Vägar till bättre översättningsprogram

Aarne Ranta, Thomas Hallgren, Krasimir Angelov

Data- och informationsteknik
Göteborgs universitet & Chalmers tekniska högskola

Vetenskapsfestivalen 8 maj 2014, Göteborg

Inom fem år, kanske tre, kan mellanspråkig meningsöverföring genom en elekronisk process inom viktiga funktionella områden av ett flertal språk mycket väl att vara verklighet.

Inom fem år, kanske tre, kan mellanspråkig meningsöverföring genom en elekronisk process inom viktiga funktionella områden av ett flertal språk mycket väl att vara verklighet.

IBM pressmeddelande 1954

Vad hände sedan?

1966 **ALPAC-rapport**: översättning med dator blir dubbelt så dyrt som för hand

1989 **IBM** igen: helautomatisk översättning byggd på statistik från texter

2014 **Google** translate med 80 språk

Men kan man lita på den?

Bäst att experimentera

- börja med språk som du förstår
- börja med korta meningar och öka längden

Kan du förstå vad dokumentet säger?

Skulle du våga publicera översättningen?

Google translate

Baserad på **statistik**

- översätter “allt”
- snabb
- bra på vanliga meningar
- oförutsägbar
  - men man kan faktiskt förutsäga en del

https://translate.google.com

Experiment med Google translate

engelska till svenska

my house / my car

my new house is very big

svenska till engelska

min far är svensk

min far är inte svensk

svenska till norska

din far är inte svensk

tyska till svenska/engelska

er bringt mich um

er bringt deinen besten Freund um

GF translate

Baserad på **grammatik**

- förutsägbar
- visar konfidensnivåer
- bra på grammatik
- kan lämna luckor
- kan vara långsam

http://www.grammaticalframework.org/demos/translation.html

Experiment med GF translate

Samma som med Google, samt

vad heter din fru

svenska till många andra språk.

Översätta med grammatik: hur fungerar det?

Beräkningsregler

Datorn kan **följa regler mekaniskt** - bättre än människan

- `2 + 2 = 4`
- `365 * 24 * 60 * 60 = 31536000`

Här väntar vi oss att datorn alltid gör rätt!

Kompilator

Programkod till maskinkod:

```
printf("hello world")
```

```
0101011011010001010101010100101001
```

Här har datorn ersatt människan som översättare!

Grammatikregler

Morfologi: böjning

- *känna -> känner, kände, känt, känd, kända*

Syntax: ordföljd och kongruens

- *den + stor + hus -> det stora huset*
- *jag + sova + inte -> jag sover inte*
- *om + jag + sova + inte -> om jag inte sover*

Kongruens på lång distans

Den allra största frågan är fortfarande öppen, och det är inte säkert att ledningen någonsin kommer att nå överenskommelse om den.

Kongruens på lång distans

Den allra största problemet är fortfarande öppen, och det är inte säkert att ledningen någonsin kommer att nå överenskommelse om den.

Kongruens på lång distans

Den allra största problemet är fortfarande öppen, och det är inte säkert att ledningen någonsin kommer att nå överenskommelse om den.

Kongruens på lång distans

Det allra största problemet är fortfarande öppet, och det är inte säkert att ledningen någonsin kommer att nå överenskommelse om det.

Detta kan vara svårt för människan men lätt för datorn.

Grammatik i översättning

Språk har

- olika ord
- olika ordföljdsregler
- olika kongruensregler

Men de kan också

- uttrycka samma mening
- ha gemensam **abstrakt struktur**

Exempel: kompilator

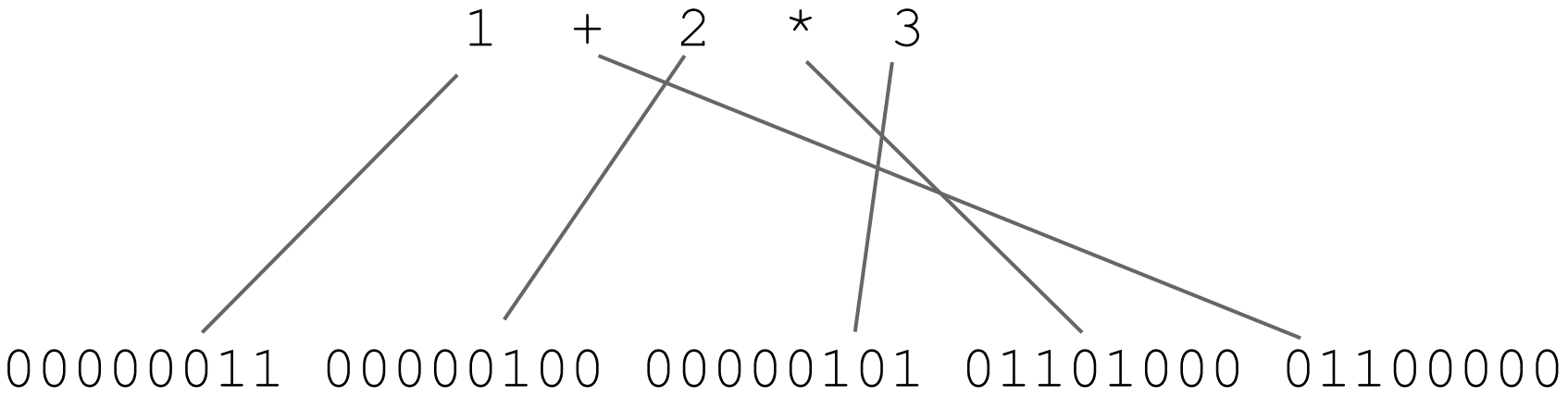
Programkod (Java)

```
1 + 2 * 3
```

Maskinkod (JVM, Java Virtual Machine)

```
00000011 00000100 00000101 01101000 01100000
```

Motsvarigheterna



Abstrakt syntax

Add : Exp -> Exp -> Exp

Mul : Exp -> Exp -> Exp

E1, E2, E3 : Exp

Add E1 (Mul E2 E3)

Konkret syntax

abstrakt	**Java**	**JVM**
Add x y	*x* “+” *y*	*x y* “01100000”
Mul x y	*x* “*” *y*	*x y* “01101000”
E1	“1”	“00000011”
E2	“2”	“00000100”
E3	“3”	“00000101”

