MOLTO: Multilingual Online Translation for Public Sector Needs

Aarne Ranta

Monnet-MOLTO Meeting, Utrecht, 19 September 2012

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.

High-quality translation

15 simultaneous languages

Web applications

2010-2013, EU-FP7

MOLTO participants

University of Gothenburg, Sweden

University of Helsinki, Finland

Universitat Politecnica de Catalunya, Spain

University of Zurich, Switzerland

Ontotext, Bulgaria

Be Informed, The Netherlands

Publisher vs. consumer of information

Consumer translation:

- use it to get an idea of a text
- Google, Bing, Babelfish
- errors tolerated

Producer translation:

- use it to publish your content
- manual work by professionals
- errors not tolerated

Difference in responsibility

Dutch e-commerce site original:

prijs 99 euros

Swedish translation:

pris 99 kronor

Is the Dutch site responsible for this price?

- no, if it was translated by the consumer
- yes, if they have published the translation

(99 kronor = 12 euros)

How to make machine translation reliable

Computer programs *are* unreliable:

- bugs
- difficult to manage complexity

But some programs are very reliable:

- calculators
- compilers

Calculators and compilers

Calculator: find the value of $1431 + 431437564/789709$

Compiler: find the machine code of a program that computes the average age of Dutch citizens

```
sum[age(x) | x <- citizens] / number(citizens)
```

All programs are mathematical formulas.

How formulas are translated

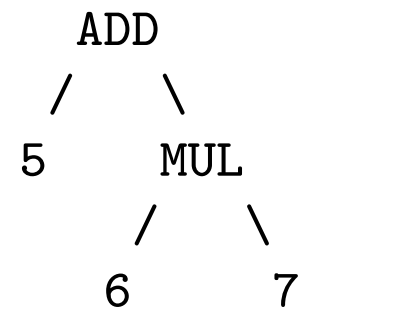
Human-readable:

```
5 + 6 * 7
```

Machine-readable:

```
0001 0000 0000 0101 0001 0000 0000 0110
0001 0000 0000 0110 0110 1000 0110 0000
```

