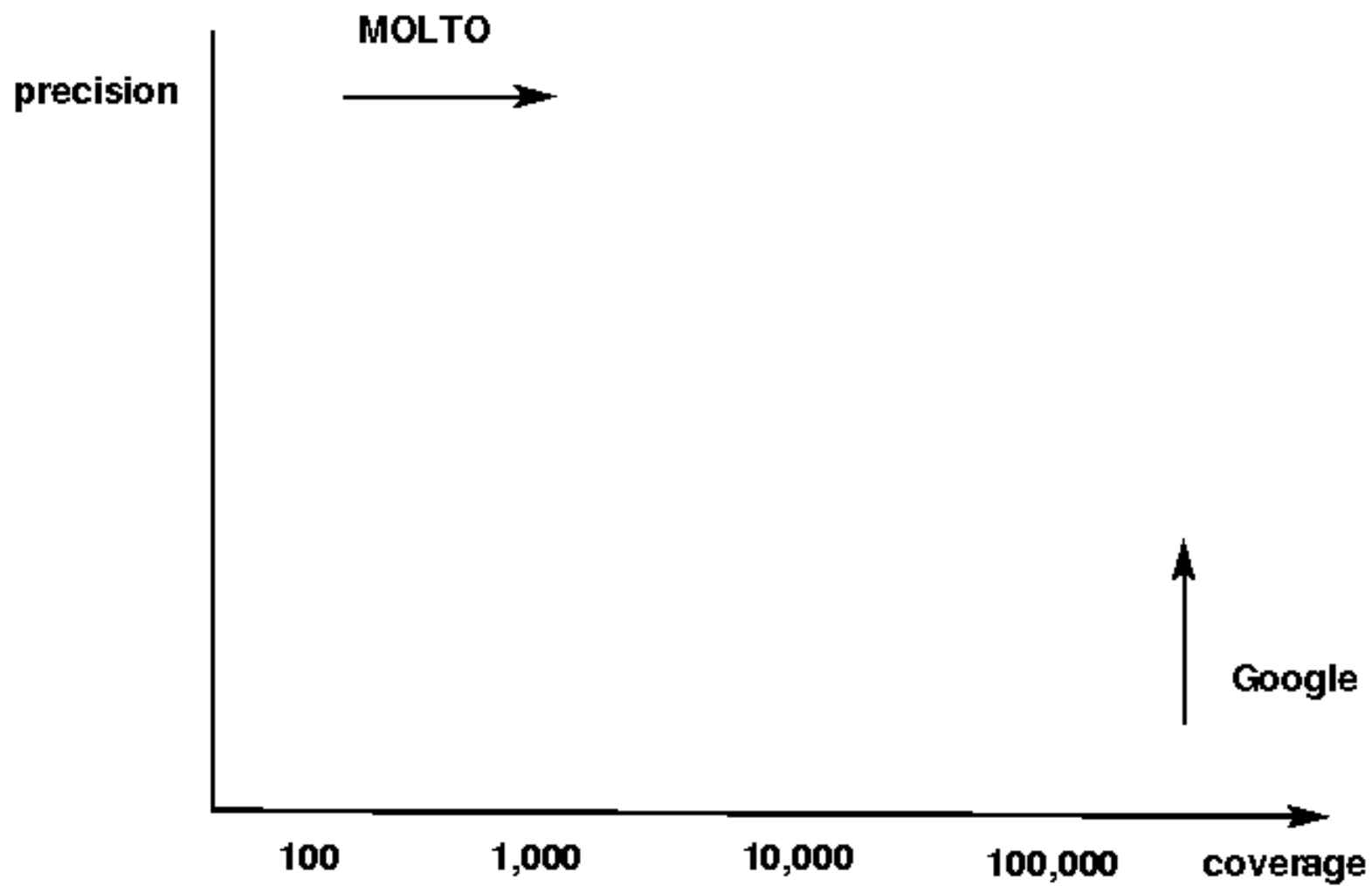
Aspects of reliability

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

Predictability (generalization for similar constructs, and over time)

Programmability / debugging and fixing bugs (vs. holism)

But there's a trade-off between coverage and precision: we cannot deal with millions of concepts.

