Data-Driven Documentation

A Technique for Reliable Multilingual Information Access

Aarne Ranta
University of Gothenburg & Digital Grammars AB

PIC-2015, Nanjing, 18-20 December 2015

digital Grammars
Language technology to rely on.

Problem

Solution

Technology

The problem:

reliable and efficient translation

Machine translation is sometimes good, sometimes bad - and you never know how it will be this time.

English | Swedish | Finnish | Detect language

Dutch | Chinese (Simplified) | English | Translate

Min mor är inte svensk.

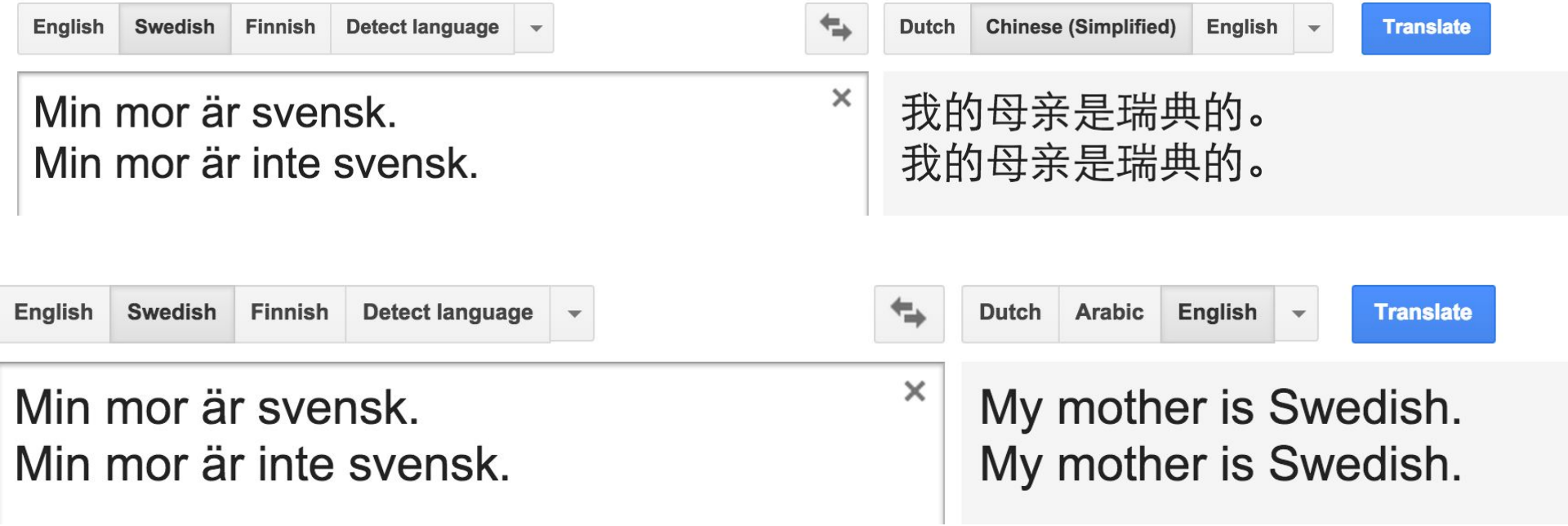
我的母亲是瑞典的。

English | Swedish | Finnish | Detect language

Dutch | Chinese (Simplified) | English | Translate

Min mor är svensk.
Min mor är inte svensk.

我的母亲是瑞典的。
我的母亲是瑞典的。


English
Swedish
Finnish
Detect language
Dutch
Chinese (Simplified)
English
Translate
Min mor är svensk.
Min mor är inte svensk.
我的母亲是瑞典的。
我的母亲是瑞典的。
English
Swedish
Finnish
Detect language
Dutch
Arabic
English
Translate
Min mor är svensk.
Min mor är inte svensk.
My mother is Swedish.
My mother is Swedish.

translate.google.com, 9 Dec 2015

Consumer translator:

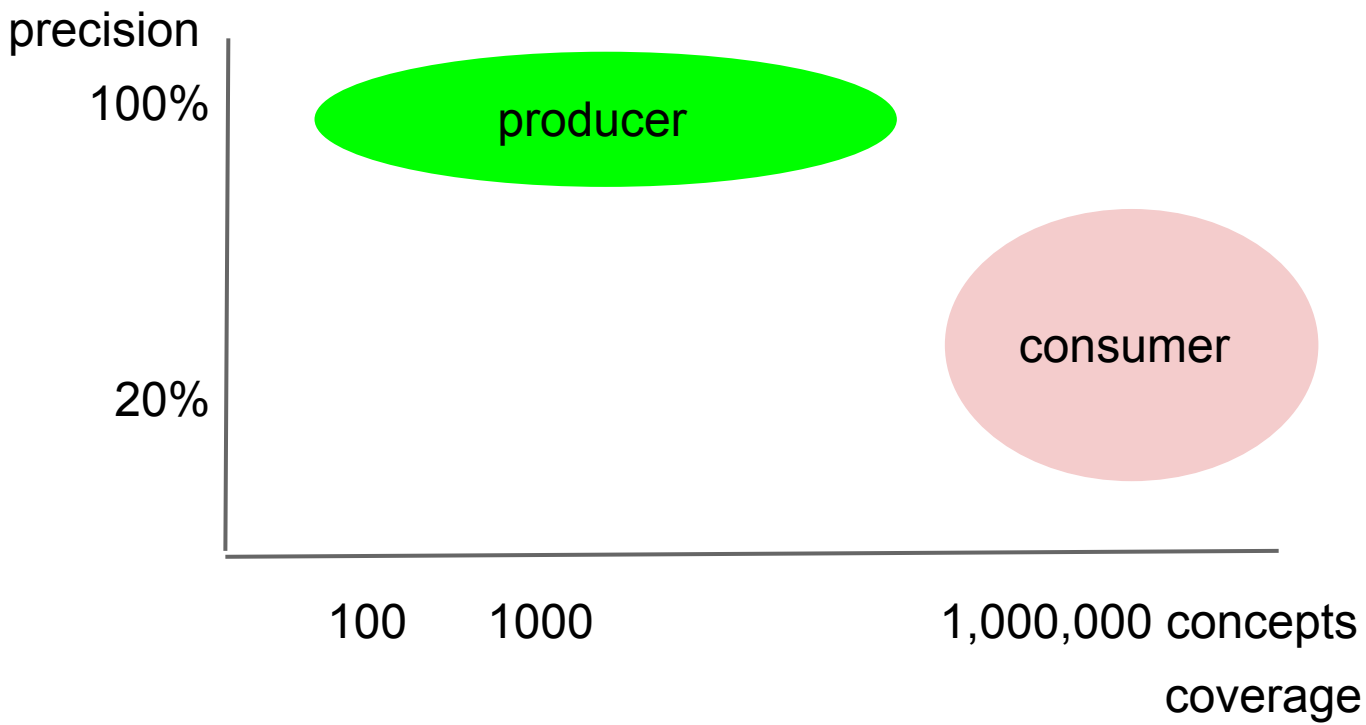
- browsing quality: to get an idea
- reader is responsible
+ translate anything