The common structure: abstract syntax tree

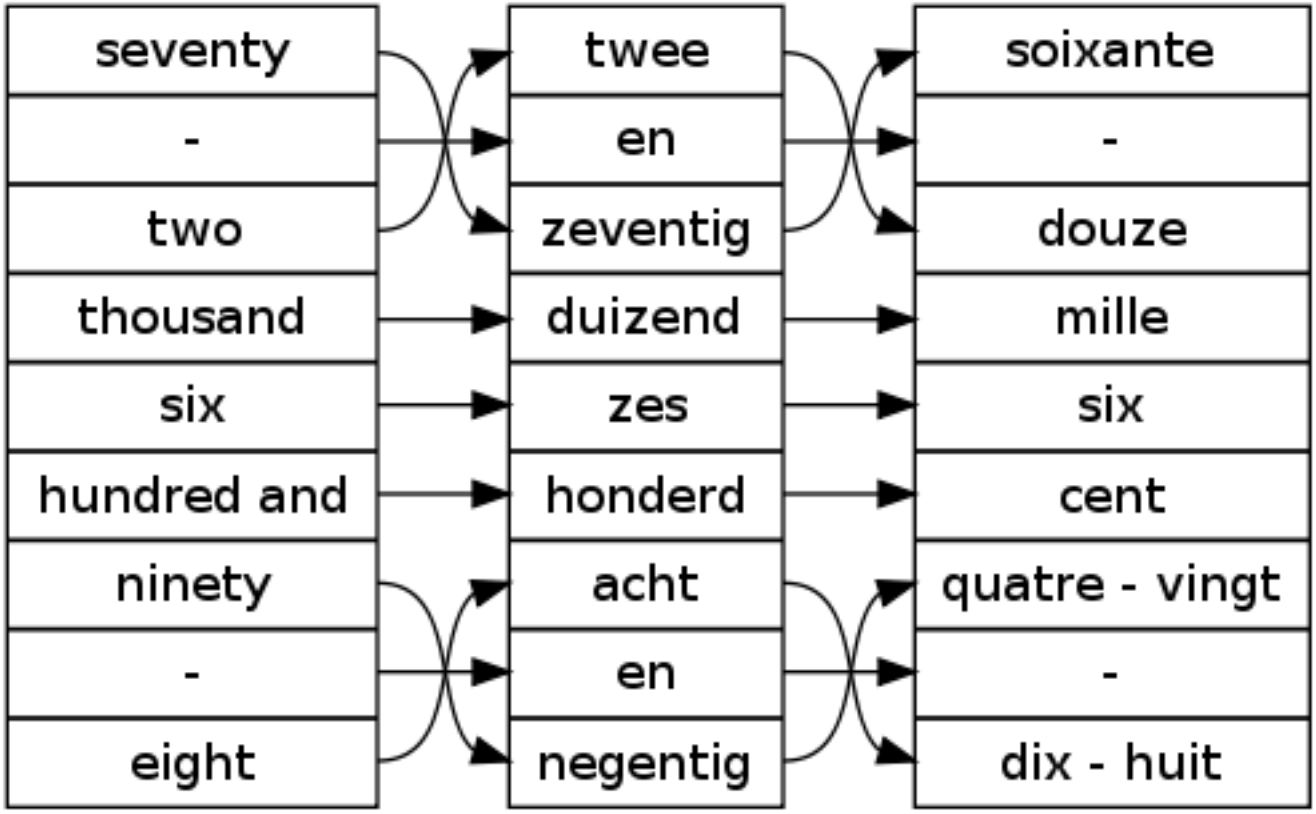


Translation in human language

seventy-two thousand six hundred and ninety-eight

tweeënzeventigduizend zeshonderdachtennegentig

soizante-douze mille six cent quatre-vingt-dix-huit


seventy
-
two
thousand
six
hundred and
ninety
-
eight
twee
en
zeventig
duizend
zes
honderd
acht
en
negentig
soixante
-
douze
mille
six
cent
quatre - vingt
-
dix - huit

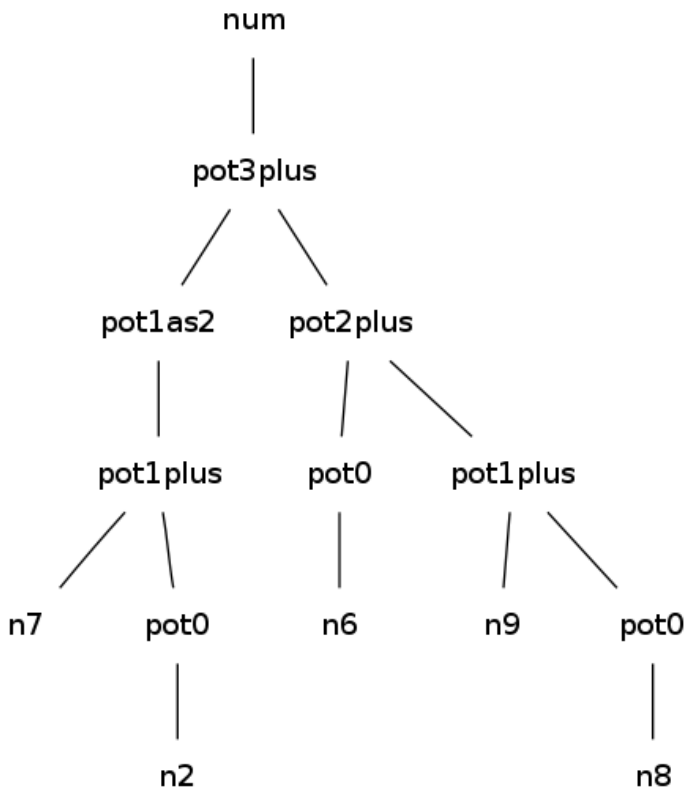
Translation word by word

From French to Dutch it would be:

