MOLTO
Multilingual Online Translation

Cristina España-Bonet

TALP Research Center

Jornada sobre la Indústria de la Traducció entre Llengües Romàniques

València, September 8th, 2010

Overview

Introduction

The project

ICT-2009.2.2

Language-Based Interaction

- Majority of EU languages
- Use of existing linguistic resources

Introduction

The project

ICT-2009.2.2

Language-Based Interaction

- Majority of EU languages
- Use of existing linguistic resources

MOLTO FP7-ICT-247914

Research Personal: 390 person-month

Timeframe: 1 March 2010 - 28 February 2013

Introduction

The consortium

Academic Partners

UNIVERSITY OF GOTHENBURG

UNIVERSITY OF HELSINKI

UNIVERSITAT POLITÈCNICA DE CATALUNYA

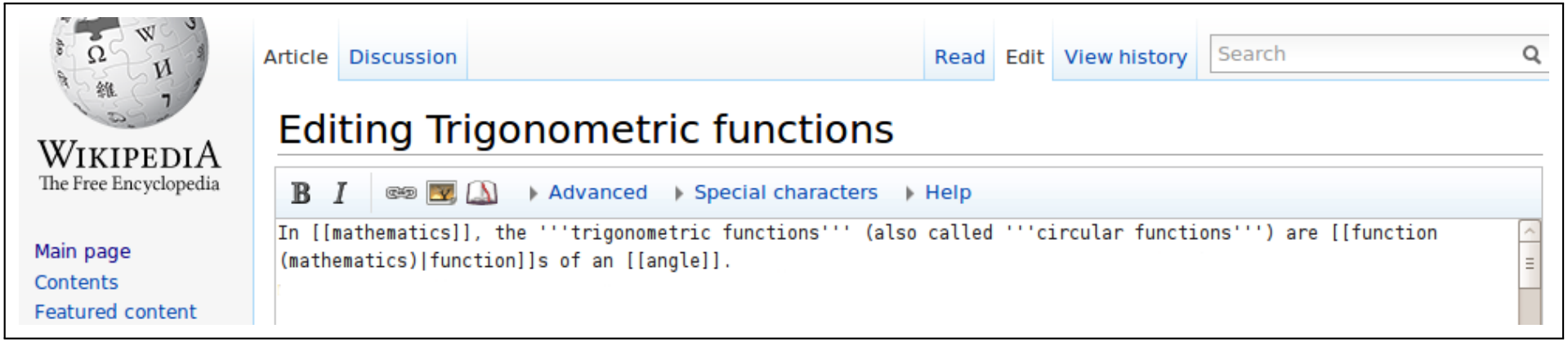
Commercial Partners

ontotext Semantic Technology Lab

MATRIXWARE INFORMATION SERVICES

Introduction

The idea



Introduction

The idea

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content

Article | Discussion | Read | Edit | View history | Search

Editing Trigonometric functions

B I | Advanced | Special characters | Help

```
In [[mathematics]], the '''trigonometric functions''' (also called '''circular functions''') are [[function
(mathematics)|function]]s of an [[angle]].
```

VIQUIPÈDIA
L'enciclopèdia lliure

Funció trigonomètrica

En matemàtiques, les **funcions trigonomètriques** són funcions d'un angle.

WIKIPEDIA
L'enciclopedia libera

Funzione trigonometrica

In matematica, le **funzioni trigonometriche** o **funzioni circolari** sono funzioni di un angolo,

WIKIPEDIA
Enciclopedia liberă

Funcție trigonometrică

În matematică, prin **funcții trigonometrice** se înțeleg niște funcții ale unui unghi oarecare.

Introduction

The idea

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content

Article | Discussion | Read | Edit | View history | Search

Editing Trigonometric functions

B I | Advanced | Special characters | Help

```
In [[mathematics]], the '''trigonometric functions''' (also called '''circular functions''') are [[function
(mathematics)|function]]s of an [[angle]]. Trigonometric functions are important in the study of triangles and
modeling periodic phenomena, among many other applications.
```

VIQUIPÈDIA
L'enciclopèdia lliure

Funció trigonomètrica

En matemàtiques, les **funcions trigonomètriques** són funcions d'un angle. Són la base per l'estudi de la trigonometria, els triangles i per la modelització dels fenòmens periòdics, entre moltes altres aplicacions

WIKIPEDIA
L'enciclopedia libera

Funzione trigonometrica

In matematica, le **funzioni trigonometriche** o **funzioni circolari** sono funzioni di un angolo. Esse sono importanti nello studio dei triangoli e nella modellizzazione dei fenomeni periodici, oltre a un gran numero di altre applicazioni.

WIKIPEDIA
Enciclopedia liberă

Funcție trigonometrică

În matematică, prin **funcții trigonometrice** se înțeleg niște funcții ale unui unghi oarecare. Ele se folosesc la studierea triunghiurilor si reprezentarea unor fenomene periodice, printre multe altel

Introduction

The idea

But, nowadays, kind of utopy.

Not viable at Wikipedia level –*large coverage*–
but one can think on a restricted language.