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

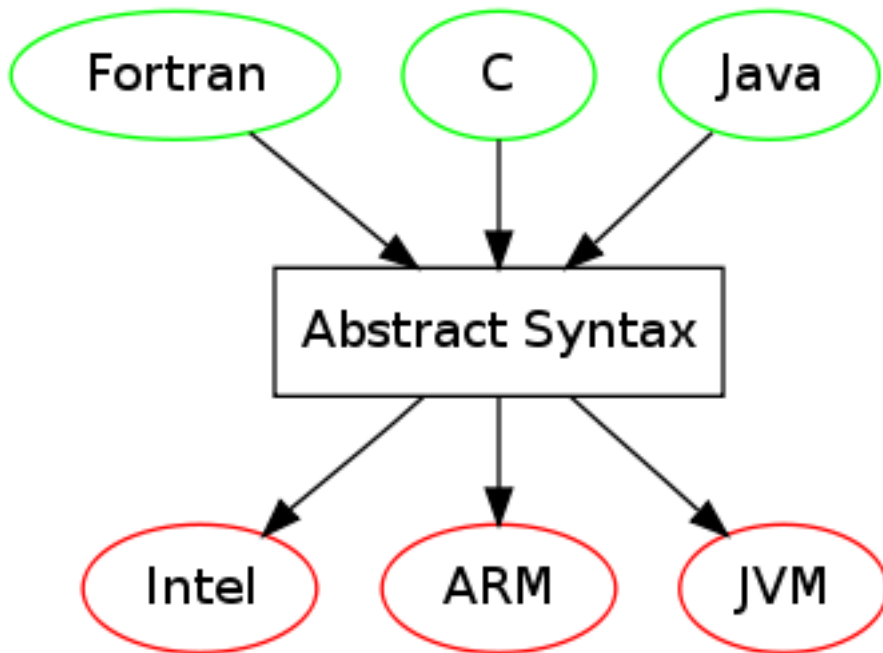- Finnish cases
- German word order

Main technologies

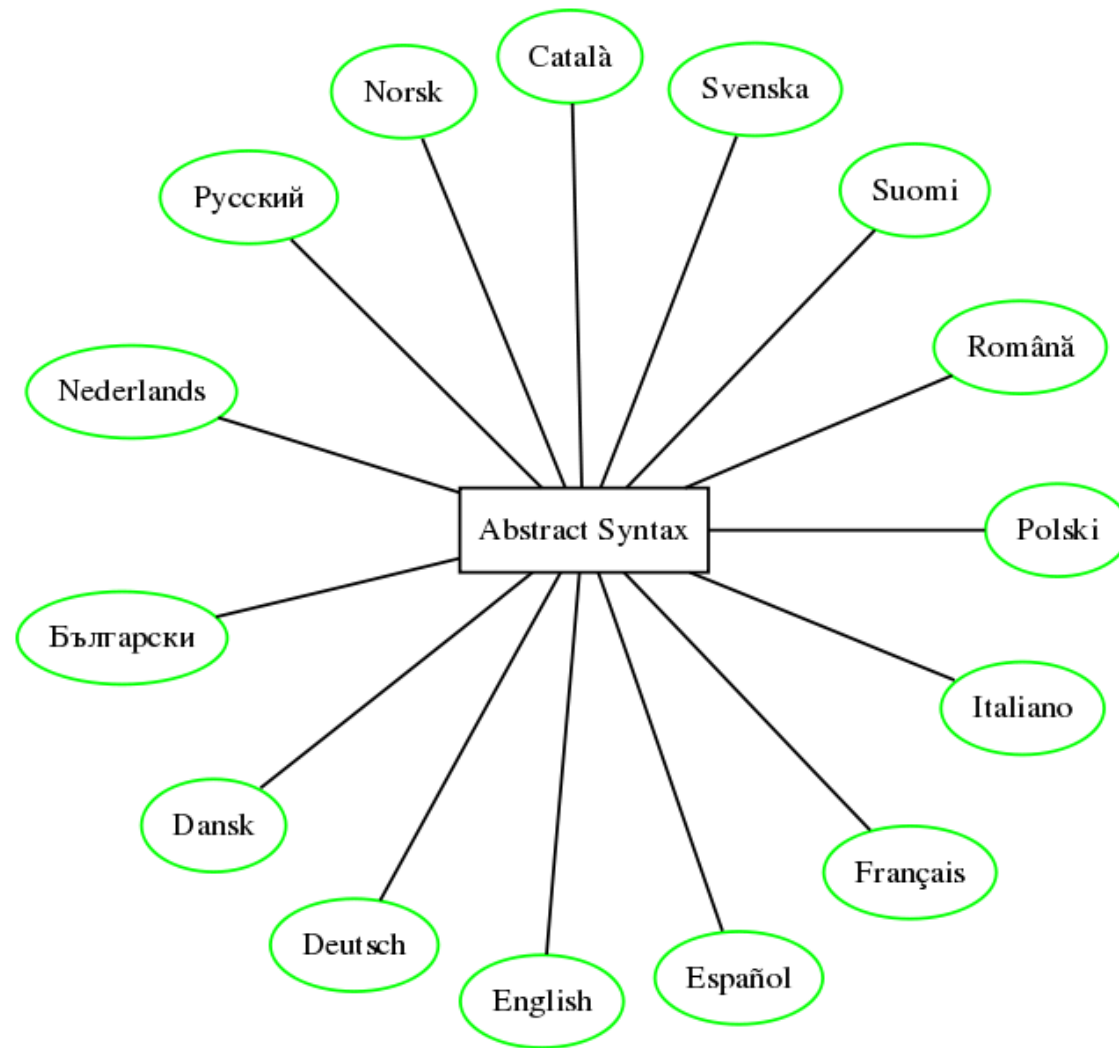GF, grammaticalframework.org

- "compiling natural languages"
- 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
