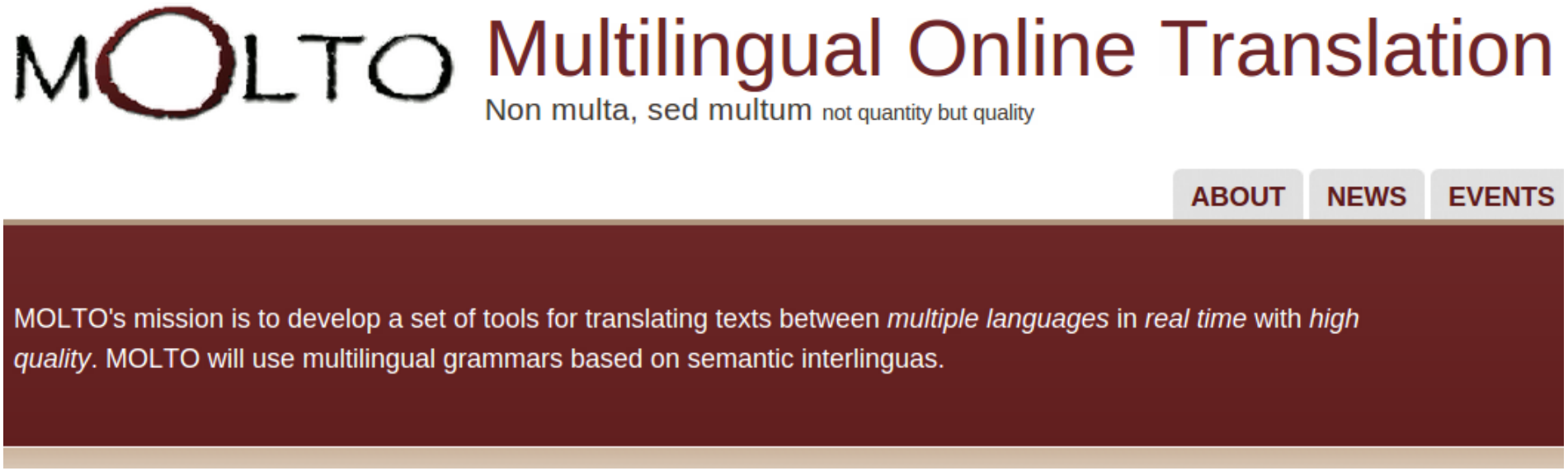
MOLTO: Multilingual Online Translation

Aarne Ranta

Monnet-MOLTO Meeting, Utrecht, 19 September 2012



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 different?

