Machine Translation for Producers of Information

Talk at Shanghai University of Finance and Economics
6 November 2014

Aarne Ranta
University of Gothenburg & Digital Grammars AB

digital Grammars
Language technology to rely on.

http://www.digitalgrammars.com

Machine translation

Translation by computer

E.g. English to Chinese

Fully automatic or interactive

Example: Google translate

Producers of information

Those who publish the original and its translations

E.g. e-commerce sites, international organizations, authorities

Responsibility for the information

Have to get the message through!

Consumers of information

Readers of the documents

E.g. customers, citizens

No responsibility, but rely on the producer

Want to know what the document says!

Main-stream machine translation

Made for consumers

Browsing quality

Can be wrong

No-one is responsible

E.g. Google translate, Bing, Apertium

A possible example

A French e-commerce site says

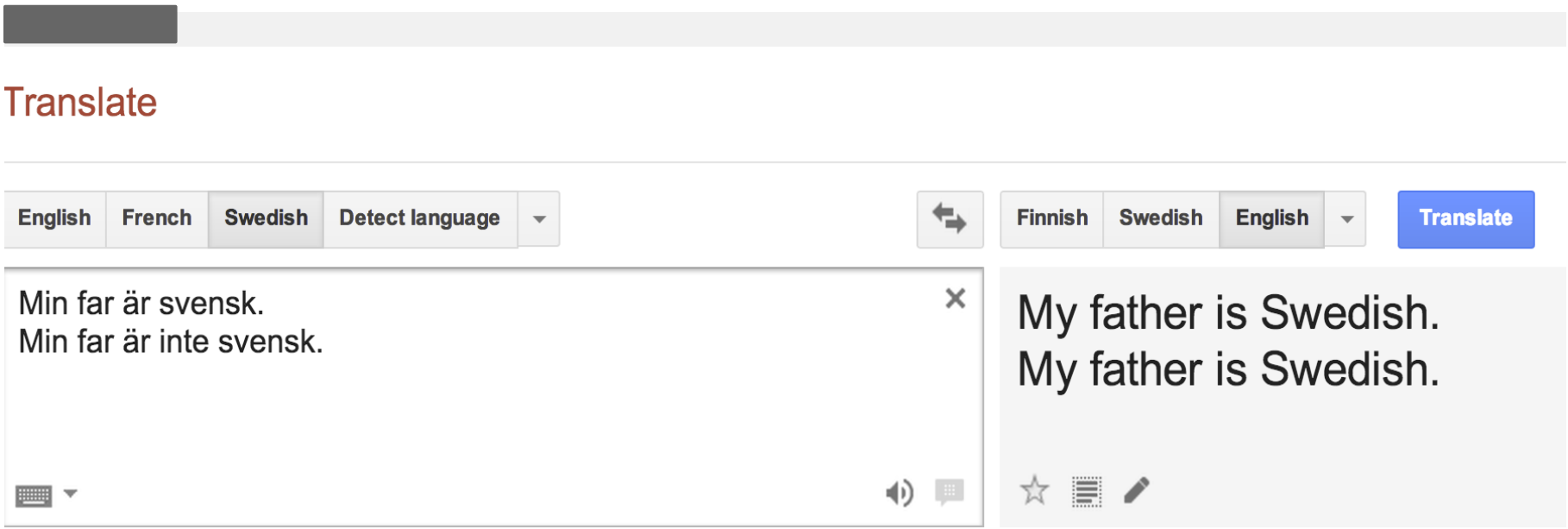
prix 99 euros

This may get translated

售价99元

Does the customer have the right to get this price?

A real example



One right, one wrong - which is which?

A real example

Min far är svensk.	我的父亲是瑞典。
Min far är inte svensk.	我的父亲是瑞典。

One right, one wrong - which is which?

What producers need

Globalization

Localization

Time to market

→ reliable, multilingual translation

Should be cheap

Should be fast

State of the art for producers

Human translation

- slow
- expensive

Localization databases

- rigid
- difficult for many languages

Typical localization problems

You have one new messages.

You have one new message.
You have five new messages.

Vous avez un nouveau message.
Vous avez cinq nouveaux messages.

The problem

It is not enough to fill templates.

One needs a **grammar** to tell how the words are changed.

Depends on language of course

你 有 一 个 新 信 息

你 有 五 个 新 信 息

Depends on language - however,...