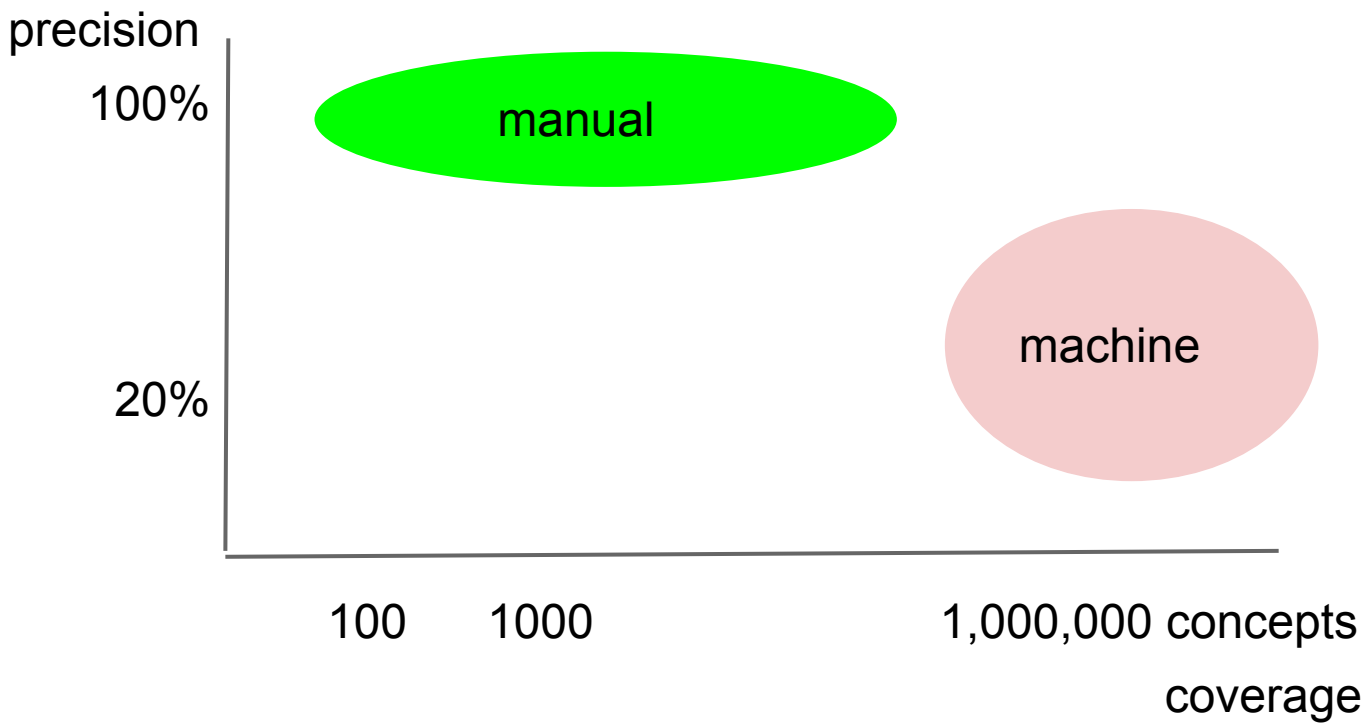
Consumer translator:

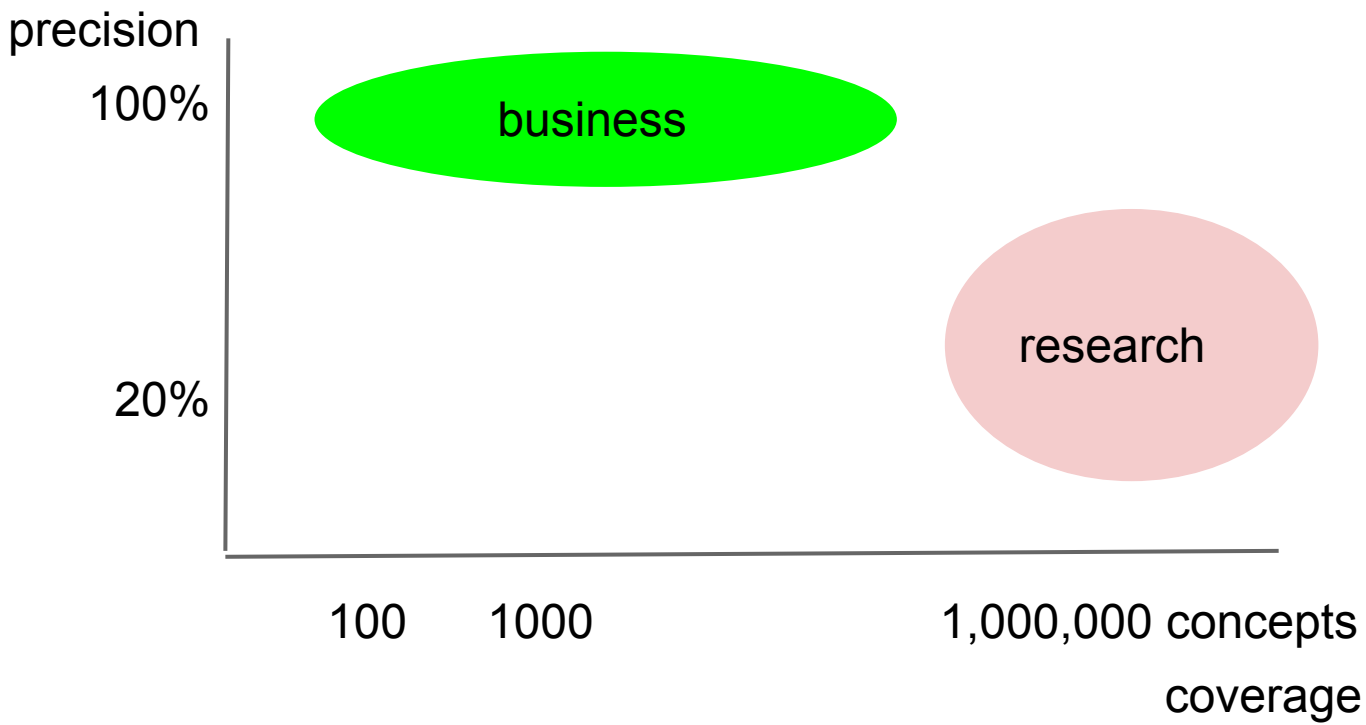
- browsing quality: to get an idea
- reader is responsible
+ translate anything