Introduction

The idea

Fridge Magnet Demo

Not viable at Wikipedia level –*large coverage*–
but one can think on a restricted language.

Introduction

The idea in a demo

Fridge Magnet Demo

http://www.grammaticalframework.org:41296/fridge/

Introduction

The idea in a demo

Fridge Magnet Demo

Introduction

System description by comparison

	GOOGLE-like	MOLTO-like
Target	consumers	translators
Input	unpredictable	predictable
Coverage	unlimited	limited
Quality	browsing	publishing

Introduction

MOLTO's goals & challenges

MOLTO's mission is to develop a set of tools for translating texts between multiple languages in real time with high quality.

Introduction

Goal: Better tools

System developers' tools

- An Integrated Development Environment (IDE)
- An example-based grammar writing component

Introduction

Goal: Better tools

System developers' tools

- An Integrated Development Environment (IDE)
- An example-based grammar writing component

Translators'/Authoring tools

- Syntax editors and word predictors as plug-ins to
  - web browsers
  - text editors
  - professional translators' tools

Introduction

Challenge: Scale up production of domain interpreters

From 100's of words to 1000's of words

Introduction

Challenge: Scale up production of domain interpreters

From 100's of words to 1000's of words

From GF experts to domain experts & translators

Introduction

Challenge: Scale up production of domain interpreters

From 100's of words to 1000's of words

From GF experts to domain experts & translators

From months to days

Introduction

Challenge: Scale up production of domain interpreters

From 100's of words to 1000's of words

From GF experts to domain experts & translators

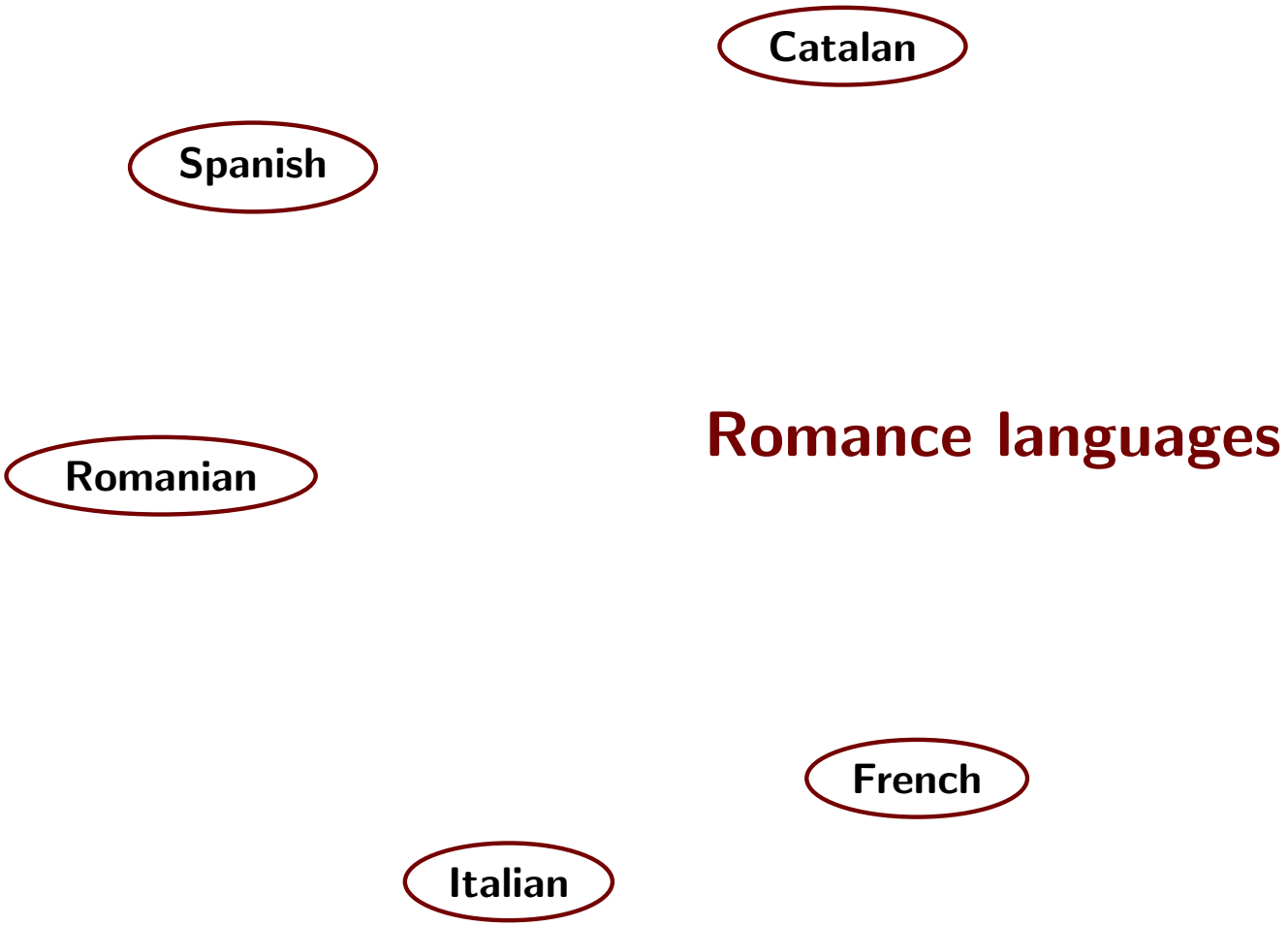
From months to days

From hand-crafting a grammar to translating a set of examples

Romance languages

Introduction

Languages



Introduction

Languages

Specific domains of application

- Description of museum items
- Mathematical problems
- Patents in biomedical and pharmaceutical domain

Specific domains of application

- Description of museum items
- Mathematical problems
- Patents in biomedical and pharmaceutical domain

Those are specific selected domains, but it is easy to think of other potential applications.