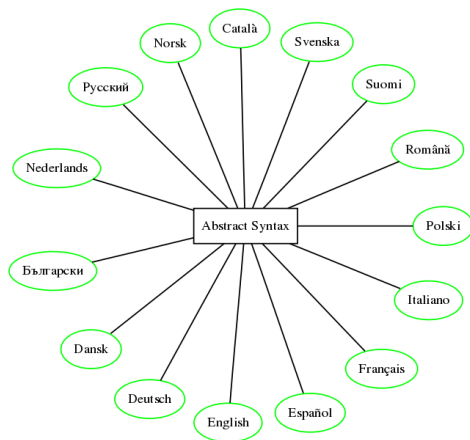
MOLTO languages

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**

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**

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, ...
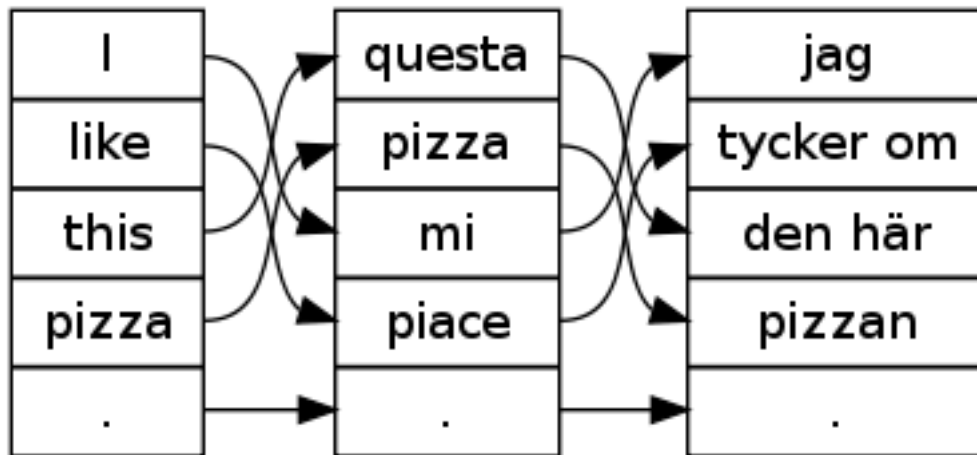
The GF Resource Grammar Library