Konpilering av naturligt språk

Abstrakt syntax

Pred : Subject -> Verb -> Object -> Sentence

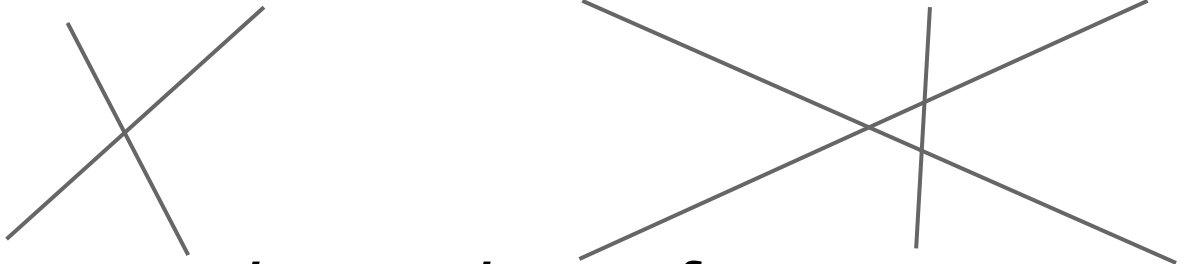
Mod : Adjective -> Noun -> Noun

Love : Verb

Konkret syntax:	**svenska**	**latin**
Pred s v o	*s v o*	*s o v*
Mod a n	*a n*	*n a*
Love	*“love”*	*“amare”*

Exempel

den kloka kvinnan älskar den fagra mannen



femina sapiens virum formosum amat

Problem: ordens flertydighet

The pen is in the box.

Pennan är i lådan.

The box is in the pen.

Lådan är i lekhagen. (Bar-Hillel 1963)

Problem: flertydig syntax

I ate a pizza with shrimps.

I ate a pizza with friends.

Problem: flertydig syntax

I ate a pizza with shrimps.

I ate a pizza with friends.

Problem: idiomatiska uttryck

what is your name

- *vad heter du* **ej** *vad är ditt namn*
- *comment t'appelles-tu* **ej** *quel est ton nom*

kick the bucket -> ta ner skylten

finlandsbåt -> ruotsinlaiva

Problem: språkens komplexitet

7000 språk

100.000 ord

10.000.000.000 tvåordskombinationer...

Bar-Hillels slutsats

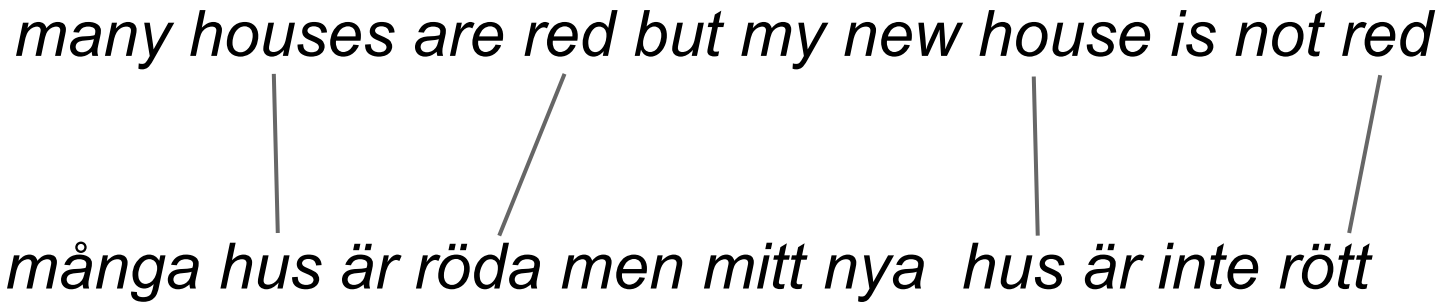
(1963)

Fullt automatisk översättning med hög kvalité kommer aldrig att vara möjligt.

Översätta med statistik: hur fungerar det?

Lexikon: ordlinjering

Hitta ord som motsvarar varandra, genom att analysera en stor mängd **parallella texter**.



Syntax: n-gram

Hjälper vid valet av ord i **kontext**

my	*min*
	mitt
	mina
new	*ny*
	nytt
	nya
house	*hus*
	huset

Problem med ordlinjering

Översättning kan inkludera **lokalisering**

this computer costs 1000 dollars

den här datorn kostar 7000 kronor

Problem: glesa data

100.000 ord

1.000.000 böjningsformer

1.000.000.000.000 2-gram

Problem: beroende på lång distans

Kongruens

mitt nya hus är mycket stor

Um+bringen = kill

er bringt deinen besten Freund um

Problem: alla ord är lika viktiga

Att negationen saknas kostar inte mycket…

min far är svensk -> my father is Swedish
min far är inte svensk -> my father is Swedish

Who is responsible for the translation?

price 500 dollars -> pris 500 kronor

producer?

consumer?

Slå ihop grammatik och statistik: hybridsystem

Ta det bästa av båda

Grammatik

- böjning
- ordföljd
- kongruens

Statistik

- val av ordbetydelse (med n-gram)
- samling av idiomer (med **fraslinjering**)