Introduction

Potential applications

Tourist phrasebooks

Introduction

Potential applications

Tourist phrasebooks

E-commerce sites

Introduction

Potential applications

Tourist phrasebooks

E-commerce sites

Medical treatment recommendations

Introduction

Potential applications

Tourist phrasebooks

Manuals

E-commerce sites

Medical treatment recommendations

Introduction

Potential applications

Tourist phrasebooks

Manuals

E-commerce sites

Wikipedia articles

Medical treatment recommendations

Multilingual translation system

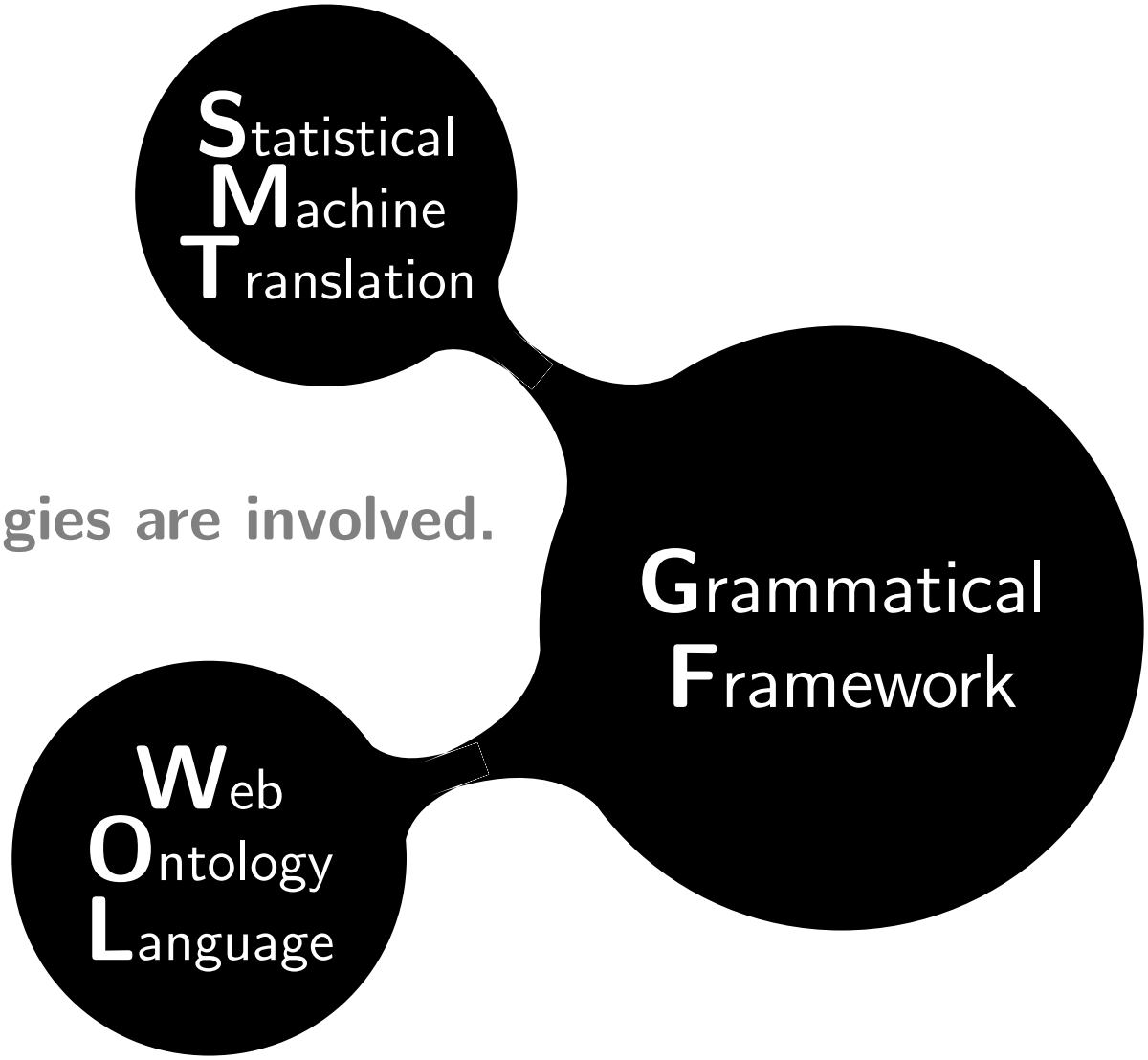
System engine

Three technologies are involved.