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

Producer's quality

Responsibility for the translation

Cannot afford translating French

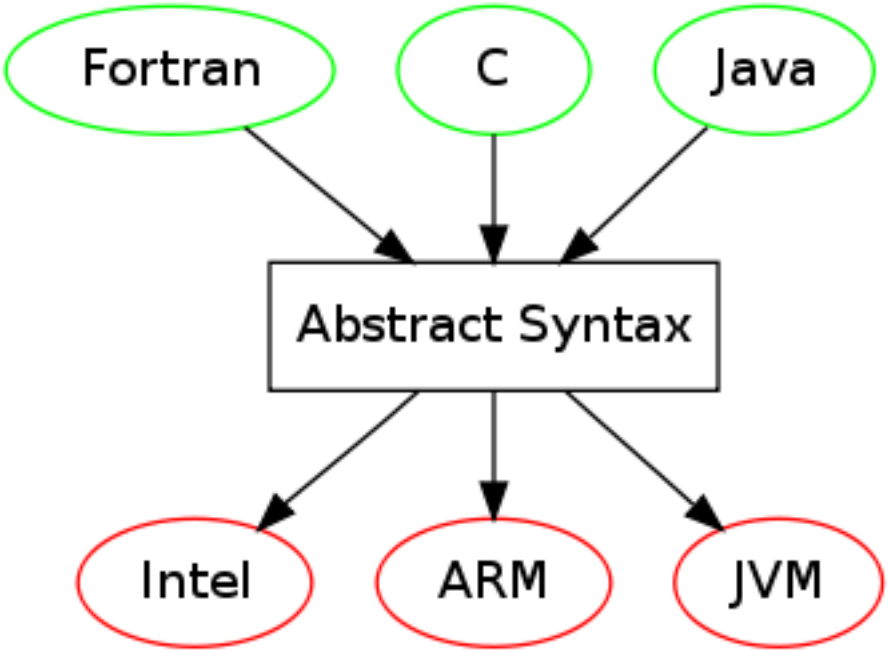
- *prix 99 euros*

to Swedish

- *pris 99 kronor*

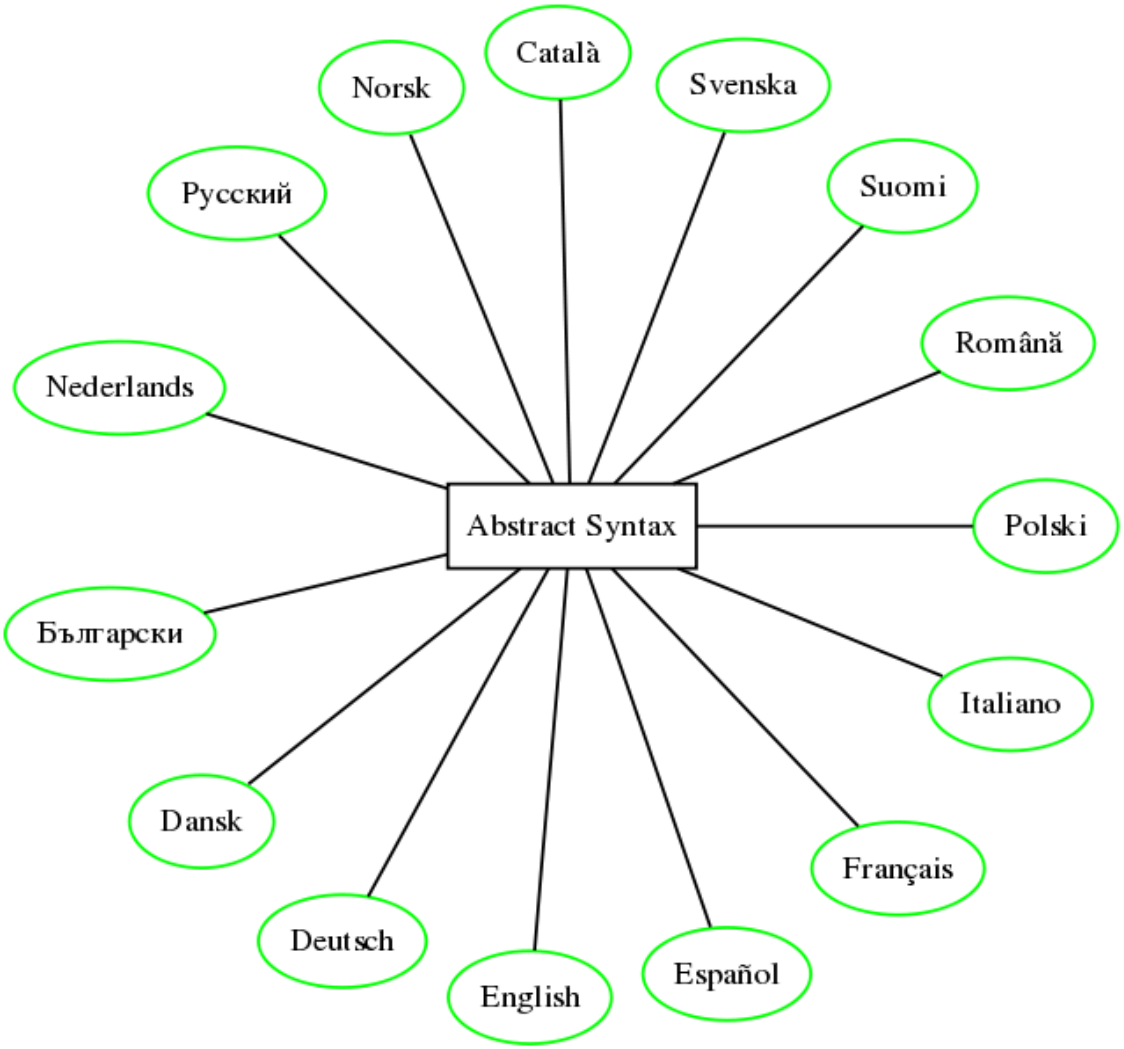
Typical SMT error due to parallel corpus containing localized texts. (N.B. 99 kronor = 11 euros)