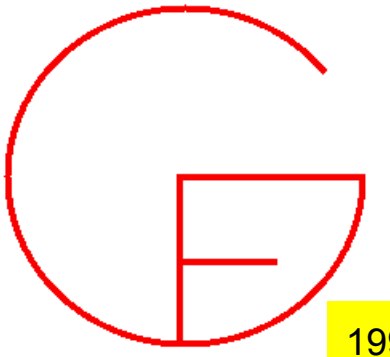
Vårt webbsystem och mobilapp

- hur de används
- hur vi byggde dem

5 March 2014 -

VR 2013-2017

EU 2010-2013



1998 -

what is your wife's name

vad heter din fru

the vice president kicked the bucket

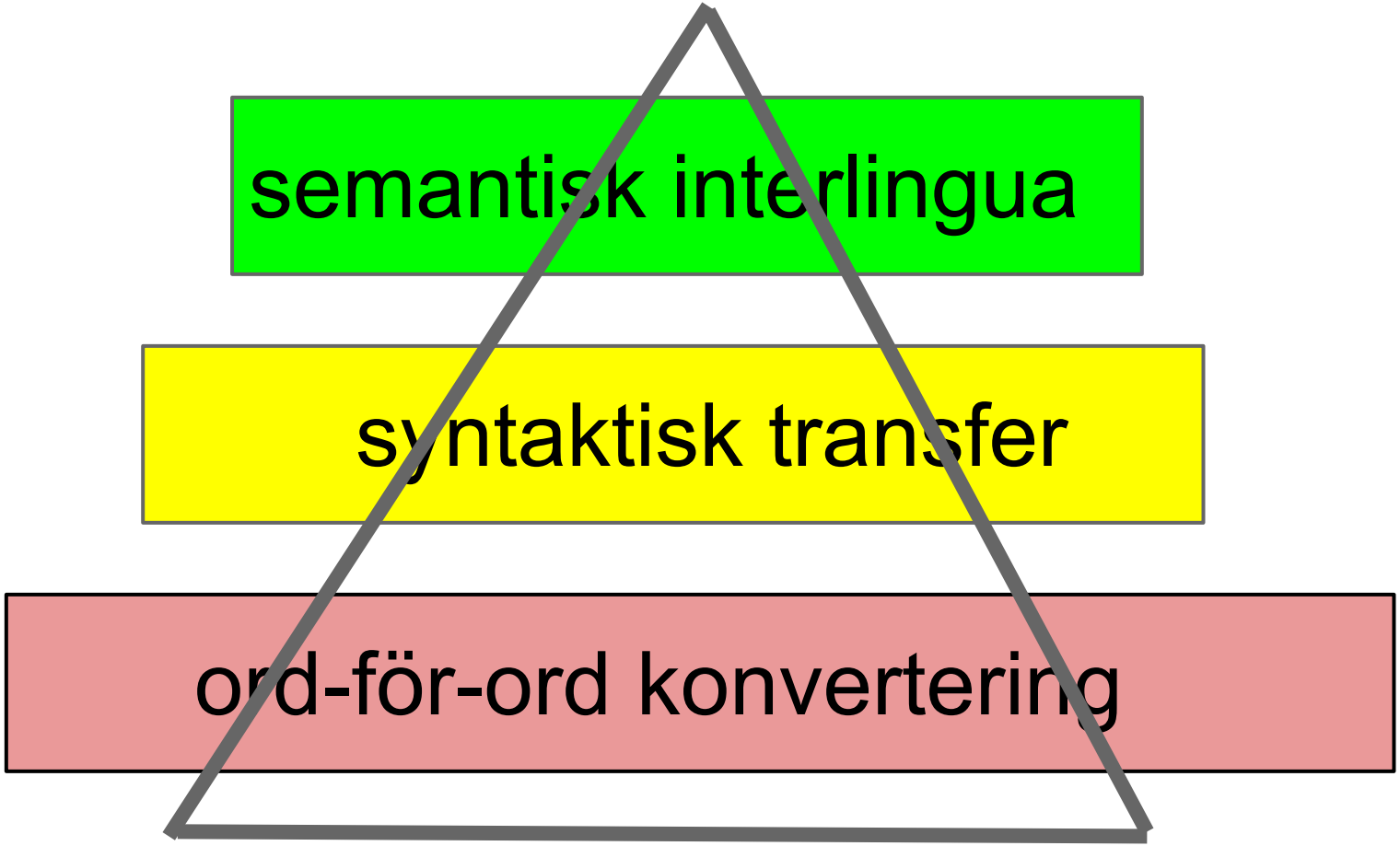
skruvstädspresidenten sparkade hinken

long time no see

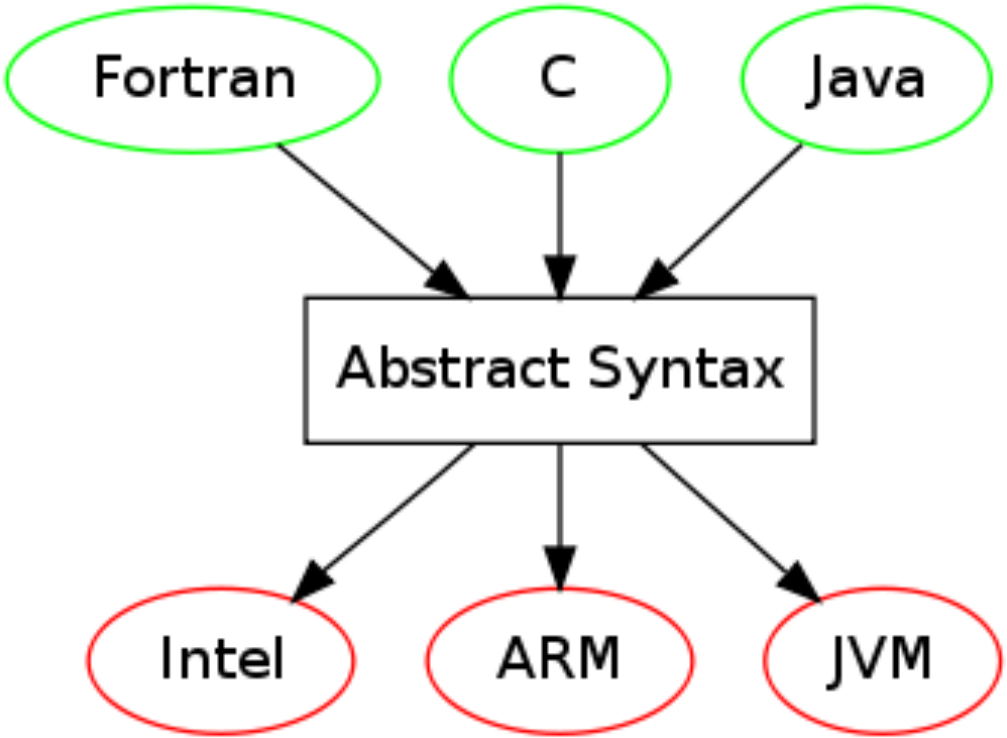
lång tid nej ser

Översättning i färg

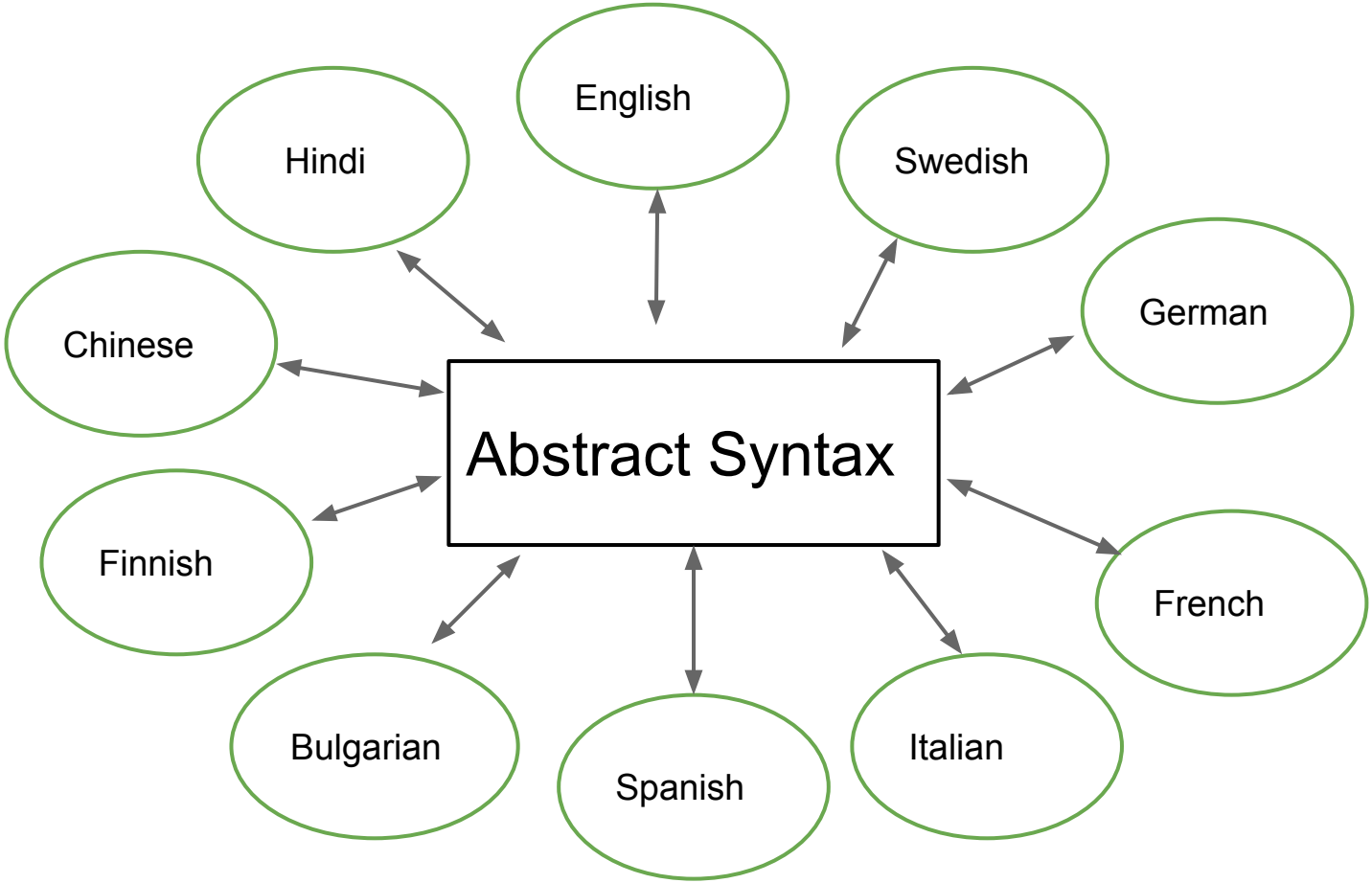
Vauquois-triangeln



Översättningsmodellen: kompilator



“The Human Language Compiler”



Semantisk interlingua