Multilingual translation system

System engine

Three technologies are involved.



What is GF?

- A **grammar formalism**: a notation for writing grammars.
- A **functional programming language**.

Multilingual translation system

The core: Grammatical Framework

What is GF?

- A **grammar formalism**: a notation for writing grammars.
- A **functional programming language**.

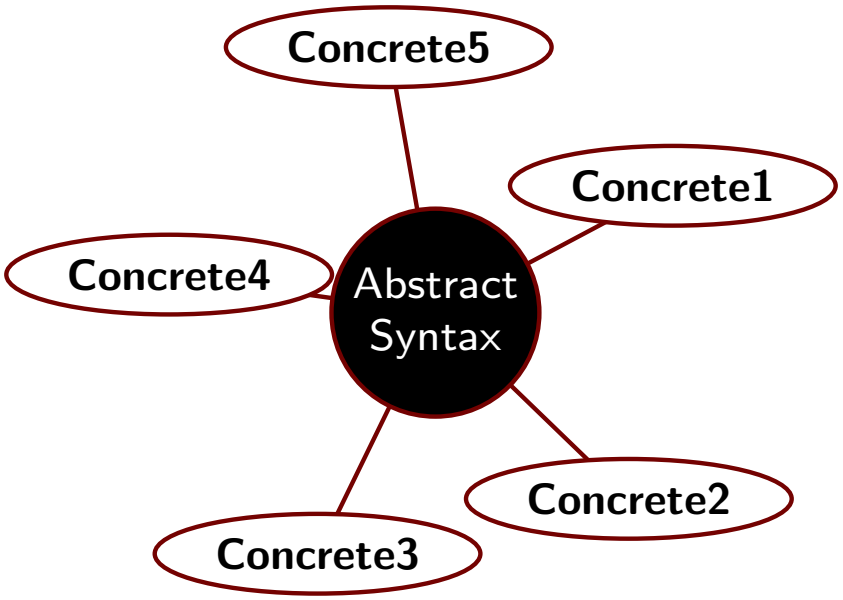
What is a multilingual grammar?

- A definition of a **parsing** and **generation** operations.
- **Concrete syntaxes** for many languages related by a common **abstract syntax**.

Multilingual translation system

Abstract and Concrete syntaxes

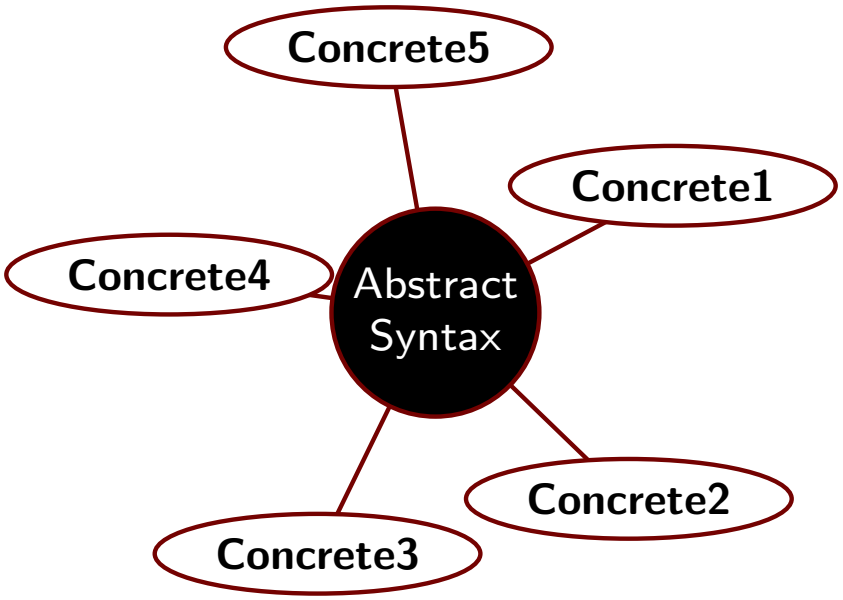
The abstract syntax acts as a **domain-specific interlingua**.



Multilingual translation system

Abstract and Concrete syntaxes

The abstract syntax acts as a **domain-specific interlingua**.



Defines not only a linguistic structure but a semantic model for translation with:

- fixed word senses
- proper idioms

Multilingual translation system

Translation with GF

Abstract Syntax

```
Nat : Set
Odd : Exp -> Prop
Gt : Exp -> Exp -> Prop
Sum : Exp -> Exp
```

Concrete Syntax (ENG)

```
Nat = "number"
Odd x = "x is odd"
Gt x y = "x is greater than y"
Sum x = "the sum of x"
```

Concrete Syntax (CAT)

```
Nat = "número"
Odd x = "x és senar"
Gt x y = "x és més gran que y"
Sum x = "la suma de x"
```

Multilingual translation system

Translation with GF

Abstract Syntax

```
Nat : Set
Odd : Exp -> Prop
Gt : Exp -> Exp -> Prop
Sum : Exp -> Exp
```

parsing

Concrete Syntax (ENG)

```
Nat = "number"
Odd x = "x is odd"
Gt x y = "x is greater than y"
Sum x = "the sum of x"
```