Where interlingua rules today: compilers




```
                                                                    bipush 6
6 + x  -->  (EAdd TInt (EInt 6) (EVar TInt "x")) -->    iload 8
                                                                    iadd
```

MOLTO languages



Two things learned from the past

No universal interlingua:

- instead, domain-specific interlinguas $\sim$ ontologies
- *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**

Grammar example: the predicate "x likes y"

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 = x ++ "pitää" ++ y ++ "stä" -- Fin
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 25 languages; 3-6 months for a new language.

40+ contributors 2001-

Complete morphology, comprehensive syntax, lexicon of irregular words.

Afrikaans	Bulgarian	Catalan	Danish	Dutch
English	Finnish	French	German	Hindi
Italian	Japanese	Latvian	Nepali	Norwegian
Persian	Punjabi	Polish	Romanian	Russian
Sindhi	Spanish	Swedish	Thai	Urdu

Smart paradigms

```
mkV "like"
mkV "cry"
mkV "go" "went" "gone"

mkV "pitää"
mkV "talo"
mkV "poika" "pojan"
```

Concrete syntax for predications

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 x (mkV2 (mkV "pitää") elative) y -- Fin
lin Like x y = mkCl y (mkV2 piacere_V dative) x       -- Ita
```

From predications to complex phrases (MOLTO Phrase-book)

Bul:
Аз знам , че нейният съпруг харесва това италианско вино.
Cat:
Sé que aquest vi italià agrada a el seu home.
Dan:
Jeg ved at hendes mand holder af denne italienske vin.
Dut:
Ik weet dat haar man van deze Italiaanse wijn houdt.
Eng:
I know that her husband likes this Italian wine.
Fin:
Tiedän että hänen miehensa pitää tästä italialaisesta viinistä.
Fre:
Je sais que ce vin italien plaît à son mari.
Ger:
Ich weiß , dass ihr Mann diesen italienischen Wein mag.

मैं जानता हूँ कि उस का पति इस इतालवी शराब को पसंद करता है.

-- I (male) know that her husband likes this Italian wine.

मैं जानती हूँ कि उस का पति इस इतालवी शराब को पसंद करता है.

-- I (female) know that her husband likes this Italian wine.

Ita:

So che questo vino italiano piace a suo marito.

Lav:

Es zinu , ka šis itāļu vīns garšo viņas vīram.

Nor:

Jeg vet at mannen hennes liker denne italienske vinen.

Pes:

من می دانم کہ شوهر او این شراب ایتالیایی را دوست دارد Û”

Pol:

Wiem , że jej mąż lubi to włoskie wino.

Ron:

Eu ştiu cã soţului sãu îi place acest vin italian.

Rus:

Я знаю , что её муж нравдит это итальянское вино.

Spa:

Sé que este vino italiano gusta a su marido.

Swe:

Jag vet att hennes man tycker om det här italienska vinet.

Tha:

ผม รู้ ว่า สามี ของ หล่อน ชอบ เหล้าองุ่น อิตาเลียน ขวด นี้

-- I (male) know that her husband likes this Italian wine.

ฉัน รู้ ว่า สามี ของ หล่อน ชอบ เหล้าองุ่น อิตาเลียน ขวด นี้

-- I (female) know that her husband likes this Italian wine.

From sentence to text (MOLTO Museum Grammar)