zestig-twaalf duizend zes honderd vier-twentig-tien-acht

soizante-douze mille six cent quatre-vingt-dix-huit

Translation by semantic model

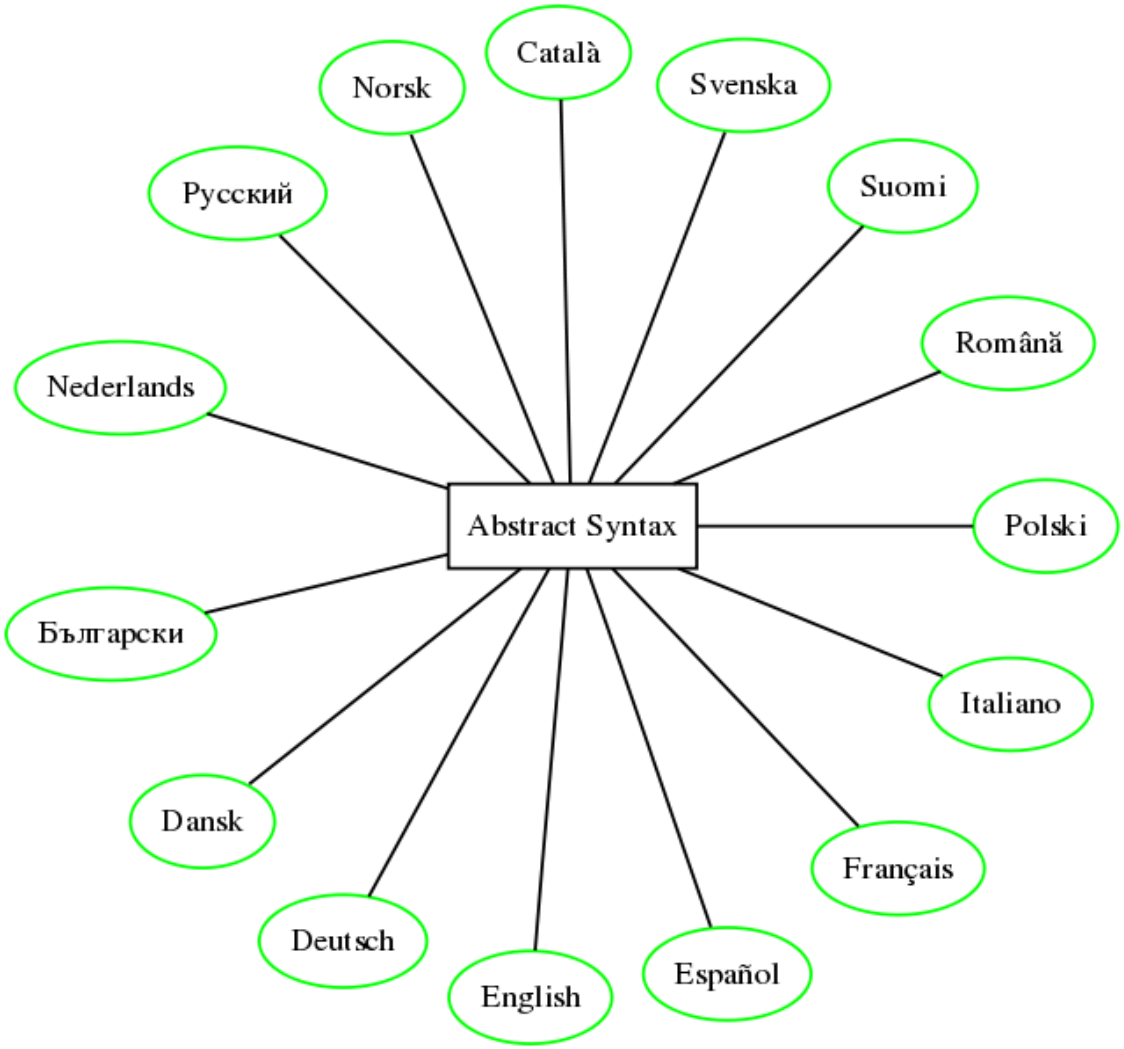


MOLTO's idea

Translations via semantic models

- made easy for new application areas

"Compiling natural language"


Català
Norsk
Svenska
Русский
Suomi
Română
Nederlands
Abstract Syntax
Polski
Български
Italiano
Dansk
Français
Deutsch
English
Español