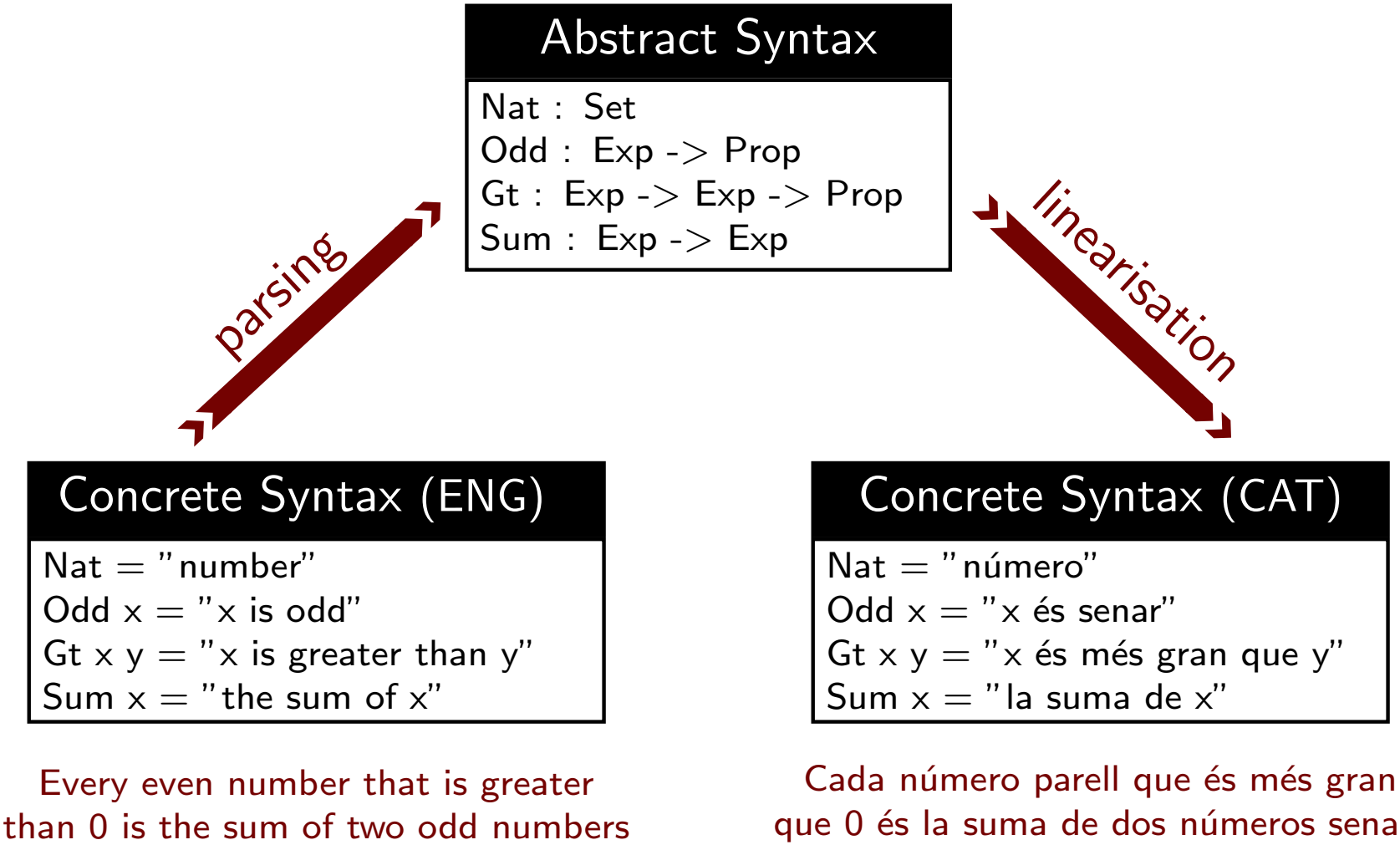
Concrete Syntax (CAT)

```
Nat = "número"
Odd x = "x és senar"
Gt x y = "x és més gran que y"
Sum x = "la suma de x"
```