Den “gröna” översättningen

Begränsad i omfattning

Oftast ett speciellt område

- matematik
- restaurangfraser

Betydelse beror på område

delivery

Betydelse beror på område

delivery

- *förlossning* (inom medicin)
- *leverans* (inom handel)

Syntaktisk transfer

Den “gula” översättningen

Obegränsad i omfattning

- men klarar inte av allt

Ej idiomatisk

- “bokstavlig” översättning

Ord för ord konvertering

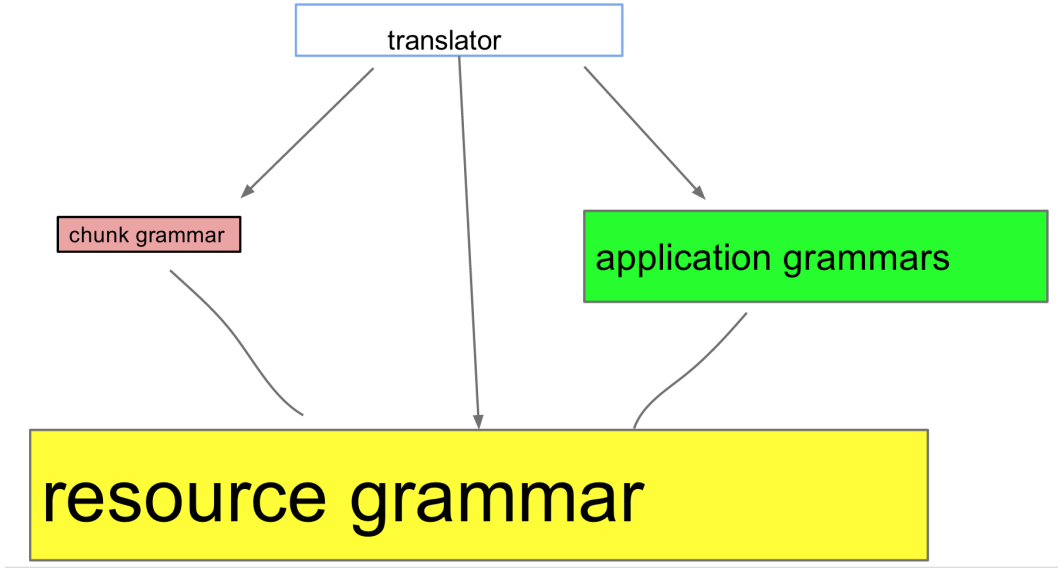
Den “röda” översättningen

Klarar ogrammatisk text

- samt luckor i grammatiken

“Something is better than nothing”