你 有 一 个 新 信 息

你 有 五 个 新 信 息

你 有 一 只 黑 猫

你 有 五 只 黑 猫

What we want to build

Reliable machine translation

- correct grammar
- correct meaning

Not necessarily **applicable** to **everything.**

But should be **adaptable** to **anything.**

Coverage vs. precision

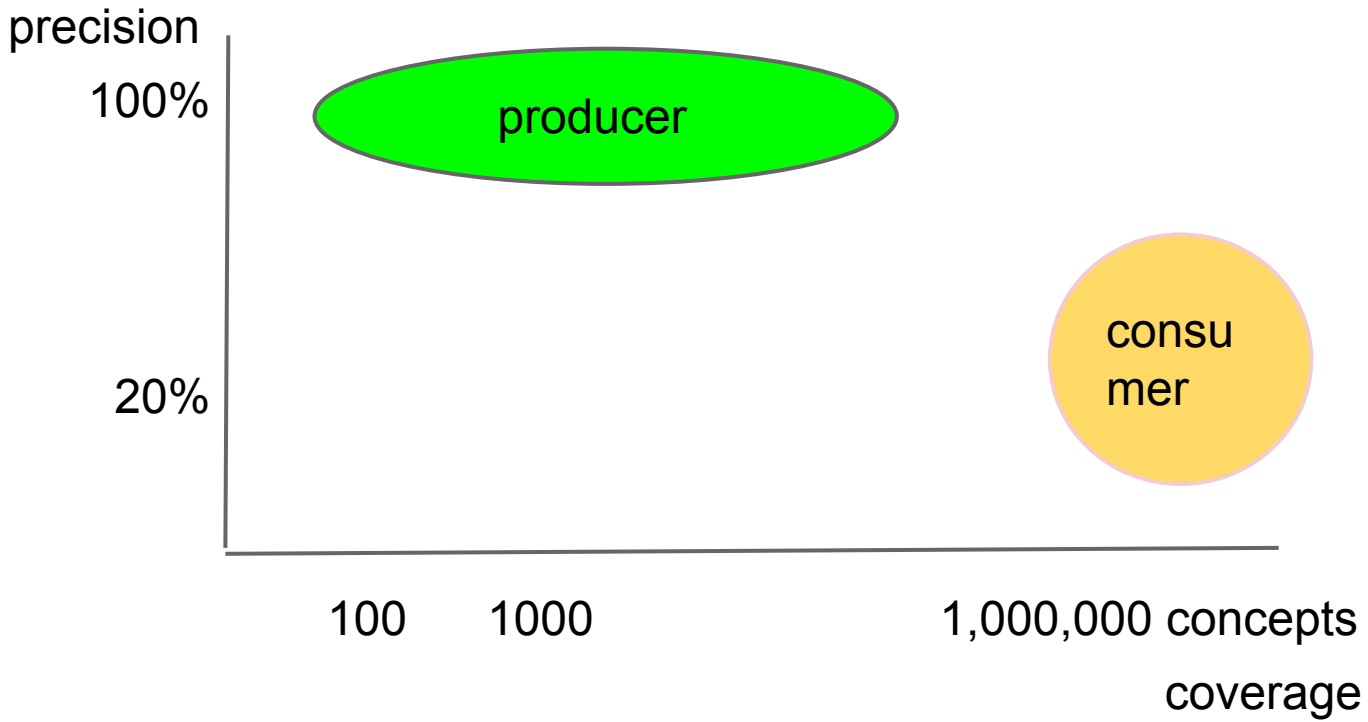
Consumers need **coverage**

- you can translate any text and get something

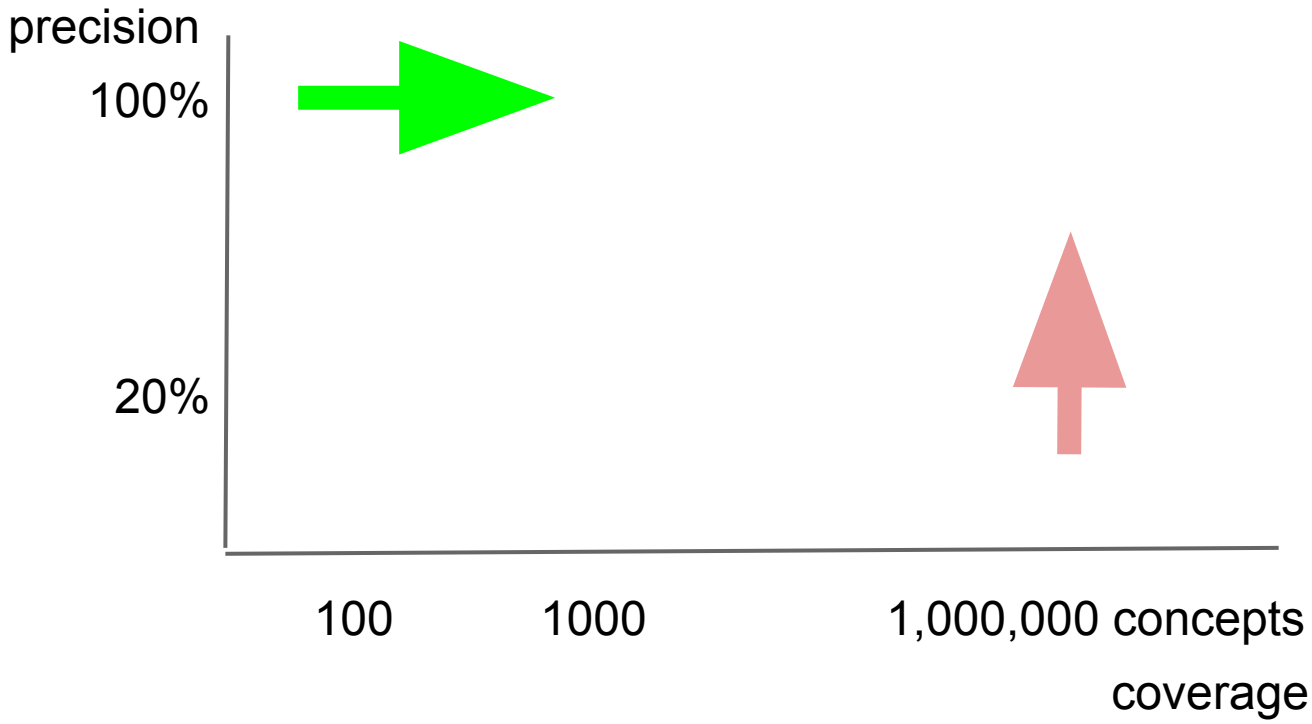
Producers need **precision**

- you only need to translate your own texts, but you have to get them right

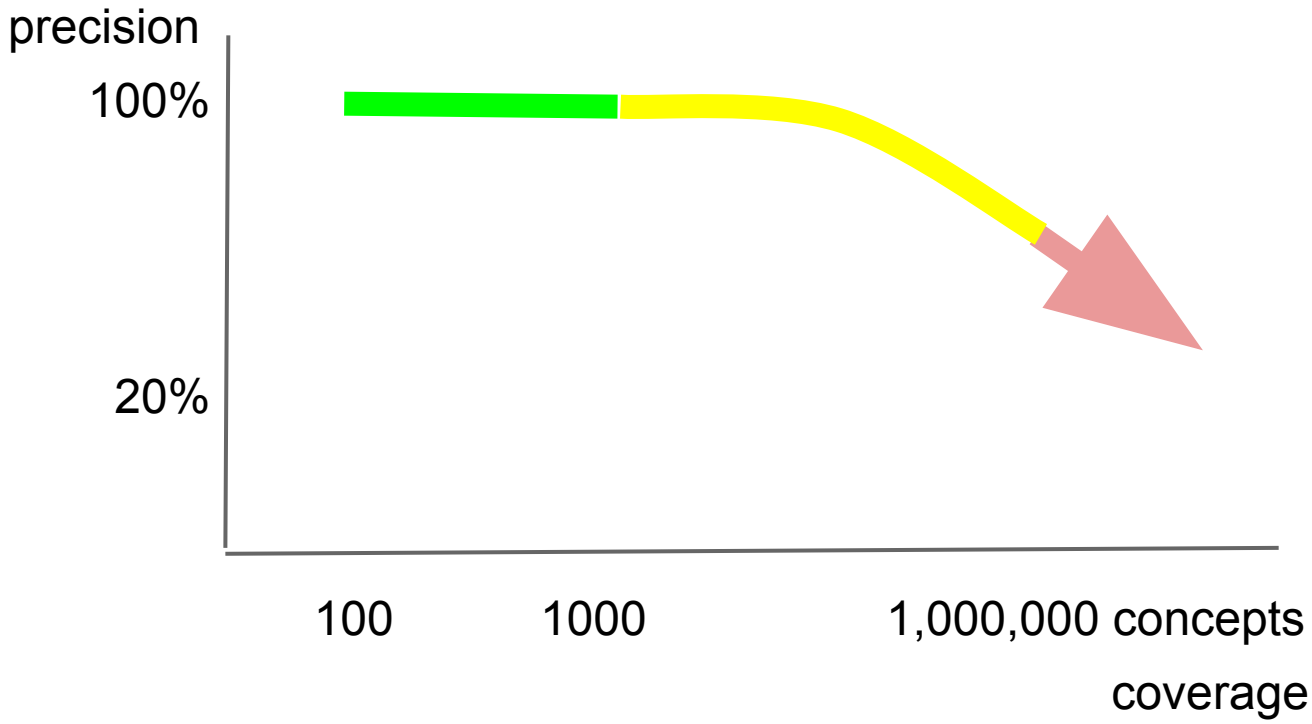
Orthogonal concepts



Two ways of developing a system



The best scenario?



This is what we want!

This is what we want

We want machine translation that

- delivers publication quality in areas where reasonable effort is invested
- degrades gracefully to browsing quality in other areas
- shows a clear distinction between these

How we do this

We use **grammars** and **type-theoretical interlinguas** implemented in **GF, Grammatical Framework**.

Started at Xerox Research in 1998, GF is a tool for **highly multilingual, precision-oriented** translation.

Latest developments have scaled it up in **productivity** and also **coverage**.

We believe GF is mature for commercial prime time!

5 March 2014 -

REMU

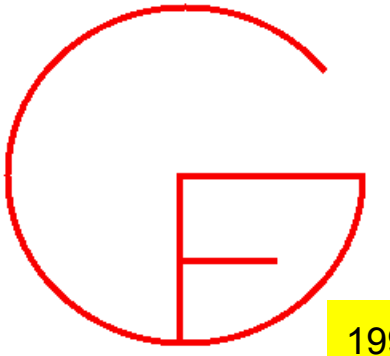
VR 2013 - 2017

MOLTO

EU 2010 - 2013

CLT

2009 -



1998 -

Demo 1: MOLTO phrasebook

Source: **controlled language input**

Always green

Based on **domain semantics**

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

Demo 2: text from data

Source: **formalized data**

Always green

Based on **ontology** (semantic web)

http://museum.ontotext.com

Demo 3: wide-coverage translation

Source: **text** in any language

Can be green, yellow, or red.

Based on **semantics, grammar, or chunks.**

http://cloud.grammaticalframework.org/wc.html

Example

How far is the airport from the hotel?

从 旅 馆 到 机 场 有 多 远?

meaning

The vice dean kicked the bucket.

副 院 长 踢 了 桶.