```
MkGenText GSM980019Obj AnnaLindskog OilPainting (MkColour Black) (MkSize (SIntInt 435 365))
  (MkMaterial Canvas) (MkYear (YInt 1885)) (MkMuseum GoteborgsCityMuseum)
```

- PaintingEng: The girl was painted on canvas by Anna Lindskog in 1885. It is of size 435 by 365 and it is painted in black. This oil painting is displayed at the City Museum of Gothenburg.

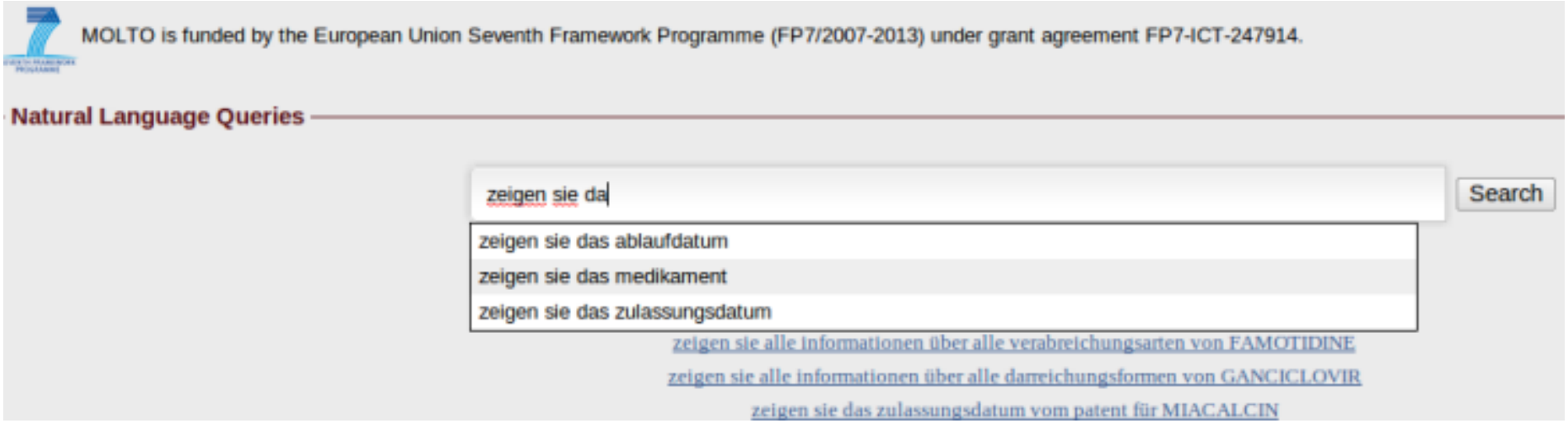
- PaintingFin: Maalauksen Flickan on maalannut Anna Lindskog kankaalle vuonna 1885. Se on kokoa 435 kertaa 365 ja se on maalattu mustalla. Tämä öljymaalaus on esillä Göteborgin kaupunginmuseossa.

- PaintingFre: Le tableau Flickan a été peint sur toile par Anna Lindskog en 1885. Il est de taille 435 sur 365 et il est peint en noir. Cette peinture à l' huile est exposée dans le musée municipal de Göteborg.