Combination of different grammars



my new house is very big

मेरा अजनबी शाला बहुत महत्वपूर्ण है

你爱我吗

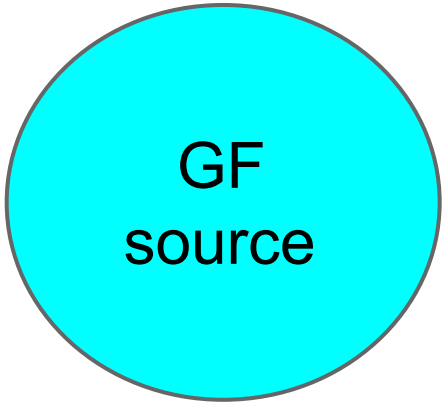
est-ce que tu m'aimes

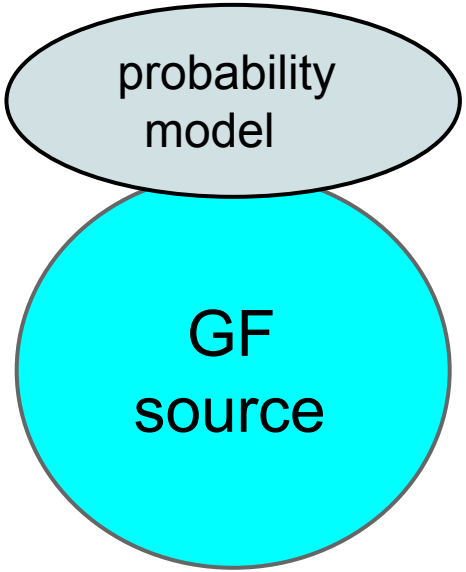
ich wohne in einem gelben Haus

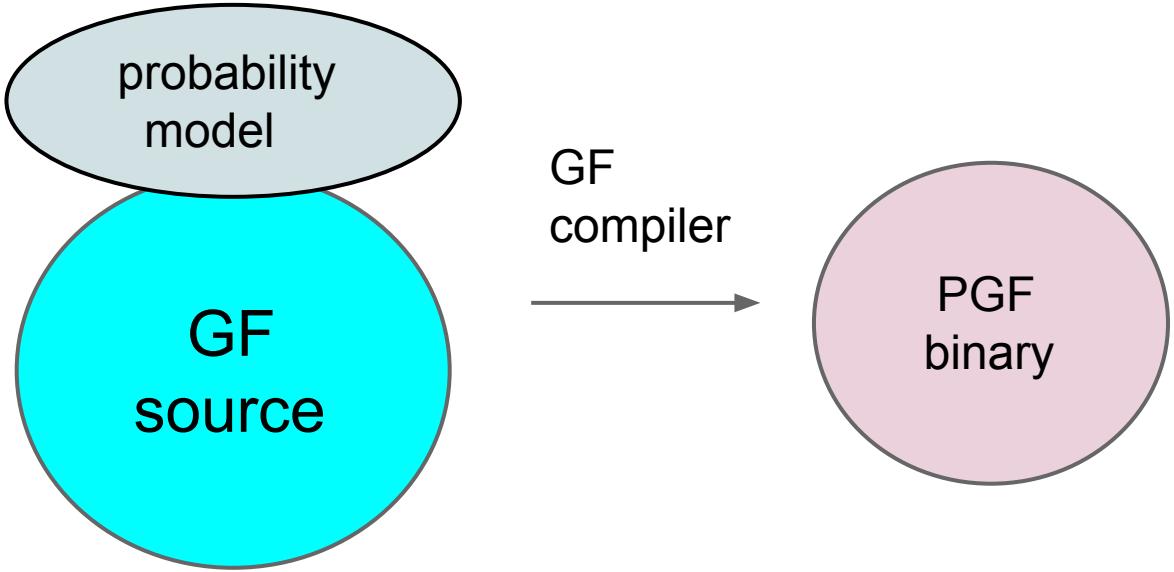
io risiedo in una casa gialla