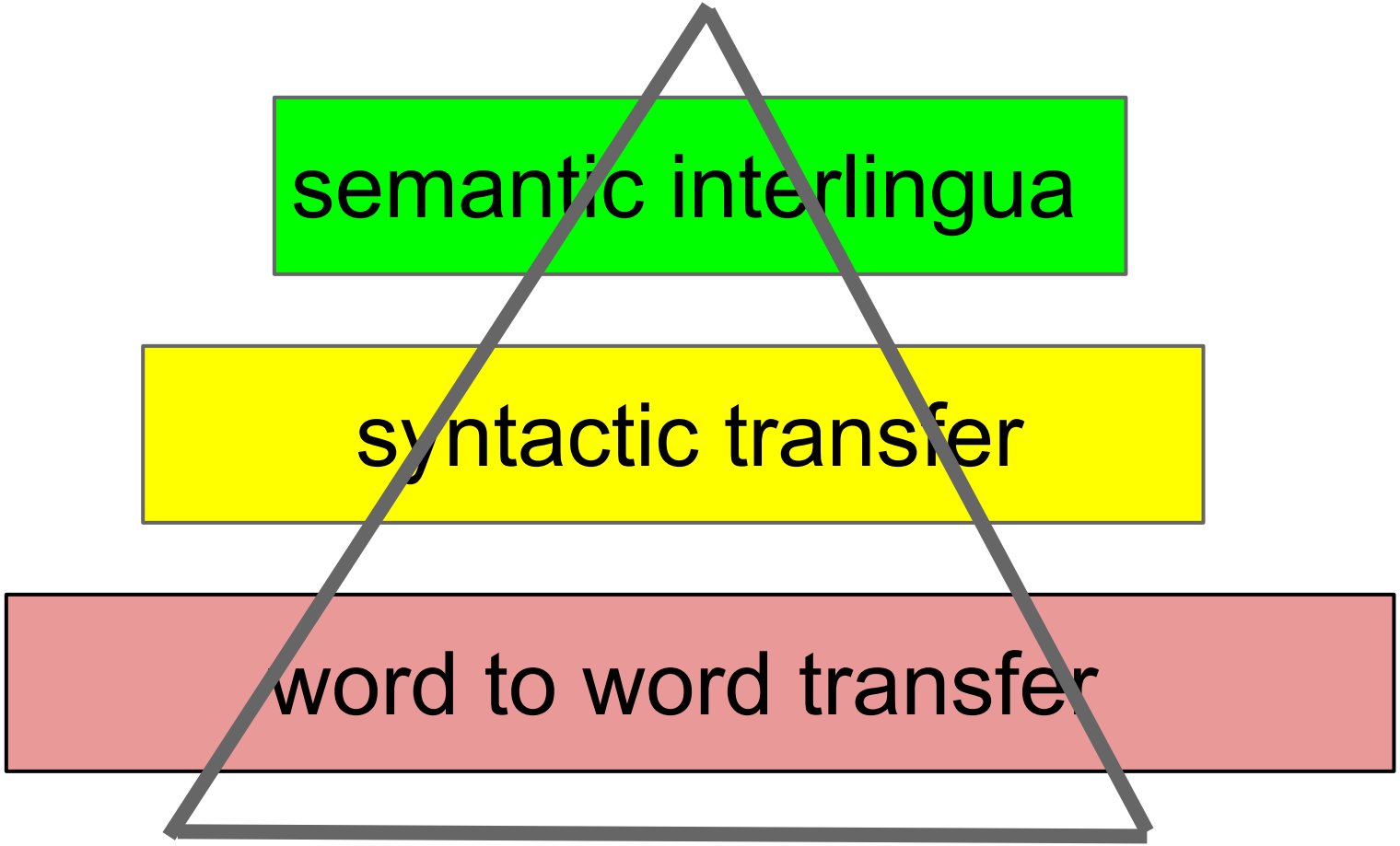
syntax

Little boy eat big snake.

小 男 孩 吃 大 蛇.

chunks

The Vauquois triangle



Demo 4: mobile translation app

Source: **text or speech** in any language

Can be green, yellow, or red.

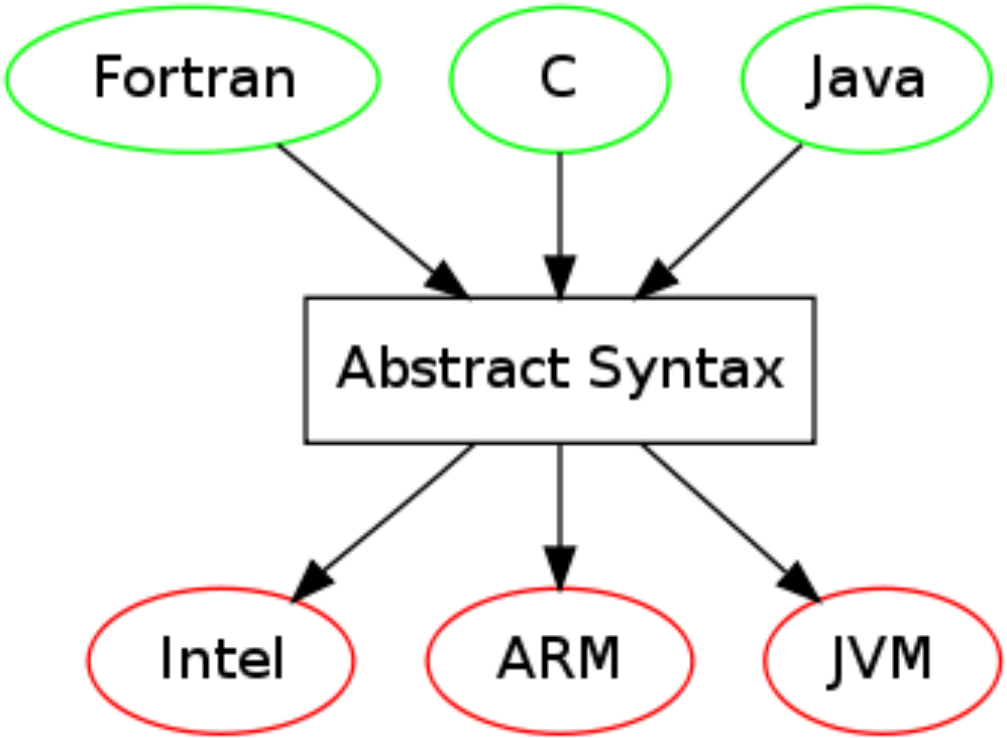
Based on **semantics, grammar,** or **chunks.**

https://play.google.com/store/apps/details?id=org.grammaticalframework.ui.android