History of the idea

Interlingua: a universal language for meaning

Need only 2*15 = 30 translators for 15 languages, not 14*15 = 210

Google uses English as interlingua for its 60 languages

But we want an exact interlingua, not a lossy one

Mathematical interlingua

From numbers to calculations to instructions

From equations to facts to stories, documents, letters

From statements to questions, wishes, requests

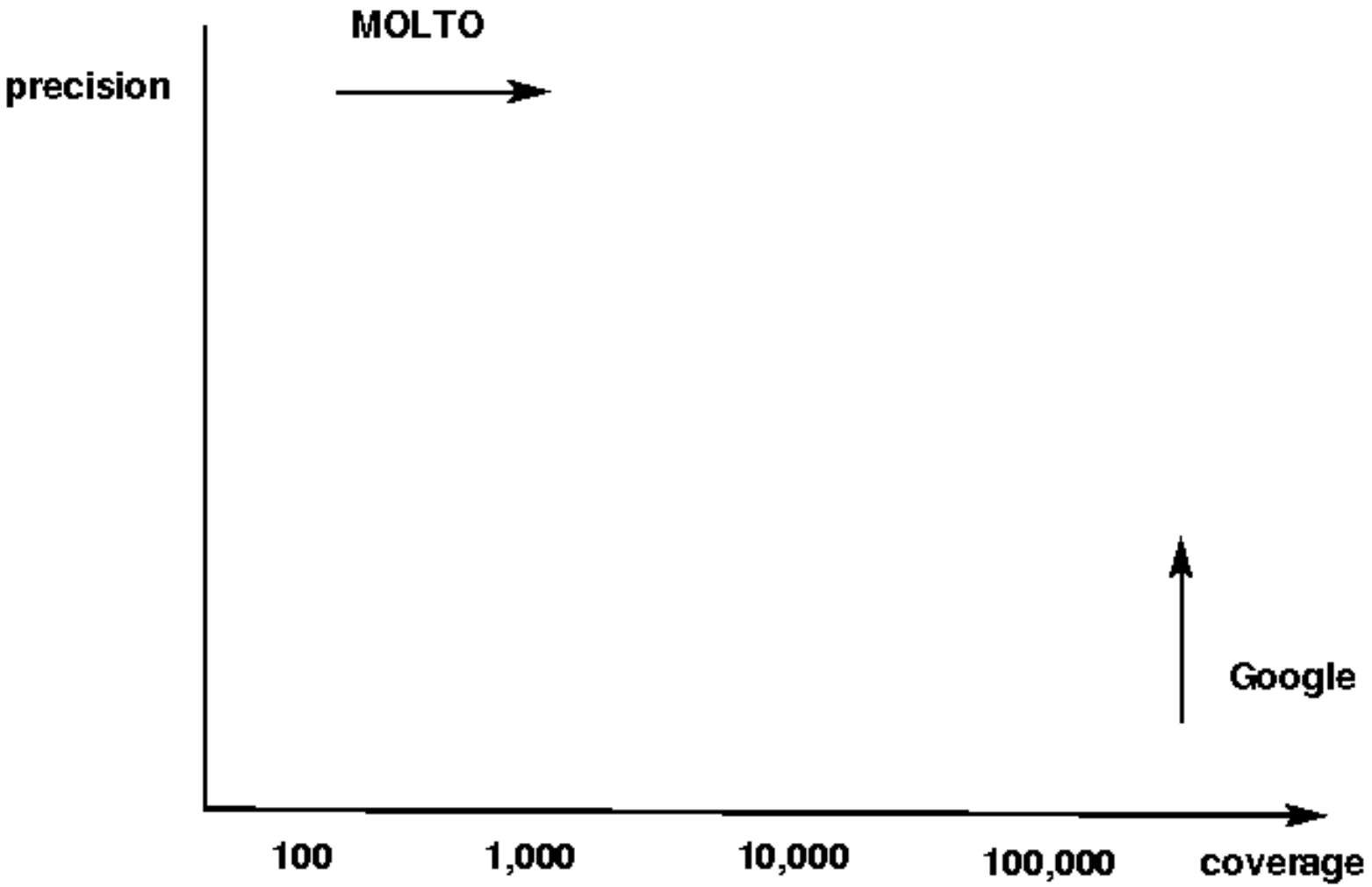
Domain vs. universal

We don't need an interlingua for everything - it would be too difficult

We build intelinguas for different **domains**

- mathematics
- museum objects
- travellers' phrases
- rule systems

The scale



The tool: GF

GF = Grammatical Framework

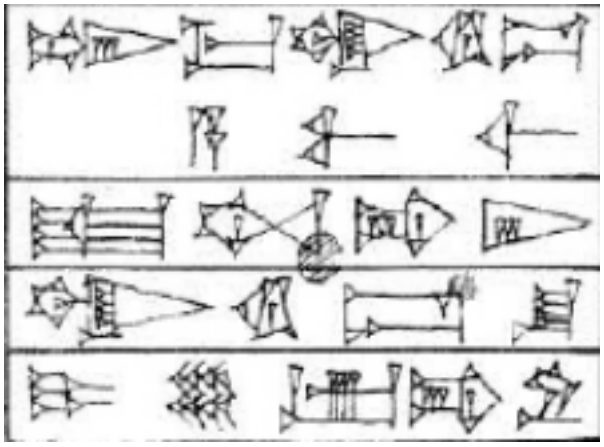
- a programming language for multilingual grammars