Every even number that is greater than 0 is the sum of two odd numbers

Multilingual translation system

Translation with GF

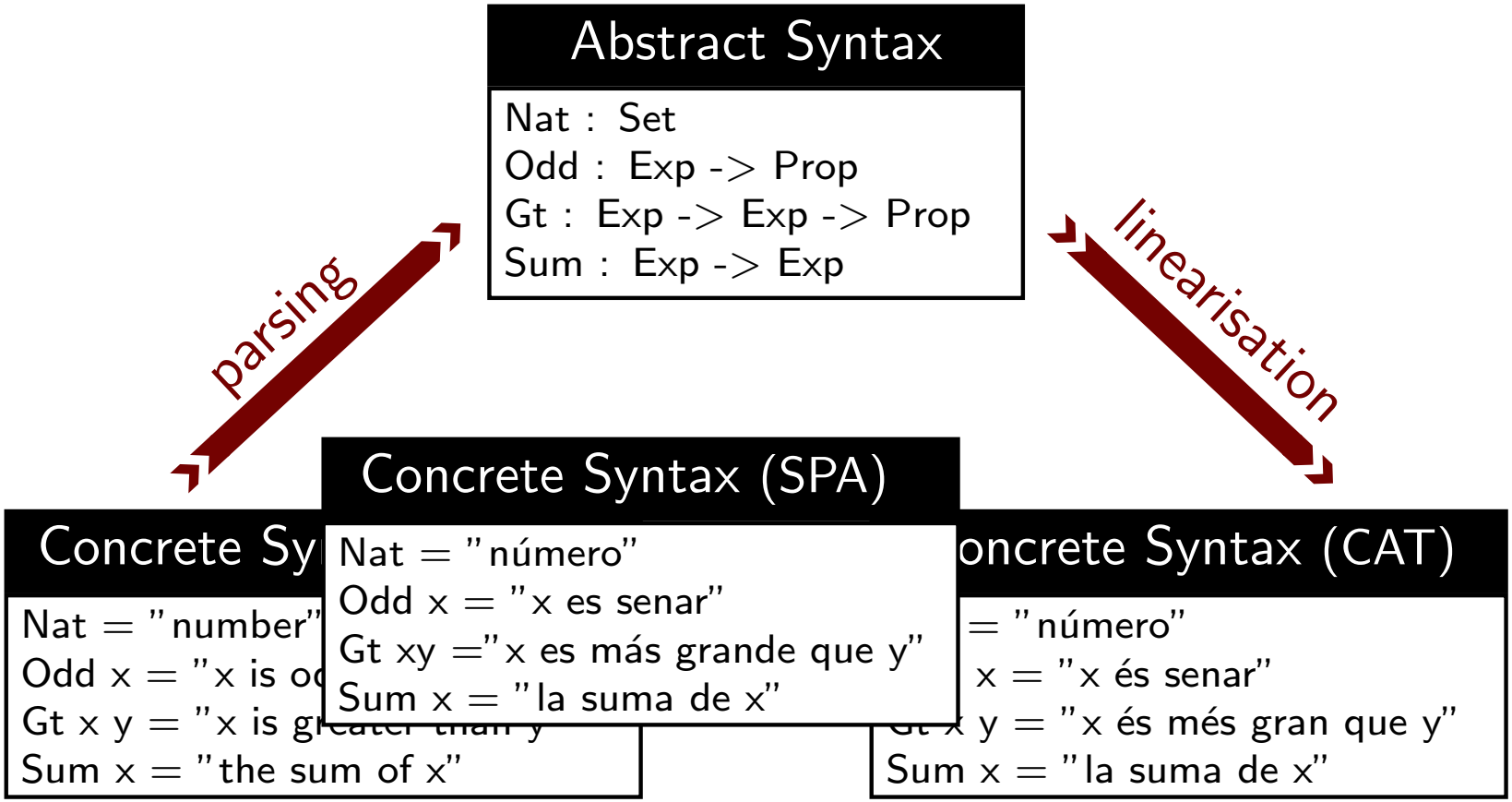


Every even number that is greater than 0 is the sum of two odd numbers

Cada número parell que és més gran que 0 és la suma de dos números senars

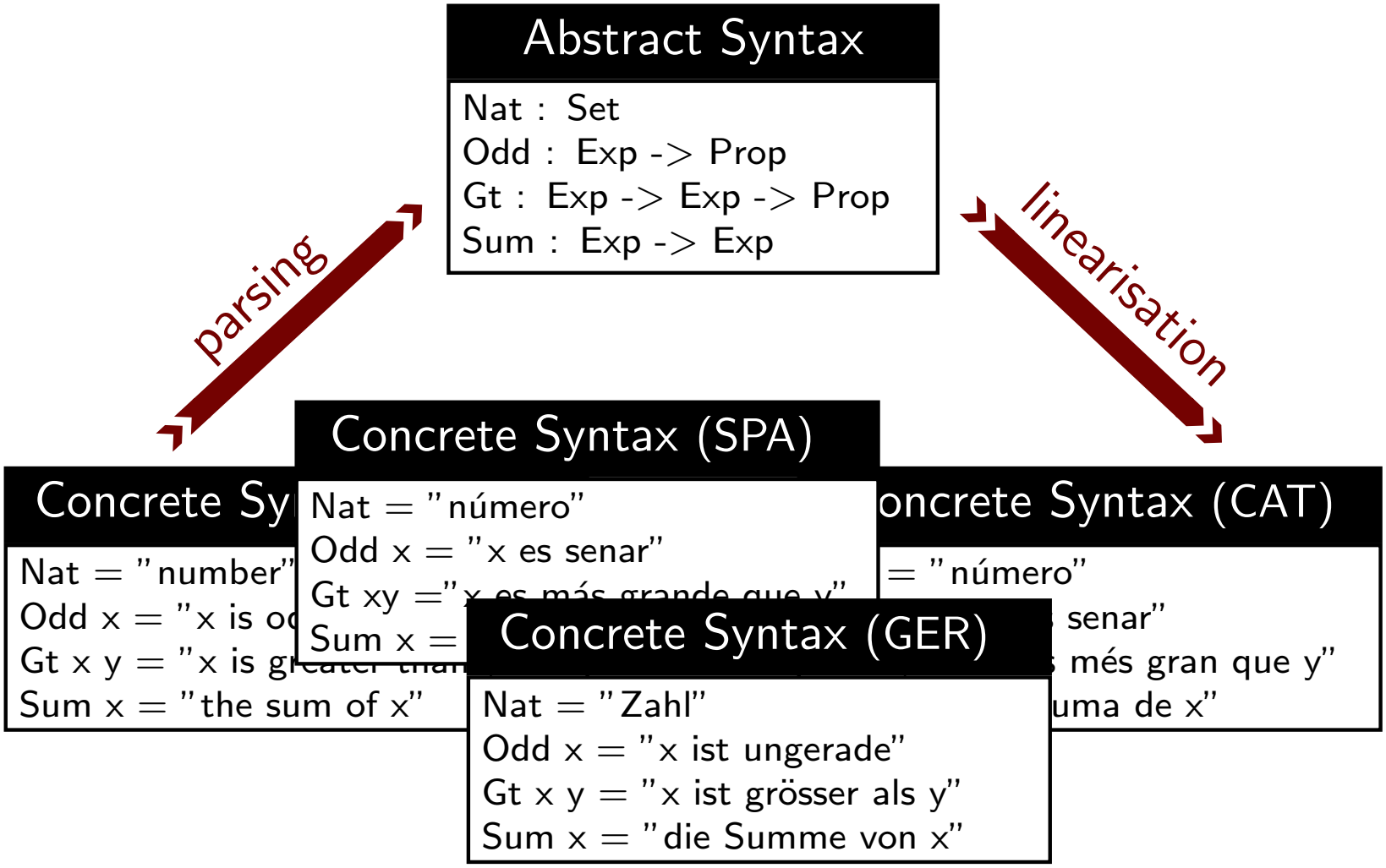
Multilingual translation system

Translation with GF



Multilingual translation system

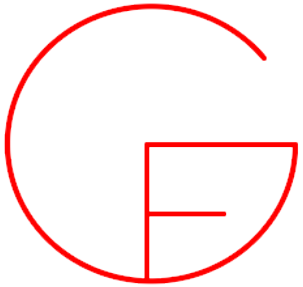
Translation with GF



Multilingual translation system

OWL-GF interoperability

OWL Express formal meaning representations (semantics) of data and content

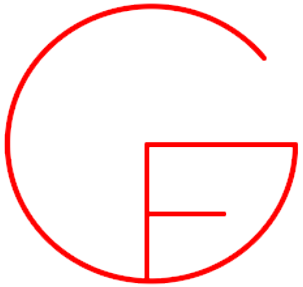


GF Renders those ontologies into natural language (and viceversa)

Multilingual translation system

OWL-GF interoperability

OWL Express formal meaning representations (semantics) of data and content



GF Renders those ontologies into natural language (and viceversa)

The mapping translates OWL's classes to GF's categories and OWL properties to GF's functions that return propositions.

Multilingual translation system

OWL-GF interoperability

Research topic

(Semi-)automatically construct GF's abstract syntax from OWL ontologies.

Multilingual translation system

OWL-GF interoperability

Research topic

(Semi-)automatically construct GF's abstract syntax from OWL ontologies.

Also, a **Research topic** not strictly related to translation:

Information retrieval from ontologies in multiple natural languages.

Multilingual translation system

OWL-GF interoperability

Research topic

(Semi-)automatically construct GF's abstract syntax from OWL ontologies.

Also, a **Research topic** not strictly related to translation:

Information retrieval from ontologies in multiple natural languages.

! Museum case data are already in OWL.

Multilingual translation system

Robustness by statistics

Research topic

Develop hybrid MT methods that complete the GF-based ones by extending their coverage in unconstrained text translation.

Multilingual translation system

Robustness by statistics

Research topic

Develop hybrid MT methods that complete the GF-based ones by extending their coverage in unconstrained text translation.

At last, **our task!**
Hybridisation is closely related to the work in the Spanish project OPEN-MT2.

Multilingual translation system

Robustness by statistics

Research topic

Develop hybrid MT methods that complete the GF-based ones by extending their coverage in unconstrained text translation.

At last, **our task!**
Hybridisation is closely related to the work in the Spanish project OPEN-MT2.

! The patents case is a quasi-open domain suitable for it.