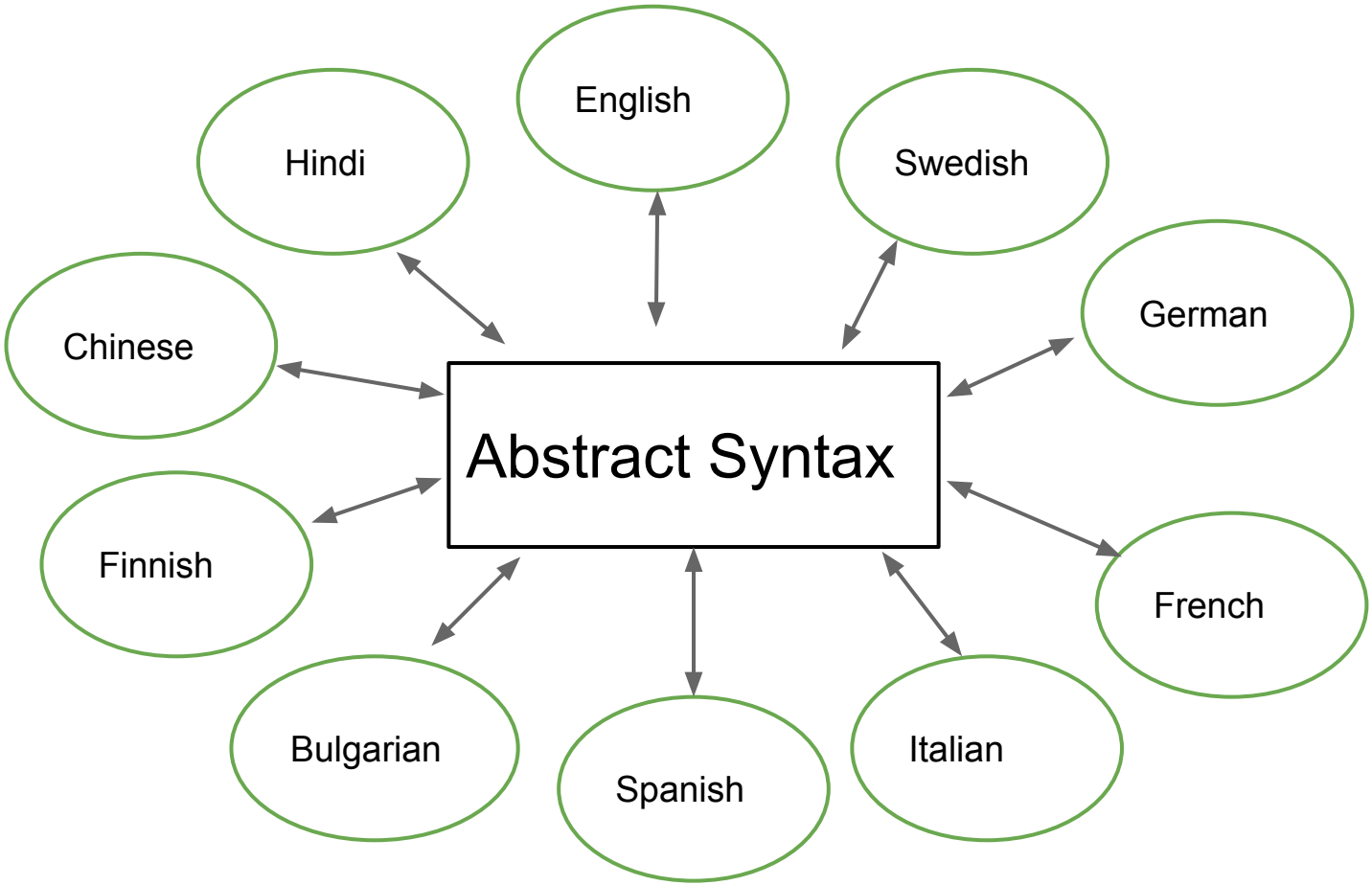
http://www.grammaticalframework.org/~aarne/App11.apk

A bit on how it works

Translation model: multi-source multi-target compiler



Translation model: multi-source multi-target compiler-**decompiler**



Abstract and concrete syntax

Abstract syntax: shared structure and semantics

Concrete syntax: language-specific details

Abstract and concrete syntax

Abstract syntax

```
fun Have : Person -> Number -> Item -> Sentence
```

Abstract and concrete syntax

Abstract syntax

```
fun Have : Person -> Number -> Item -> Sentence
```

Concrete syntax, English

```
lin Have p n i = p ++ “have” ++ n.s ++ i ! n.n
```

Abstract and concrete syntax

Abstract syntax

```
fun Have : Person -> Number -> Item -> Sentence
```

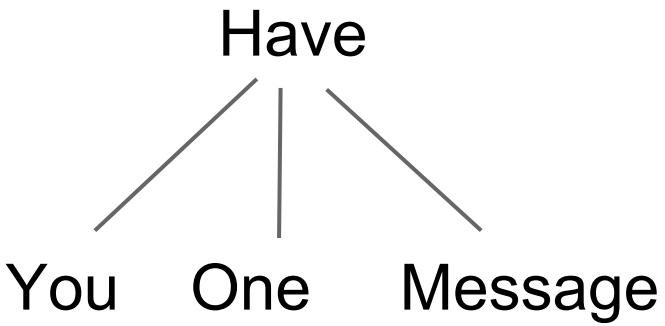
Concrete syntax, English

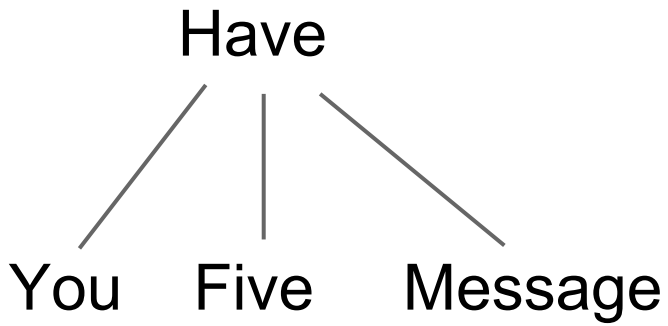
```
lin Have p n i = p ++ “have” ++ n.s ++ i ! n.n
```

Concrete syntax, Chinese

```
lin Have p n i = p ++ “有” ++ n ++ i.c ++ i.s
```

Abstract trees and linearizations





you have one message

你 有 一 个 信 息

you have five messages

你 有 五 个 信 息

A very small complete GF grammar

Abstract syntax

```
cat
 Sentence ;
 Item ;
 Person ;
 Number ;
fun
 Have :
  Person ->
  Number -> Item ->
  Sentence ;
 You : Person ;
 One : Number ;
 Five : Number ;
 Message : Item ;
```

Concrete syntax: English

```
lincat
 Sentence = Str ;
 Item = Num => Str ;
 Person = Str ;
 Number =
   {s : Str ; n : Num} ;
lin
 Have p n i =
   p ++ "have" ++
   n.s ++ i ! n.n ;
 You = "you" ;
 One = {s="one"; n=Sg} ;
 Five = {s="five"; n=Pl};
 Message = table {
   Sg = "message" ;
   Pl => "messages"
   } ;
param Num = Sg | Pl ;
```

Concrete syntax: Chinese

```
lincat
 Sentence = Str ;
 Item =
   {s : Str ; c : Str} ;
 Person = Str ;
 Number = Str ;
lin
 Have p n i =
   p ++ "有" ++
   n ++ i.c ++ i.s ;
 You = "你" ;
 One = "一" ;
 Five = "五" ;
 Message =
  {s = "信 息"; c = "个"};
```