Currently for 22 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
```

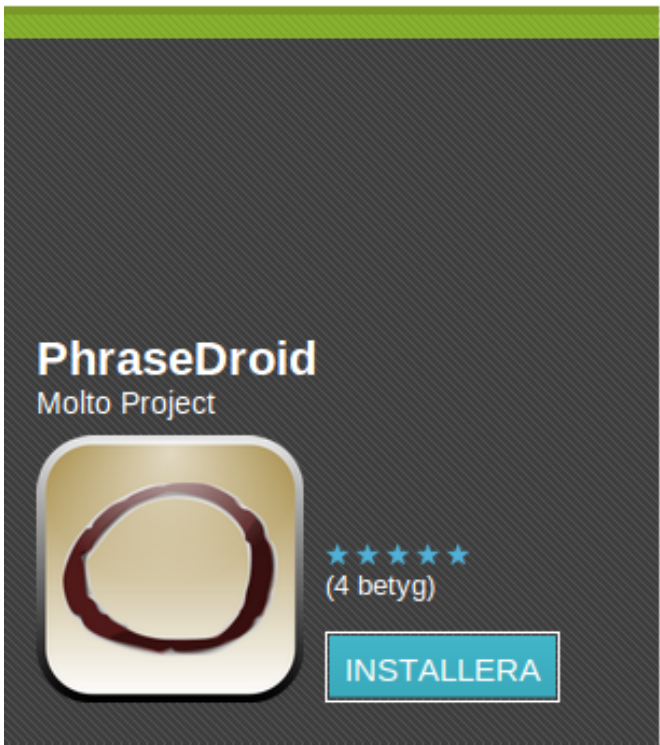# Word/phrase alignments via abstract syntax

Domains for case studies

Mathematical exercises (<- WebALT)

Patents in biomedical and pharmaceutical domain

Museum object descriptions

Demo: a tourist phrasebook (web and Android phones)

Android Market

Sök

PhraseDroid
Molto Project

★ ★ ★ ★ ★
(4 betyg)

INSTALLERA

Other potential uses

Wikipedia articles

E-commerce sites

Medical treatment recommendations

Social media

SMS

Contracts

Controlled language

Almost what MOLTO is, except that we

- generalize this to **multilingual controlled language systems**
- support ambiguous language

Prime example: Attempto Controlled English (U Zurich)

- generalized to 5 languages in GF (CNL 2009)
- will be extended to 15 and a **multilingual semantic wiki** in MOLTO

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_NP gefallen_V2 she_NP`	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: plug-ins in standard translator tools, web, mobile phones...

Demo: the MOLTO phrasebook

http://www.grammaticalframework.org/demos/phrasebook/

text input + prediction