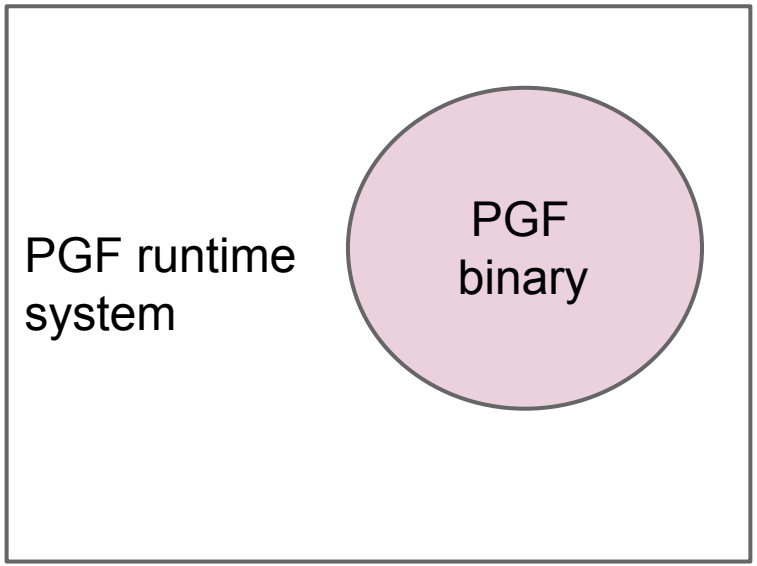
jag är inte en älg

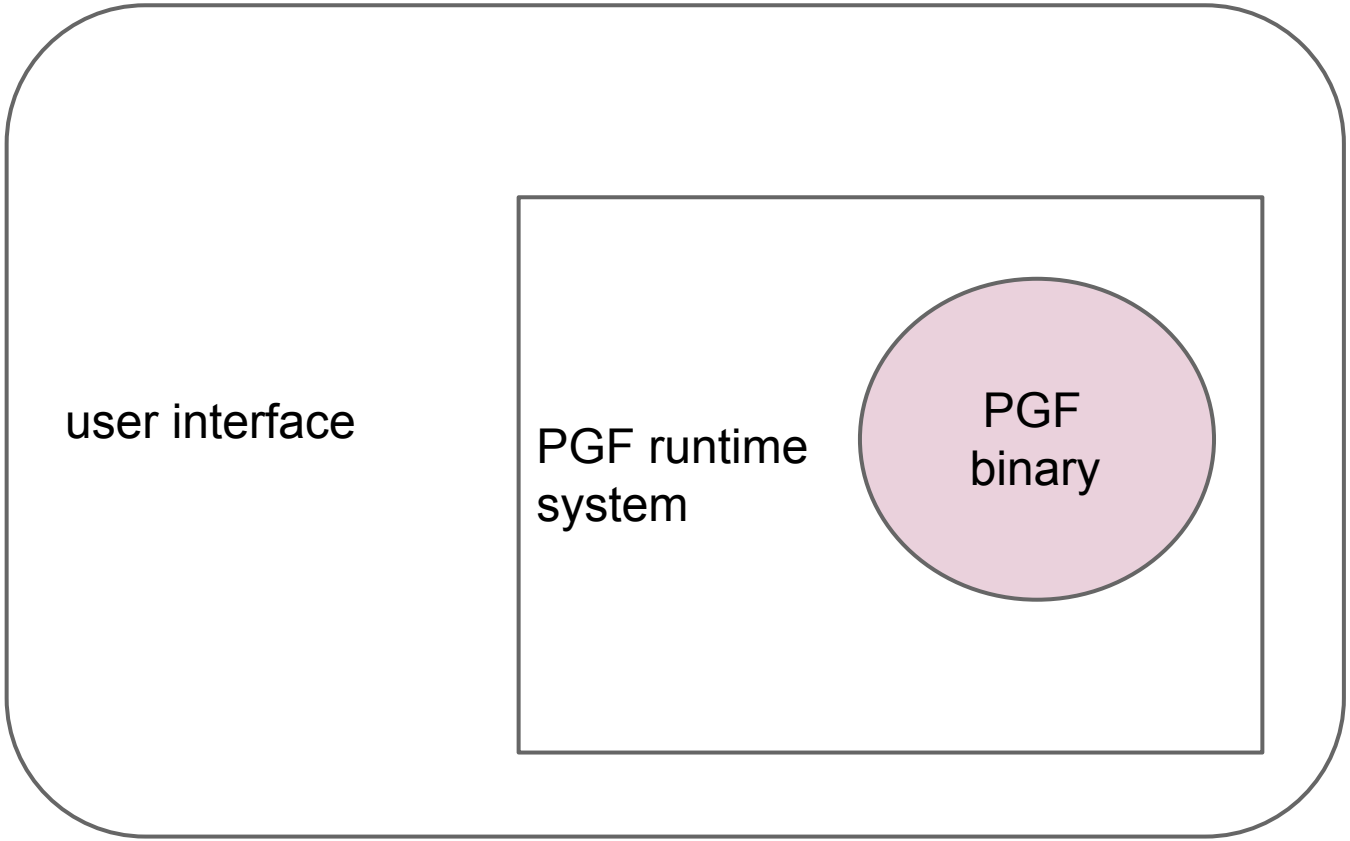
minä en ole hirvi

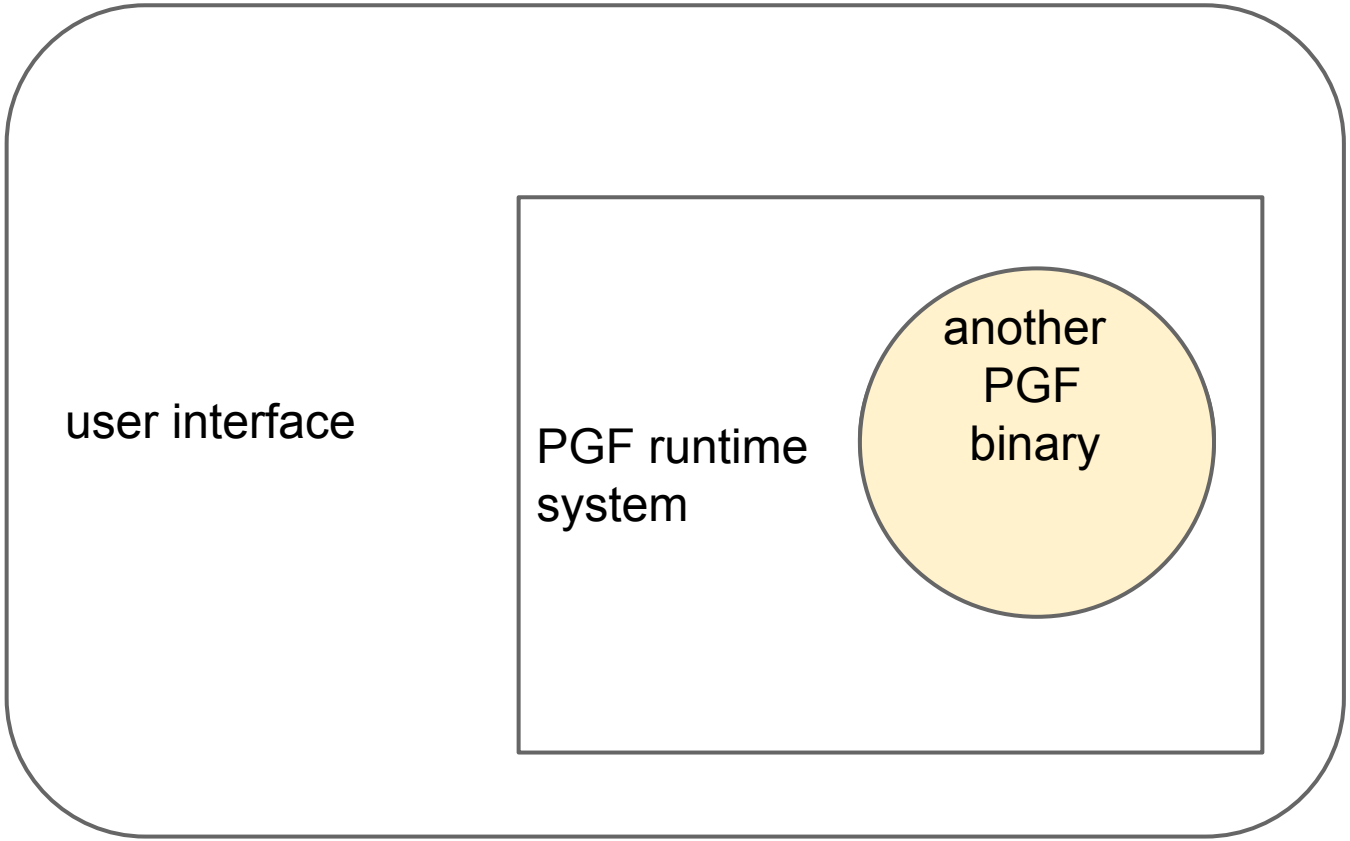

GF
source

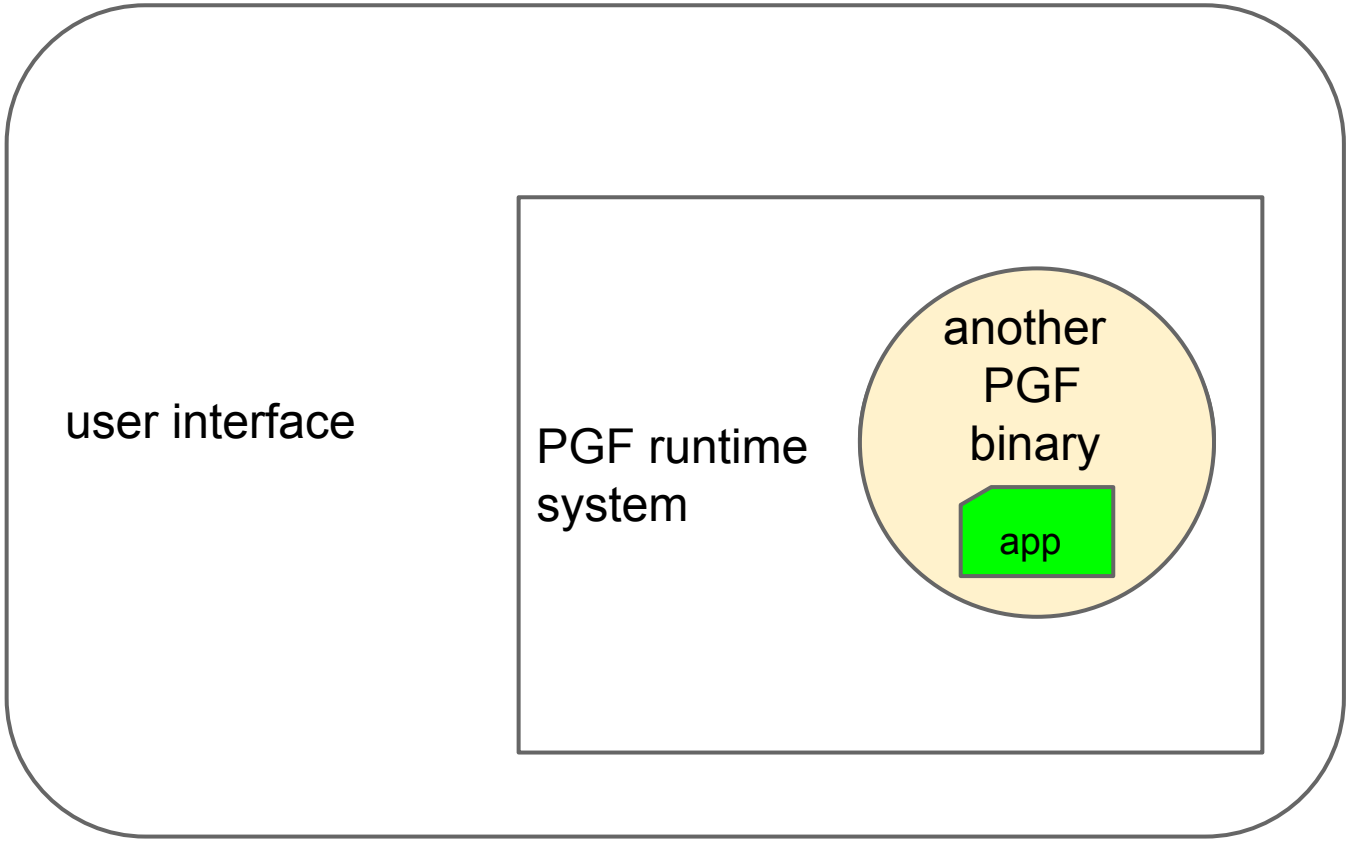

probability
model
GF
source

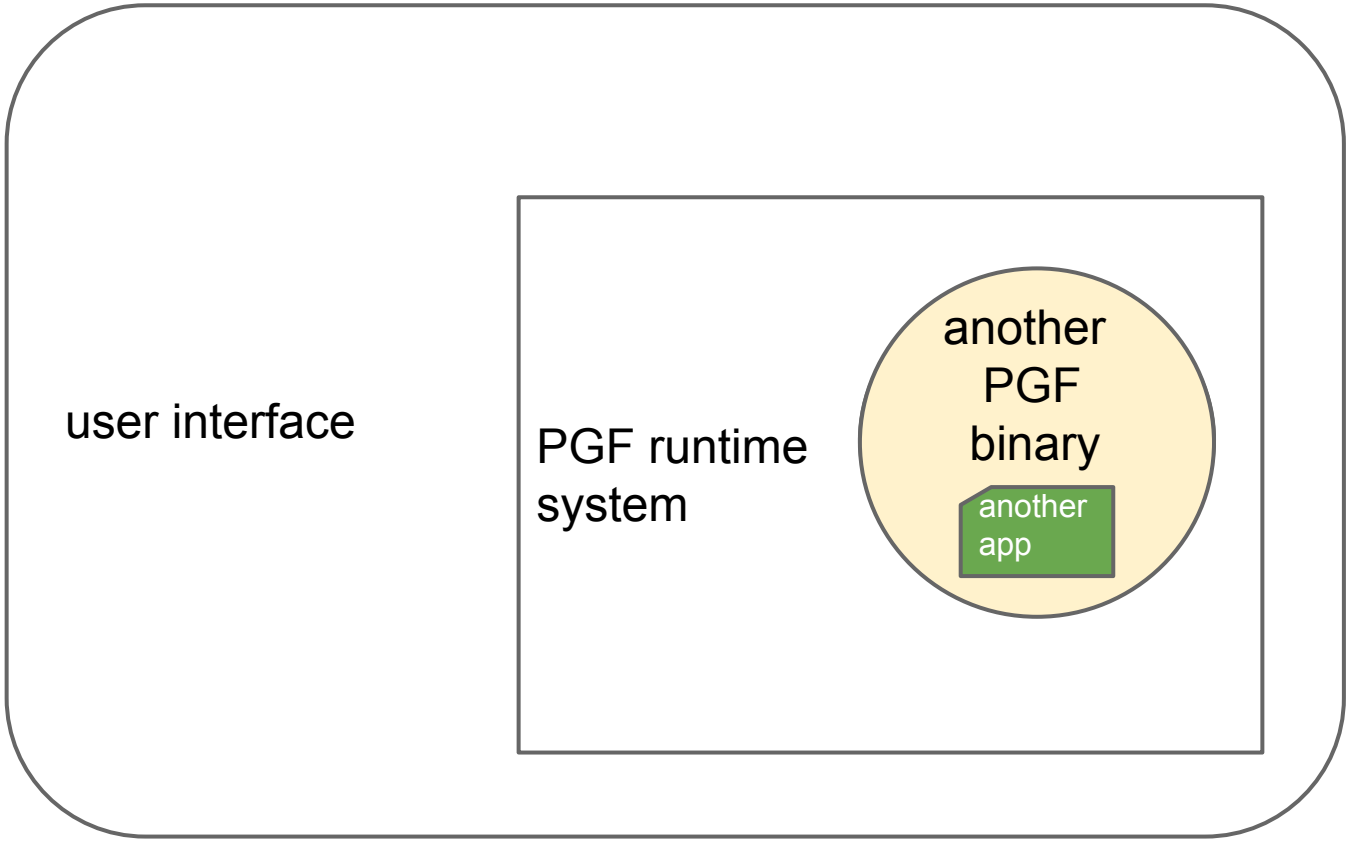