RGL = Resource Grammar Library

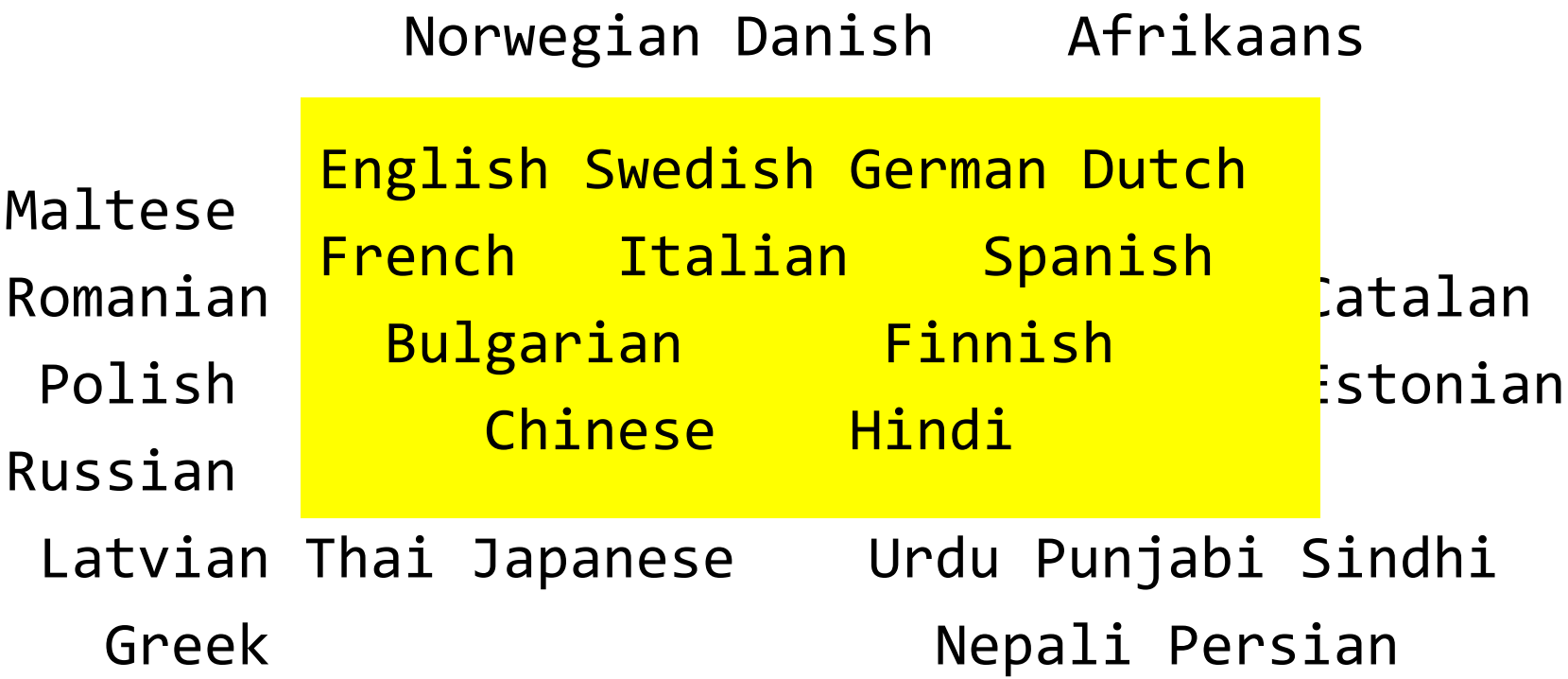
The standard library of GF

Takes care of linguistic details:

- morphology
- syntax

Makes GF productive and feasible

The RGL language potential



The English grammar with RGL

```
lincat
 Sentence = S ;
 Item = N ;
 Person = NP ;
 Number = Numeral ;

lin
 Have p n i = mkS (mkCl p have_V2 (mkNP n i)) ;
 You = you_NP ;
 One = mkNumeral “1” ;
 Five = mkNumeral “5” ;
 Message = mkN “message” ;
```

The Chinese grammar with RGL

```
lincat
 Sentence = S ;
 Item = N ;
 Person = NP ;
 Number = Numeral ;

lin
 Have p n i = mkS (mkCl p have_V2 (mkNP n i)) ;
 You = you_NP ;
 One = mkNumeral “1” ;
 Five = mkNumeral “5” ;
 Message = mkN “信 息” ;
```

The French grammar with RGL

```
lincat
 Sentence = S ;
 Item = N ;
 Person = NP ;
 Number = Numeral ;

lin
 Have p n i = mkS (mkCl p have_V2 (mkNP n i)) ;
 You = you_NP ;
 One = mkNumeral “1” ;
 Five = mkNumeral “5” ;
 Message = mkN “message” masculine ;
```