Interlingua: **abstract syntax**

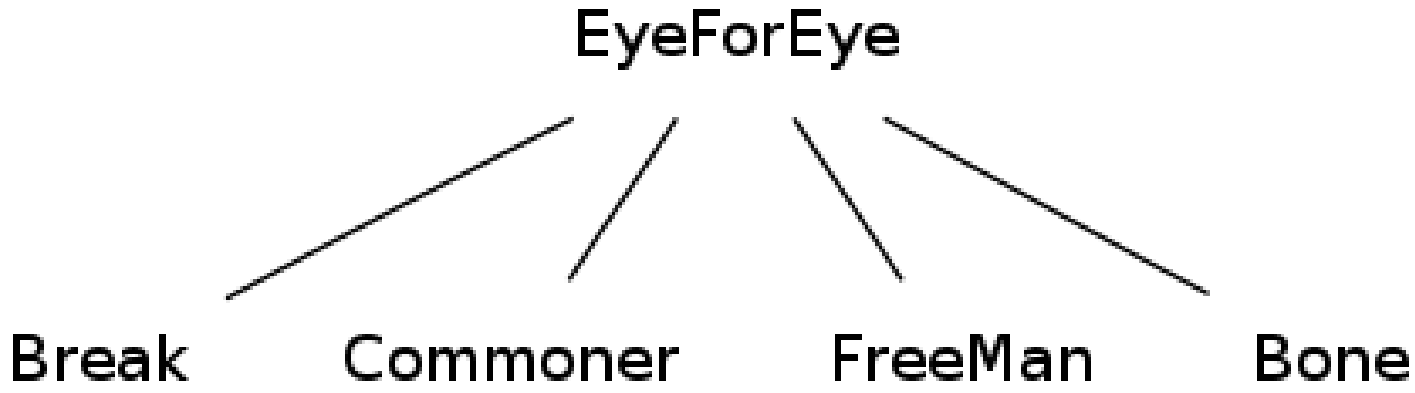
Translations to languages: **concrete syntaxes**

Example: The Code of Hammurabi

1772 BC, "an eye for an eye, a tooth for a tooth"

§ 197: 