Multilingual translation system

Statistics: methodology

1. Probabilistic extension of a GF **domain grammar**.

Multilingual translation system

Statistics: methodology

1. Probabilistic extension of a GF **domain grammar**.

2. Adapt base SMT systems to the **patents domain**.

Multilingual translation system

Statistics: methodology

1. Probabilistic extension of a GF **domain grammar**.

2. Adapt base SMT systems to the **patents domain**.

3. Develop and test **hybrid GF-SMT** translation methods.

Multilingual translation system

Training and adaptation of a SMT system

2.1 Base SMT system builded with out-of-domain corpora.

!	Out-of-domain

Multilingual translation system

Training and adaptation of a SMT system

2.1 Base SMT system builded with out-of-domain corpora.

! Out-of-domain

2.2 Use of small patents parallel corpora for adaptation.

! Maybe too small

Multilingual translation system

Training and adaptation of a SMT system

2.1 Base SMT system builded with out-of-domain corpora.

! Out-of-domain

2.2 Use of small patents parallel corpora for adaptation.

! Maybe too small

2.3 Explore the usage of synthetic corpora generated by GF.

! Ongoing work

Multilingual translation system

Training and adaptation of a SMT system

2.3 Explore the usage of synthetic corpora generated by GF.

Use domain grammar to generate **correct** translations in the patent domain, which serve as more training examples for SMT

Multilingual translation system

Training and adaptation of a SMT system

2.3 Explore the usage of synthetic corpora generated by GF.

Use domain grammar to generate **correct** translations in the patent domain, which serve as more training examples for SMT

But, some **requirements** are needed:

- Translations have to be varied.
- The balance between in-domain and general training corpora has to be properly set.

Multilingual translation system

Hybrid GF-SMT system

GF translation is high quality, thus there is no need of SMT when GF parses the input and generates a complete translation.

Multilingual translation system

Hybrid GF-SMT system

GF translation is high quality, thus there is no need of SMT when GF parses the input and generates a complete translation.

3.1 Baseline combination

Fall-back / back-off / cascaded approach, i.e., use pure SMT whenever GF fails to produce a translation of the source sentence or a source phrase.

3.2 Hard integration

Fix translation phrases produced by the partial GF analyses in a probabilistic decoding.

> **!** It constraints the search space with **secure translations** of some phrases, but GF predictions do not really interact with the SMT model.

3.3 Soft integration

GF scored partial output as **new features** in SMT decoding.

$$\log P(e|f) \sim \lambda_{lm} \log P(e) + \lambda_g \log P(f|e) + \lambda_d \log P(e|f)$$
$$+\lambda_{di} \log P_{di}(e, f) + \lambda_w \log w(e) + \lambda_{\mathbf{GF}} \mathbf{log\, P_{GF}(e|f)}$$

Multilingual translation system

Hybrid GF-SMT system

3.3 Soft integration

GF scored partial output as **new features** in SMT decoding.

$$\log P(e|f) \sim \lambda_{lm} \log P(e) + \lambda_g \log P(f|e) + \lambda_d \log P(e|f)$$
$$+\lambda_{di} \log P_{di}(e, f) + \lambda_w \log w(e) + \boldsymbol{\lambda_{GF} \log P_{GF}(e|f)}$$

But, some **requirements** are needed:

- GF predictions have to be probabilistic.
- Phrase pairs without prediction must be complemented.

Final notes

In summary

Three innovations.

Find useful ways of combining GF with statistical translation methods

Scale up grammar-based interlingual translation with GF from a set of successful experiments to a productive tool

Link GF grammars with web ontology standards and exploit ontologies in translation

MOLTO

Final notes

Results

Three families of results.

- A tool **for creating** domain-specific translation systems.
- A set of tools **for translators** and the general public to translate documents.
- Three extensive **case studies** (mathematical exercises, biomedical patents, museum objects).

Final notes

Results

MOLTO software will be released as open-source software under GNU LGPL license, except for the patent translator which will be exploited by one of the partner companies.

Final notes

Thanks!

GRÀCIES!

More about MOLTO at

`http://www.molto-project.eu/`

MOLTO

MOLTO
Multilingual Online Translation

Cristina España-Bonet

TALP Research Center

Jornada sobre la Indústria de la Traducció entre Llengües Romàniques

València, September 8th, 2010

Partners

People & Commitments

UNIVERSITY OF GOTHENBURG

Aarne Rantra *et al.*

- Grammar development tools
- Museum case

UNIVERSITY OF HELSINKI

Lauri Carlson *et al.*

- Translation tools
- Evaluation

Partners

People & Commitments

Borislav Popov *et al.*

- Ontology tools
- Web interfaces

Neil Tipper *et al.*

- Patents data

Partners

People & Commitments

Jordi Saludes *et al.*

- Mathematic problems case

Lluís Màrquez *et al.*

- Statistical methods

Partners

People & Commitments

Jordi Saludes *et al.*

- Mathematic problems case

Lluís Màrquez *et al.*

- Statistical methods

Cases of study in depth

Mathematical exercises

Enhance the multilingual mathematical GF library by adding a grammar for commanding a Computer Algebra System by natural language imperative sentences. Using ontologies to describe word problems, the system will be able to carry out a dialog with the student solving the problem.

Cases of study in depth

Museum object descriptions

Build an ontology-based multilingual grammar starting from a CRM ontology for artifacts at Gothenburg City Museum.
The prototype will be tested for cross-language retrieval and representation, and for automatic generation of Wikipedia-like articles for museum artifacts in 5 languages.

Cases of study in depth

Biomedical and pharmaceutical patents

Create a commercially viable prototype of a system for multilingual translation and cross-language retrieval of patent abstracts and claims in at least 3 languages.