my new house is very big

मेरा अजनबी शाला बहुत महत्वपूर्ण है

你爱我吗

est-ce que tu m'aimes

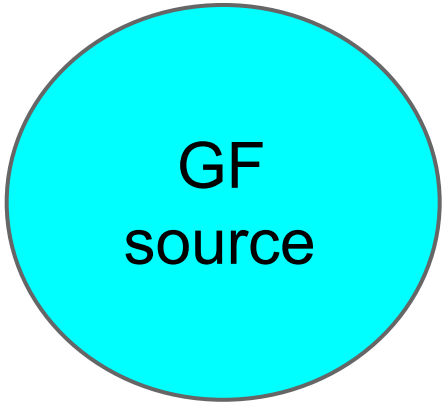
ich wohne in einem gelben Haus

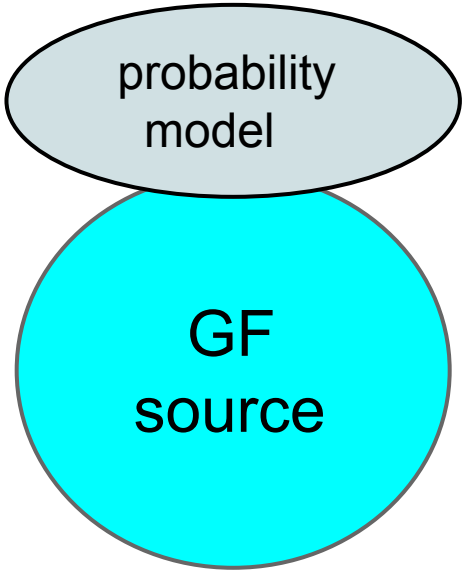
io risiedo in una casa gialla

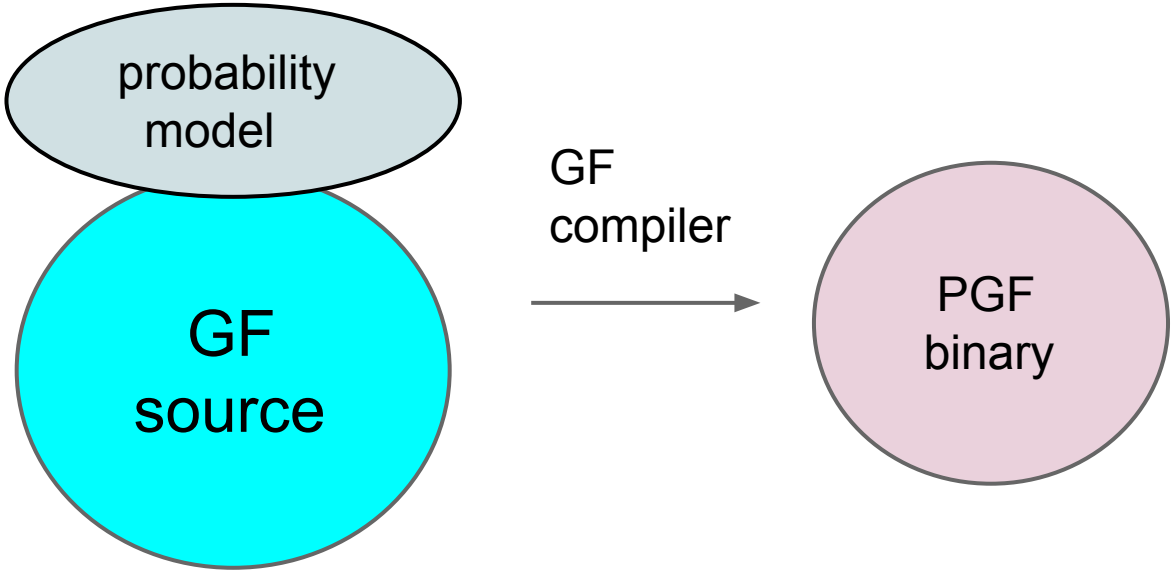
jag är inte en älg

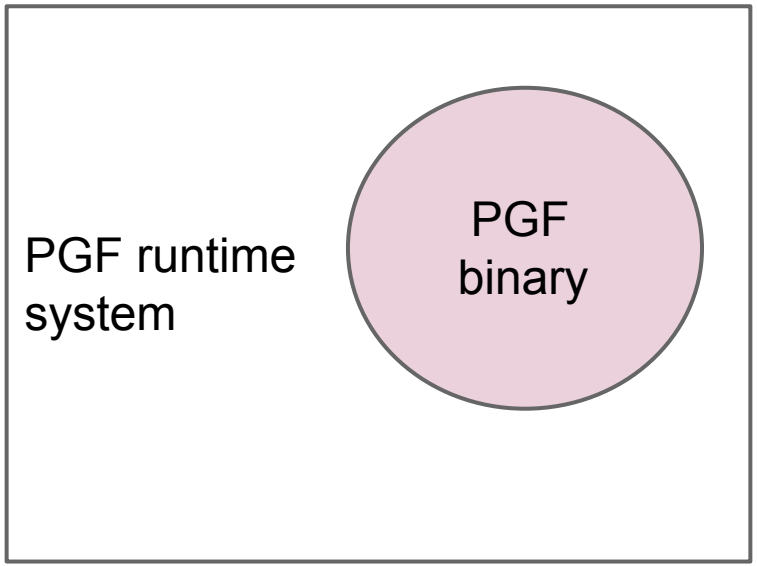
minä en ole hirvi

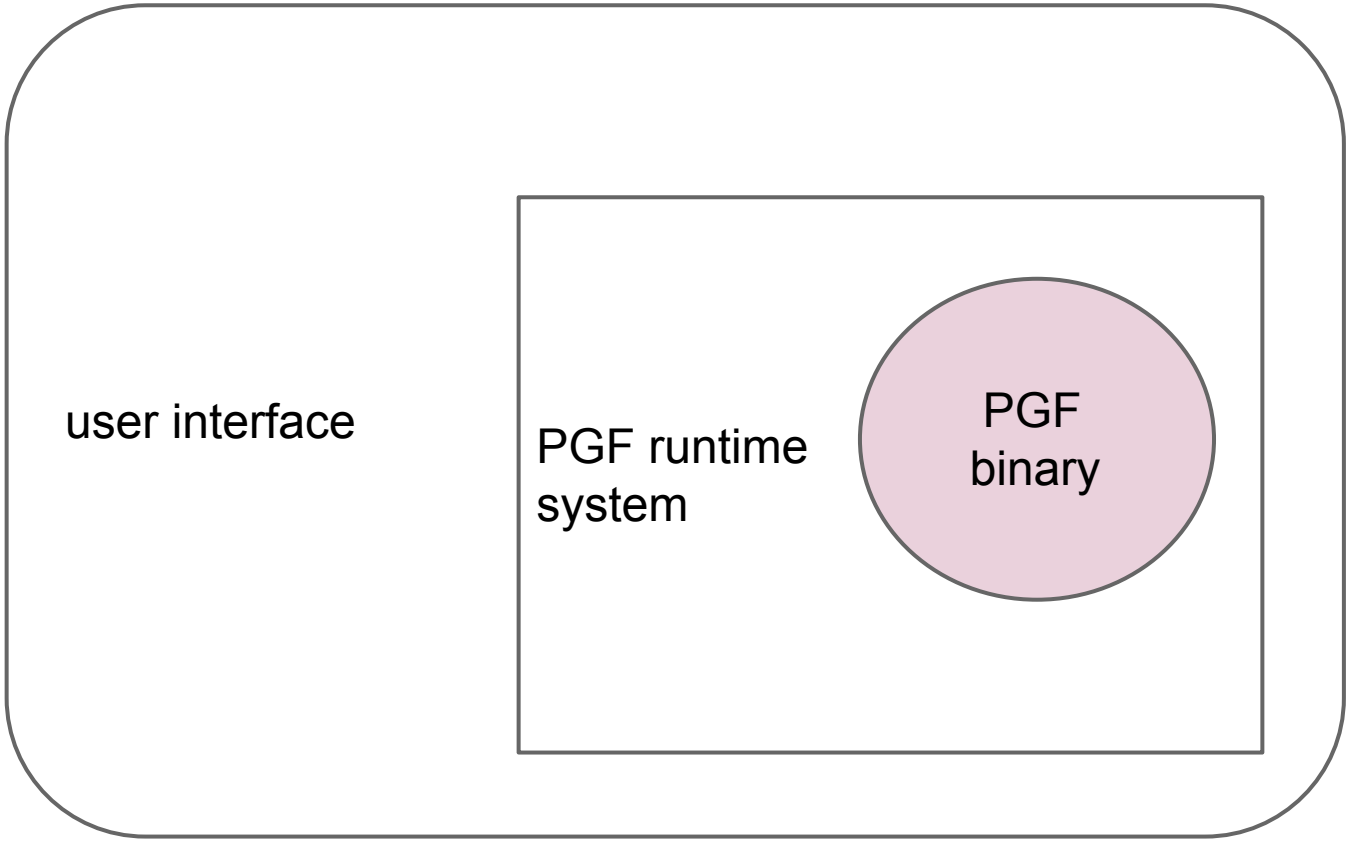
Building and maintaining GF applications