(not yet: syntax editor for modification)

disambiguation

(not yet: on the fly extension)

normal workflows: plug-ins in standard translator 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	*she likes him*	first grammarian
German translation	*er gefällt ihr*	**SMT system**
resource tree	`mkCl he_NP gefallen_V2 she_NP`	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*
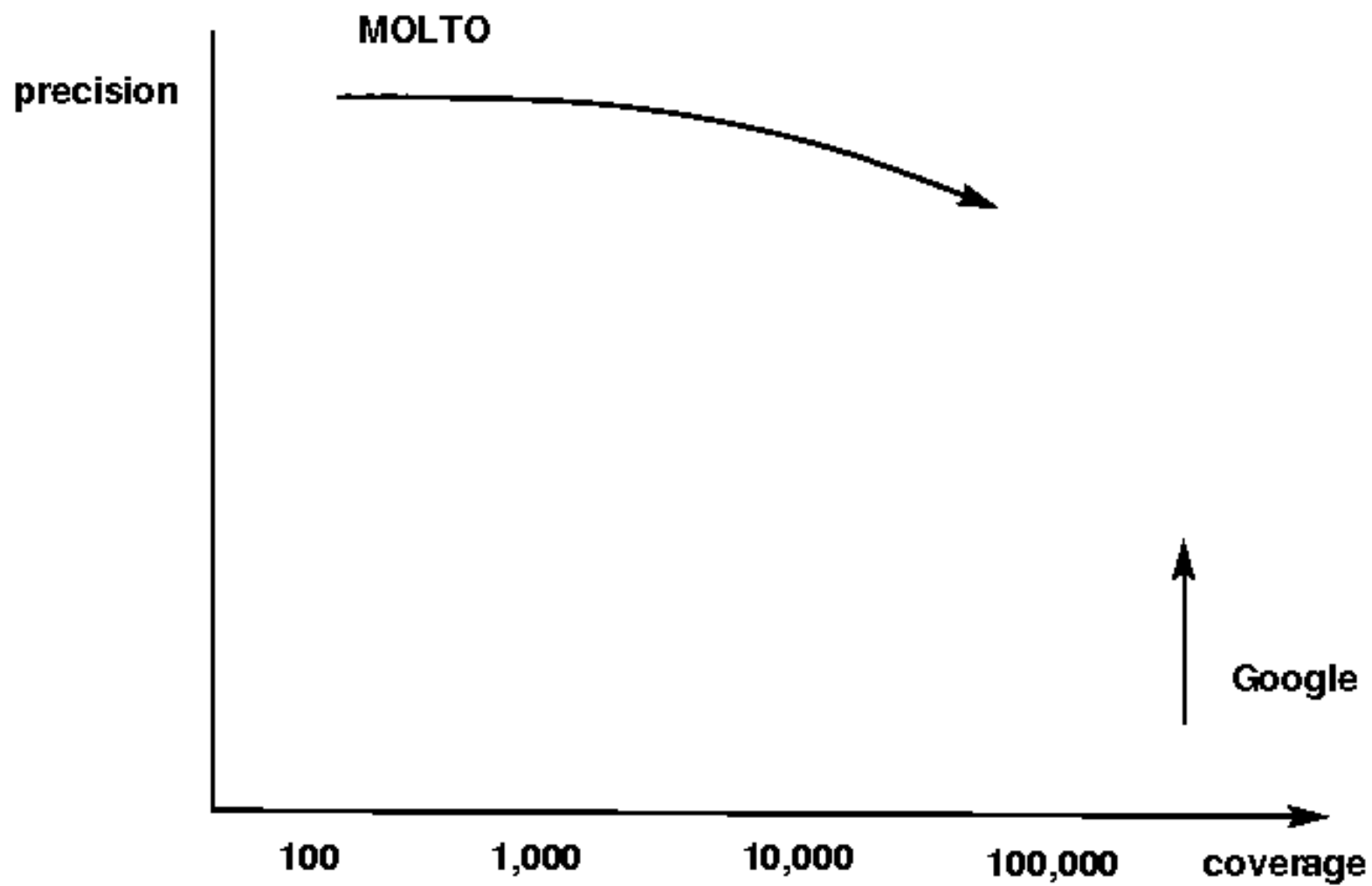
Improving SMT by grammars

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

A possible scenario: controlled trade-off precision/quality

Availability of MOLTO tools

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

Web demos

Mobile applications (Android)

Demos

Phrasebook: http://www.grammaticalframework.org/demos/phrasebook/

Grammar editor: http://www.grammaticalframework.org/demos/gfse/

Eclipse IDE: http://www.grammaticalframework.org/eclipse/

Ontology queries: "MOLTO KRI" in http://www.molto-project.eu/

Mathematics: http://www.grammaticalframework.org/demos/minibar/mathba

Phrasedroid: https://market.android.com/details?id=org.grammaticalframewo

The RGL API: http://www.grammaticalframework.org/lib/synopsis.html

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)