- PaintingIta: Il quadro Flickan è stato dipinto su tela da Anna Lindskog nel 1885. Misura 435 per 365 ed è dipinto in nero. Questo dipinto ad olio è esposto nel museo municipale di Goteburgo.

- PaintingSwe: Flickan målades på duk av Anna Lindskog år 1885. Den är av storlek 435 gånger 365 och den är målad i svart. Den här oljemålningen är utställd på Göteborgs stadsmuseum.

Using grammars for querying (MOLTO Patent Grammar)



Workflow summary

Give: ontology

- **Obtain**: abstract syntax

Give: concrete syntax

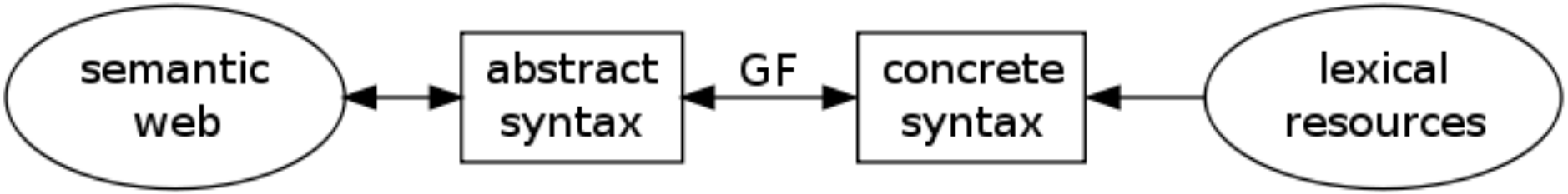
- **Obtain**: verbalization, querying

Vision

Web pages equipped with ontologies...

will soon be extended to translation systems!

GF, Semantic Web, lexical resources



Hands-on example

Localize the painting grammar to Dutch.

The grammar code is 196 lines.

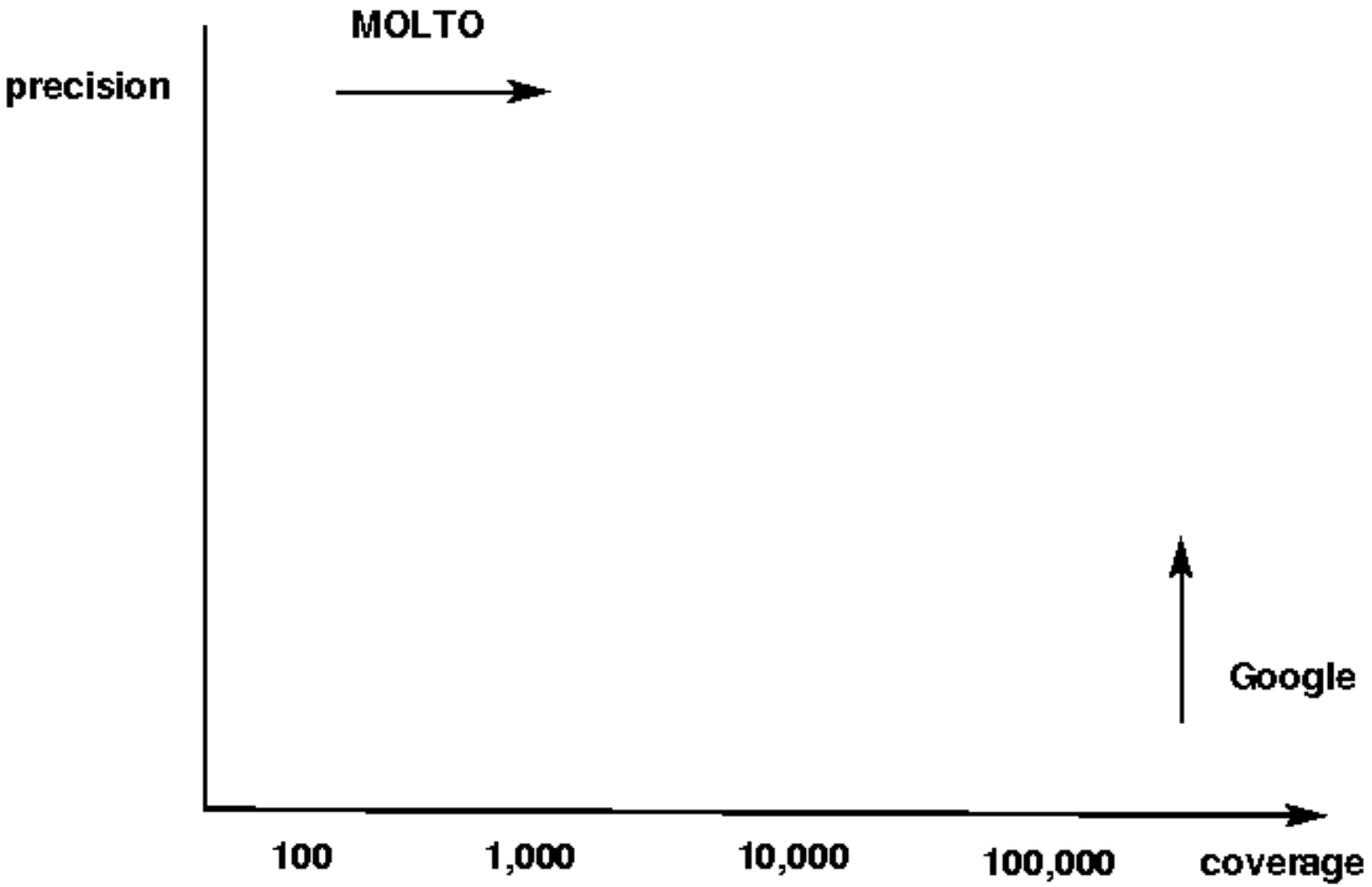
Changes needed on 19 lines.