Paragraph 197 in GF

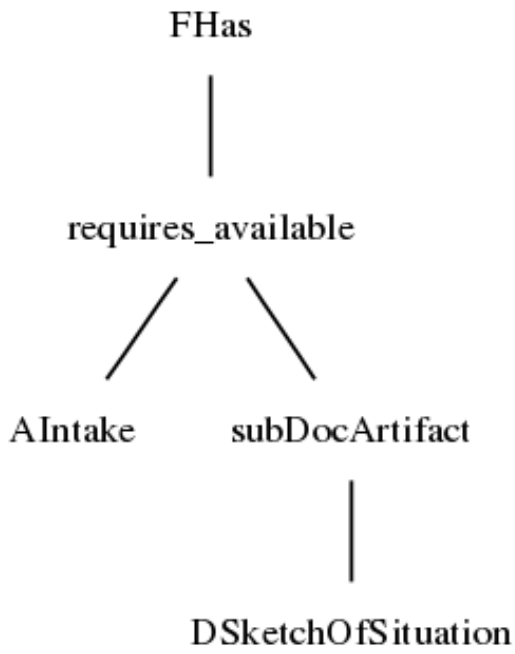


Akkadian: *šumma muškeenum esemti awiilim ištebir, esemti muškeenim išebbiruu*

English: *If a commoner breaks a bone of a free man, his bone shall be broken.*

Dutch: *Als een boer een bot van een vrije man breekt, wordt zijn bot gebroken.*

A modern variant: Be Informed rule grammar

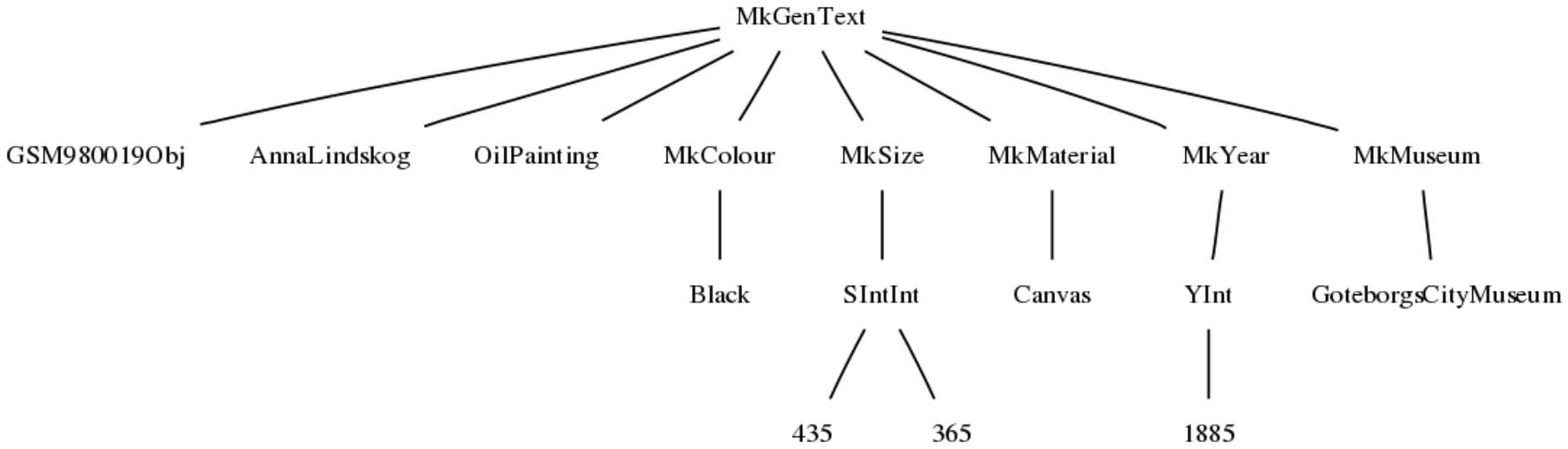


Als de inname afgerond geworden is, is een schets van de situatie beschikbaar.

If the intake has been completed, a sketch of the situation is available.

Om intagningen har blivit kompletterad, är ett skiss över situationen tillgängligt.

Another variant: museum object descriptions



- Eng: 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.

- Fin: 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.

- Fre: 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.

- Ita: 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.

- Swe: 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.

A demo: the MOLTO Phrase-book

Highly idiomatic - and mission critical!

Predictive parsing, disambiguation

18 languages

Web-based and mobile (Phrasedroid)

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

Needed for a translation system

A semantic model

- knowledge about the domain of application

Concrete syntaxes

- knowledge about languages

GF Grammar

Abstract syntax:

```
Break : Person -> Object -> Action
```

Concrete syntaxes, first version

```
Break person object = person ++ "breaks" ++ object   -- English
Break person object = person ++ "breekt" ++ object   -- Dutch
Break person object = person ++ object ++ "ištebir"  -- Akkadian
```

Grammar complications

In Dutch, we can't do with just

een boer breekt een bot

We also need

als **een boer een bot breekt**, **breekt een boer een bot**

More grammar complications

Number and person: *ik breek*, *wij breken*

Tense: *ik brak*, *ik heb gebroken*, *ik zal breken*

Passive: *wordt zijn bot gebroken*

GF code (don't look!)

```
Break person object = table {
  <Main,   Simul, tense> => person ++ breken_V ! tense ! person.agr ++ object ;
  <Invert, Simul, tense> => breken_V ! tense ! person.agr ++ person ++ object ;
  <Main,   Simul, tense> => person ++ object ++ breken_V ! tense ! person.agr ;
  <Main,   Anter, tense> => person ++ hebben_V ! tense ! person.agr ++ object ++ breken_V
  ..
  } ;
breken_V = table {
  Inf => "breken" ;
  Pres Sg P1 => "breek" ;
  Pres Sg P3 => "breekt" ;
  Past Sg P1 => "brak" ;
  PastPart   => "gebroken" ;
  ...
  } ;
```

Relieving the requirements

GF Resource Grammar Library: details of concrete syntax.

We just need to write

```
Break person object = mkCl person breken_V object
```

This gives us

- different forms of the clause (`mkCl`)
- different forms of the verb (`breken_V`)

Available languages

Currently for 25 languages

40+ contributors 2001-

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

3-6 months for a new language

To build a translation system

Abstract syntax & first language:

- domain expert, programming skills
- 3 days GF training
- days of work

Further languages:

- domain expert, language skills
- 3 hours GF training
- hours of work