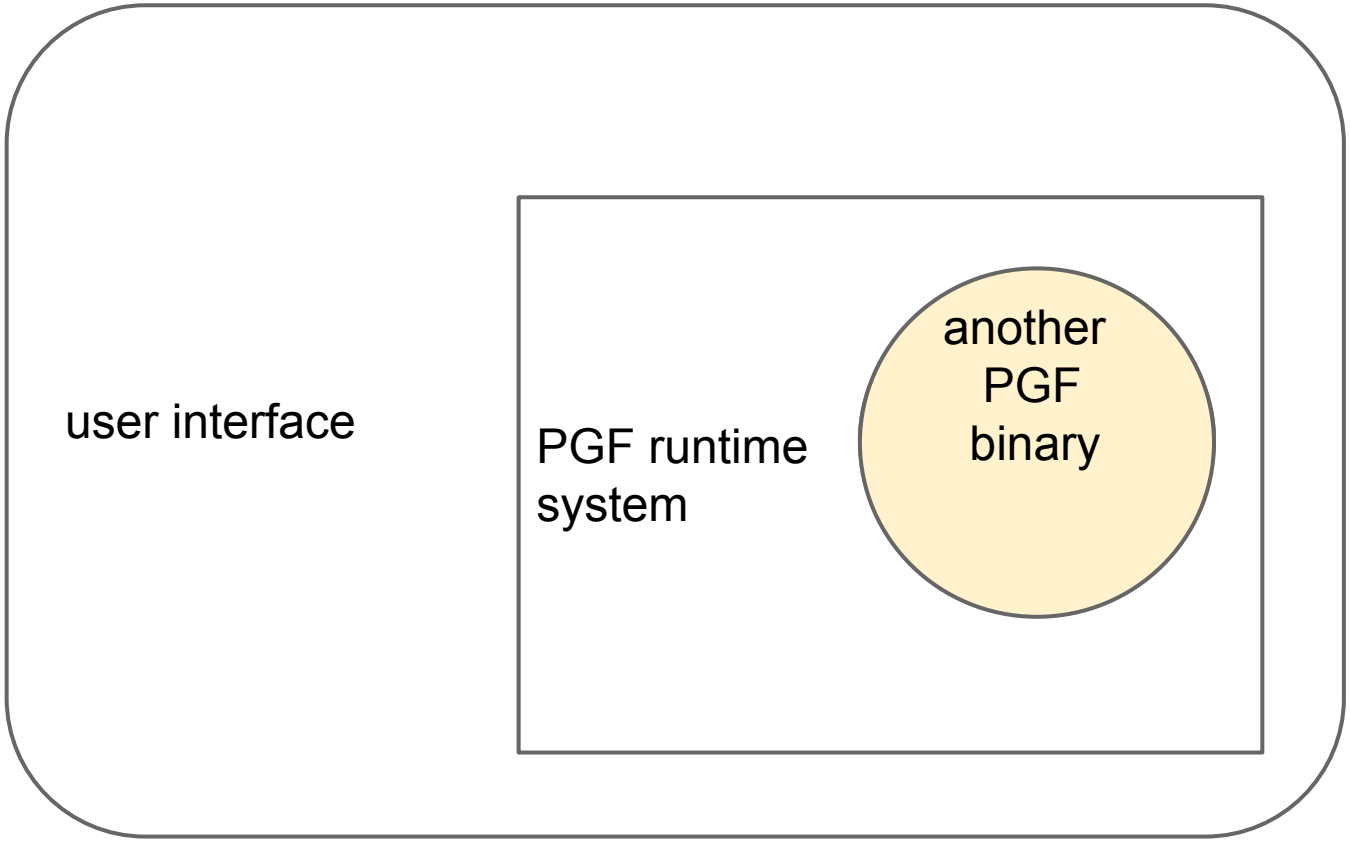

GF
source

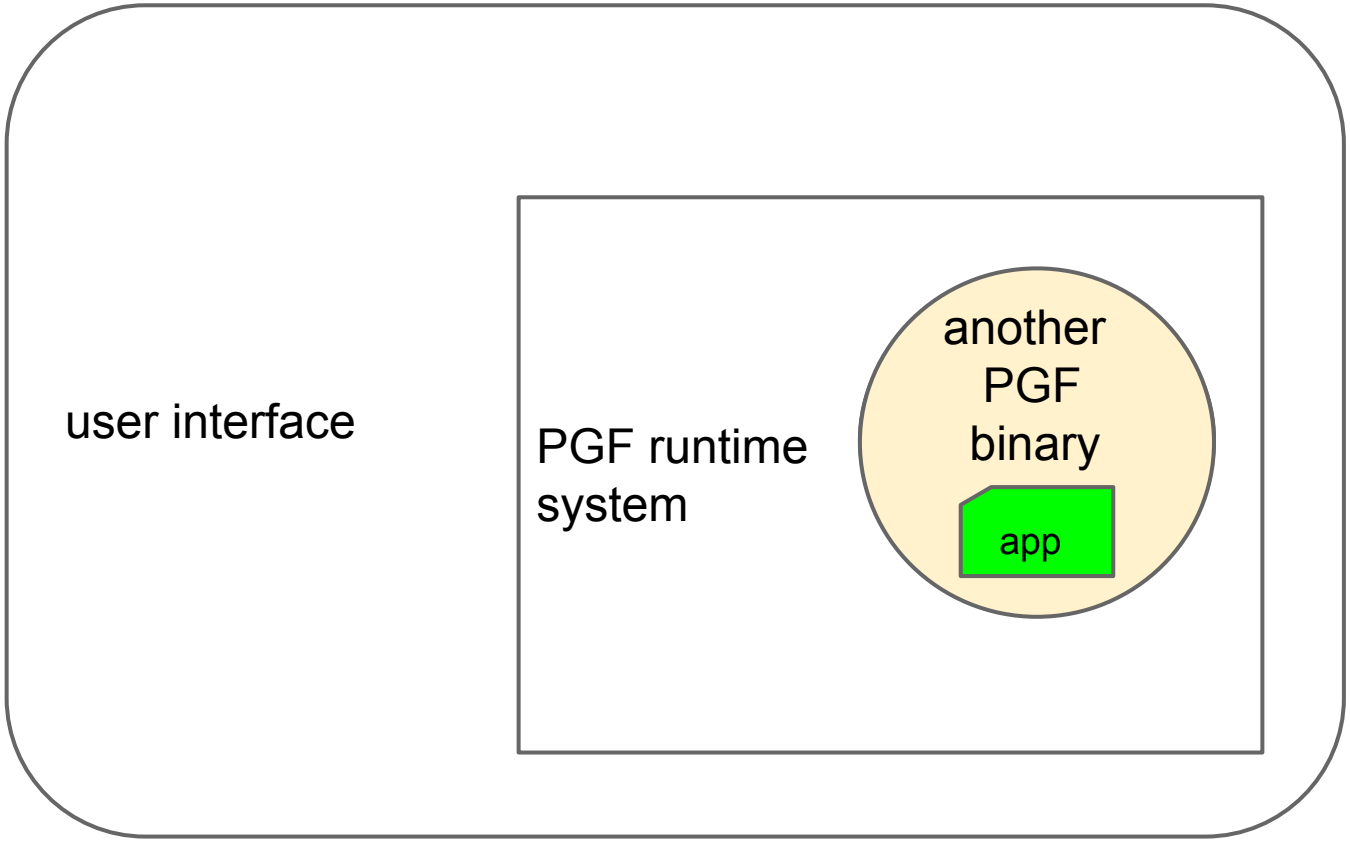

probability
model
GF
source

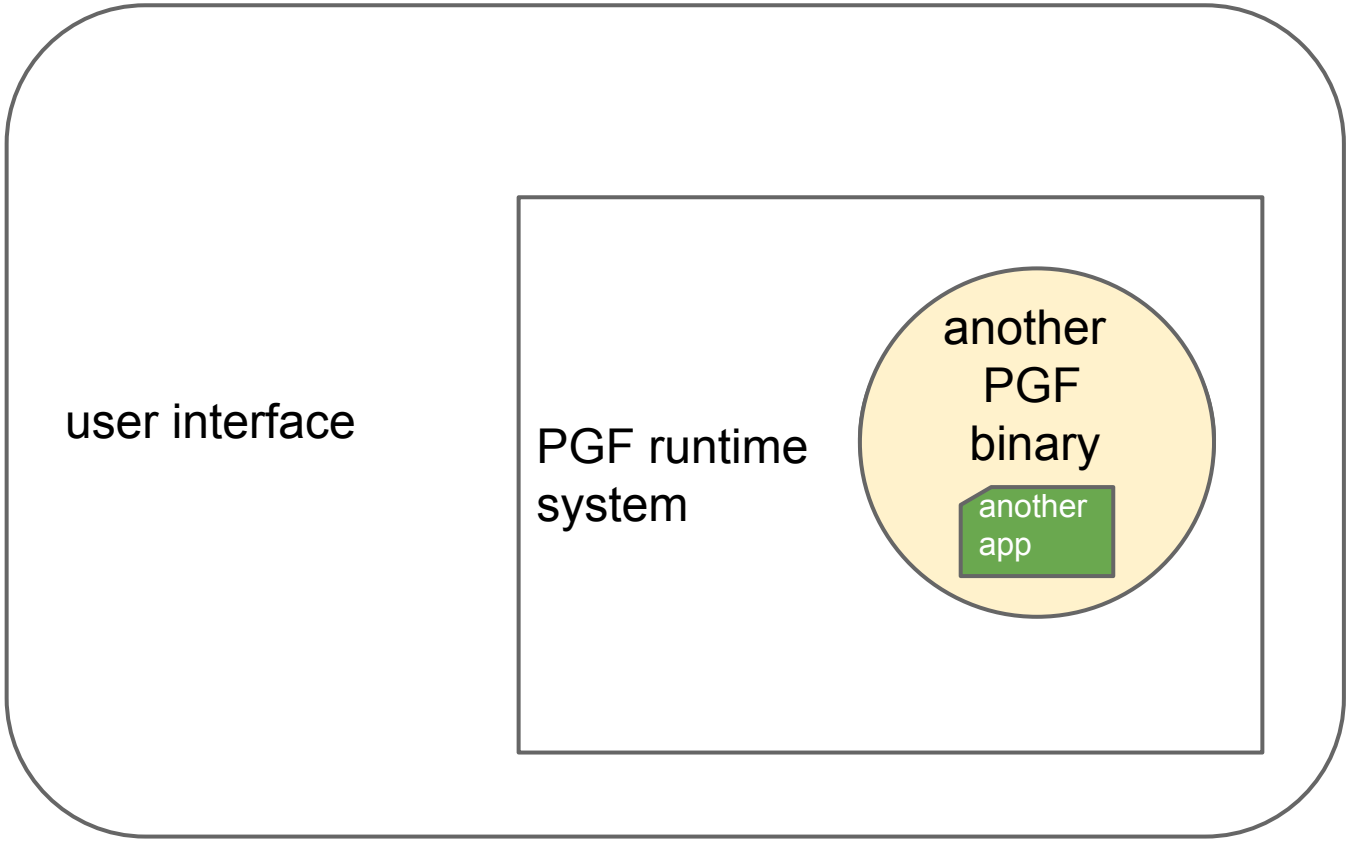

probability
model
GF
source
GF
compiler
PGF
binary


PGF runtime system
PGF binary

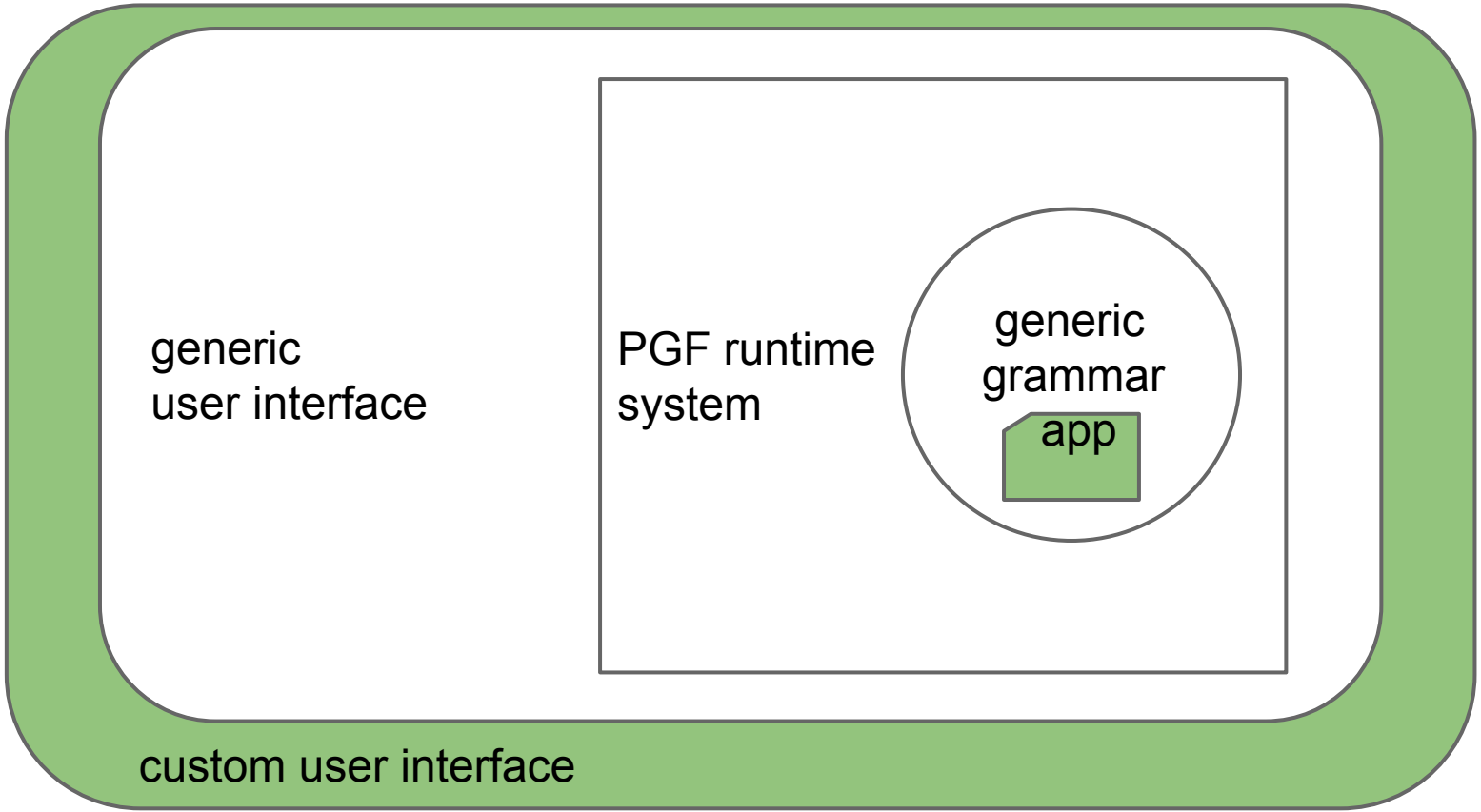

user interface
PGF runtime system
PGF binary


user interface
PGF runtime system
another PGF binary


user interface
PGF runtime system
another PGF binary
app


user interface
PGF runtime system
another PGF binary
another app

White: free, open-source. **Green**: what we sell.



Open source policy

Created with public funding:

- open source, free: also for other companies
- GF platform and language resources

Proprietary extensions allowed