```
paint_V2 = mkV2 "paint" ;
display_V2 = mkV2 "display" ;
at_Prep = mkPrep "at" ;
by_Prep = mkPrep "by" ;  -- the size is 50 by 40 cm
painting_N = mkN "painting" ;
work_N = mkN "work" ;  -- work of art
size_N = mkN "size" ;

PaintingEntity =  mkCN (mkN "painting") ;
PortraitPainting  =  mkCN (mkN "portrait") ;
OilPainting = mkCN (mkN "oil painting") ;
```

```
WatercolorPainting = mkCN (mkN "watercolour painting") ;
Paper = mkCN (mkN "paper") ;
Wood = mkCN (mkN "wood") ;
Linen = mkCN (mkN "linen") ;
Canvas = mkCN (mkN "canvas") ;
White = mkNP (mkPN "white") ; -- X is painted in white
Black = mkNP (mkPN "black") ;
Grey =  mkNP (mkPN "grey") ;
GoteborgsCityMuseum = mkNP (mkPN "the City Museum of Gothenburg") ;
```

Scaling up

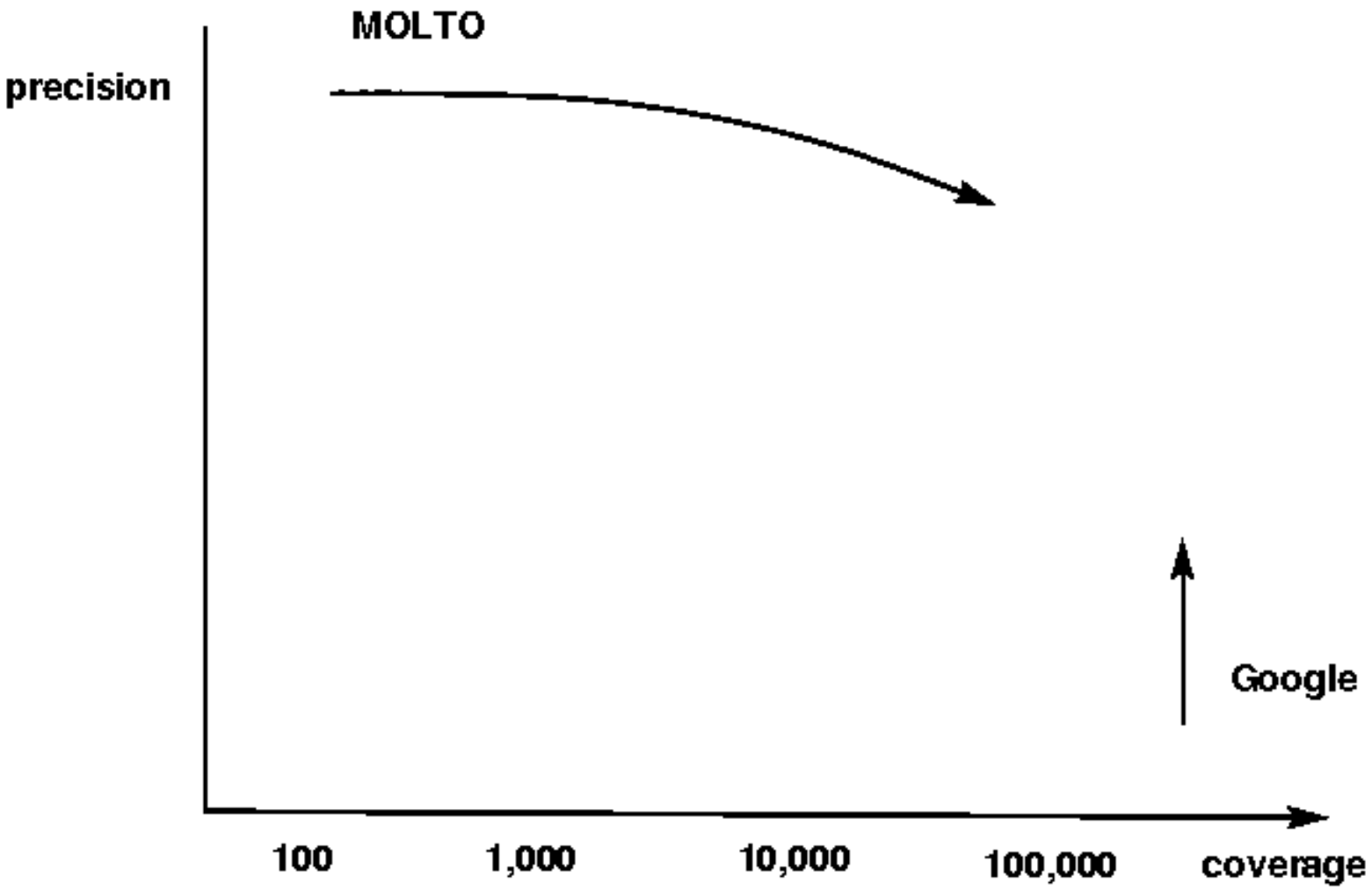


Learning GF grammars from examples

Abstract syntax	`Like She He`	first grammarian
English example	*she likes him*	first grammarian
German translation	*er gefällt ihr*	translator/SMT syste
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*

A possible scenario: controlled trade-off precision/quality


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

MOLTO Flagships

The focus for the last year of MOLTO

"Popularization" of MOLTO technology

More directly useful than the "Work Packages"

Evaluations for each of them, showing that high quality can be reached with reasonable effort

8 flagships

Flagship 1: IDE

IDE's for GF

- cloud-based, for introductory use
  - push-button web app creation
  - collaboration

- Eclipse, for corporate use, in particular for Java developers
  - library access
  - regression testing

Flagship 2: translator tool