probability
model
GF
source
GF
compiler
PGF
binary


PGF runtime system
PGF binary

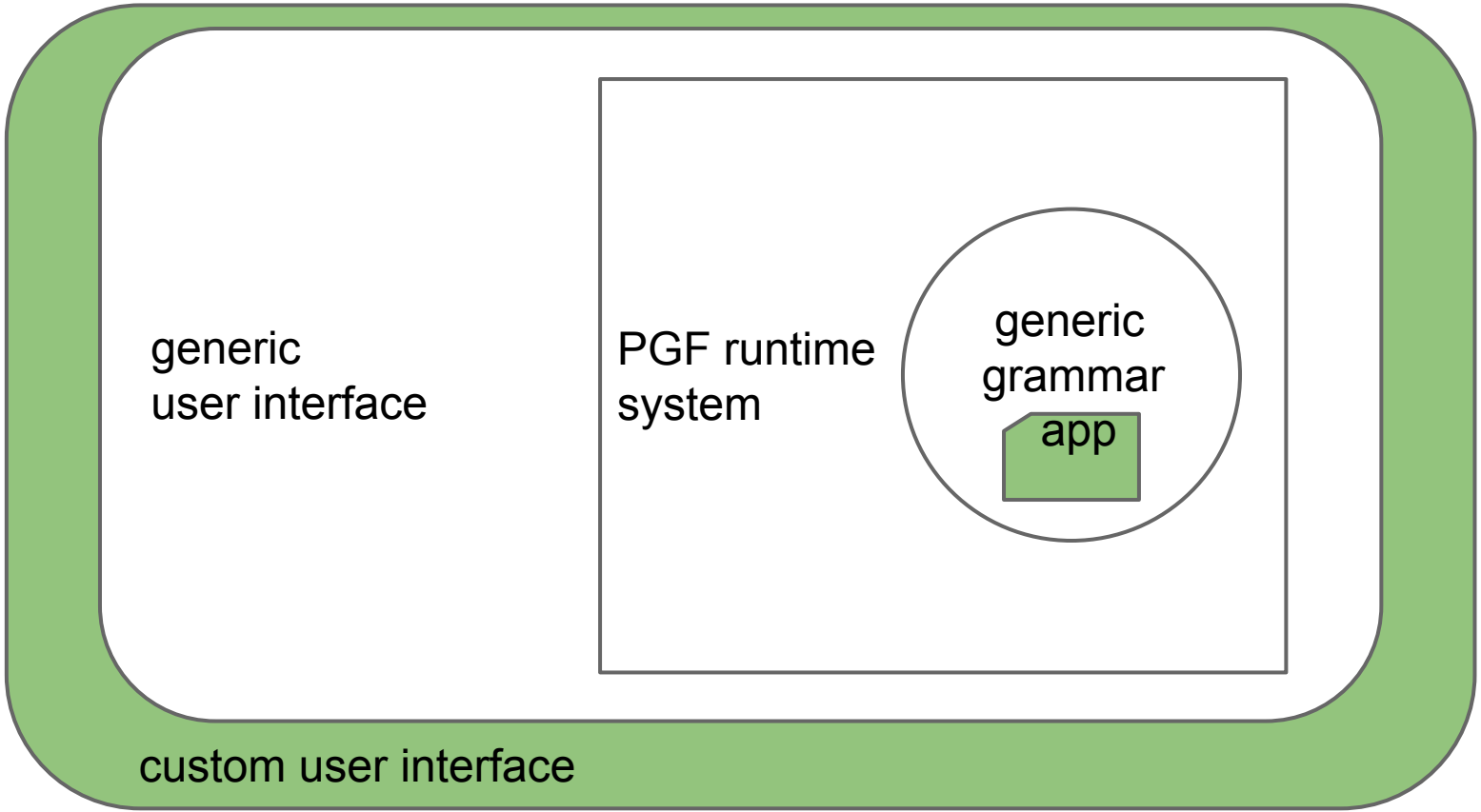

user interface
PGF runtime system
PGF binary


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

Customized translation system



Testa den själv!

Android app: Human Language Compiler

http://www.grammaticalframework.org/demos/app.html

- fungerar utan nätuppkoppling
- både text och tal
- konfidensfärger
- grammatisk information
- 23 MB för 110 språkpar (Google 100MB for 1 pair)

Web app

http://www.grammaticalframework.org/demos/translation.html

The Human Language Compiler