- customer-paid work
- customer’s data

Anyone is allowed to build a business on this!

Some existing application domains

Tourist phrasebook (MOLTO)

Multilingual Wiki (ACE)

Patent query language (Ontotext)

Museum query language and texts (Ontotext)

Business models (Be Informed)

Medical examination journals (Lingsoft)

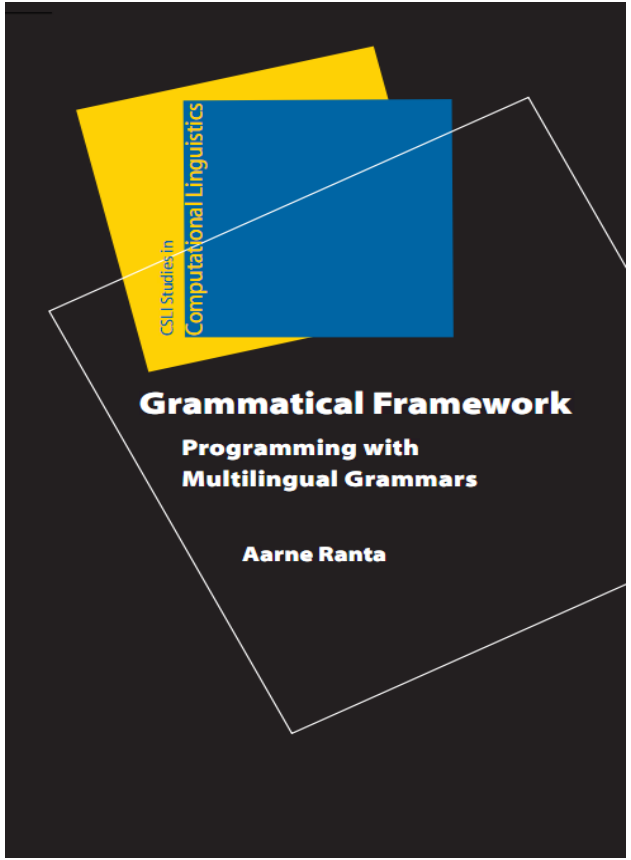
Speech commands in cars (Talkamatic)

Resources

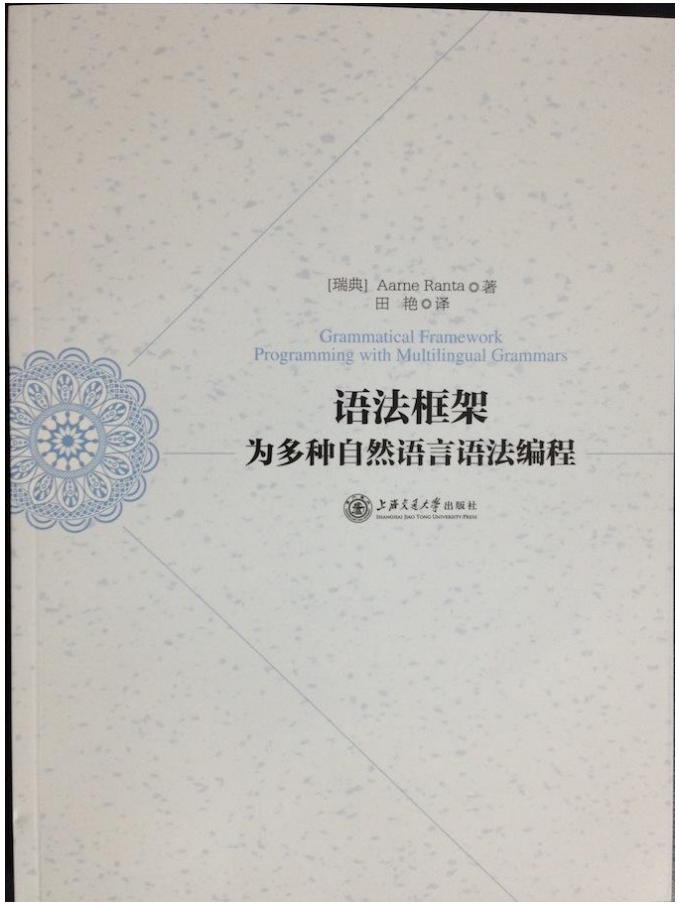
GF homepage and community

Digital Grammars AB

The GF book



CSLI, Stanford, 2011



Shanghai Jiao Tong University press, 2014

GF World Map

Conclusion

GF: translation for 29 languages

Control on translation quality

Easily tailored to new domains to ensure production quality