Professional translator tool

- integrated in work flows
- terminology extension
- predictive parsing, syntax editing, robust parsing
- manual post-editing, alternative translation servers

Flagship 3: query language

Ontology query system

- generic reusable grammar
- automatically derived domain-specific extensions.
- 15 languages

Flagship 4: hybrid translation

Hybrid patent translation system that beats its competitors

- GF+SMT soft integration
- built from existing components without explicit transfer

Flagship 5: mathematics

Mathematics Grammar Library

- math documentation (Wiki)
- tutorial (Sage, word problems)

Flagship 6: museum

Museum visitors' system

- generated descriptions from data
- query system
- 15 languages

Flagship 7: Wiki

Multilingual Semantic Wiki

- collaborative construction of documents and grammars
- integrating almost all MOLTO aspects
- generalized Attempto ACE Wiki

Flagship 8: a business case

Be Informed platform for natural language user interfaces

- ontology verbalizations
- user input + reasoning + explanations
- a tool to be handed to customers

What we have achieved

A standard approach to Controlled Natural Language

- increasing share of CNL contributions
- take care of the "concrete syntax", to focus on semantics, reasoning, etc
- use the potential for multilinguality

A technology of choice for Multilingual Semantic Web

- ontology verbalization
- ontology retrieval

Emerging trends

Hybrid systems

- this is becoming the main stream in MT
- MOLTO has found its own approach (WP5, WP7)

High-quality translation

- ”This is where the money is” (Martin Kay, FreeRBMT 2012)
- the next European call