Producer translator:

+ publication quality: to get everything right
+ publisher is responsible
- translate my content


precision
100%
20%
producer
consumer
100
1000
1,000,000 concepts
coverage


precision
100%
20%
manual
machine
100
1000
1,000,000 concepts
coverage


precision
100%
20%
business
research
100
1000
1,000,000 concepts
coverage

A solution:

Data-Driven Documentation

digital Grammars

Language ely on.

2014 -

REMU

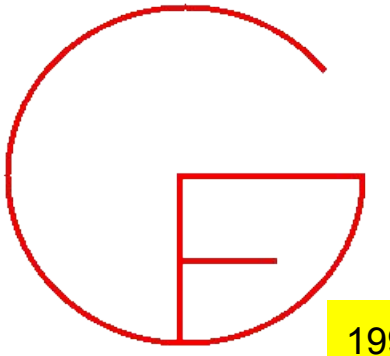
VR 2013 - 2017

MOLTO

EU 2010 - 2013

CLT

2009 - 2015



1998 -

Data

object	property	value
door	free width	121cm
walking area	tilt sideways	0.5%

Data

object	property	value
door	free width	121cm
walking area	tilt sideways	0.5%

Documentation: Eng

The free width of the door is 121cm.
The walking area tilts 0.5% sideways.

Data

object	property	value
door	free width	121cm
walking area	tilt sideways	0.5%

Documentation: Eng

The free width of the door is 121cm.
The walking area tilts 0.5% sideways.

Documentation: Swe

Dörrens fria bredd är 121cm.
Gångytan lutar 0.5% i sidled.

Data

object	property	value
door	free width	121cm
walking area	tilt sideways	0.5%

Documentation: Eng

The free width of the door is 121cm.
The walking area tilts 0.5% sideways.

Documentation: Swe

Dörrens fria bredd är 121cm.
Gångytan lutar 0.5% i sidled.

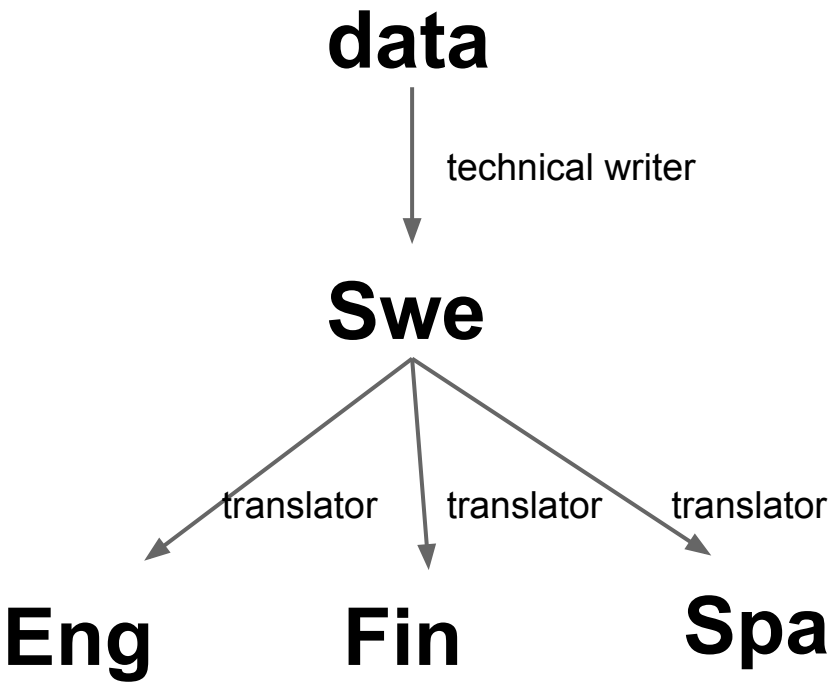
Documentation: Fin

Oven vapaa leveys on 121cm.
Kävelypinta kallistuu 0.5% siv…

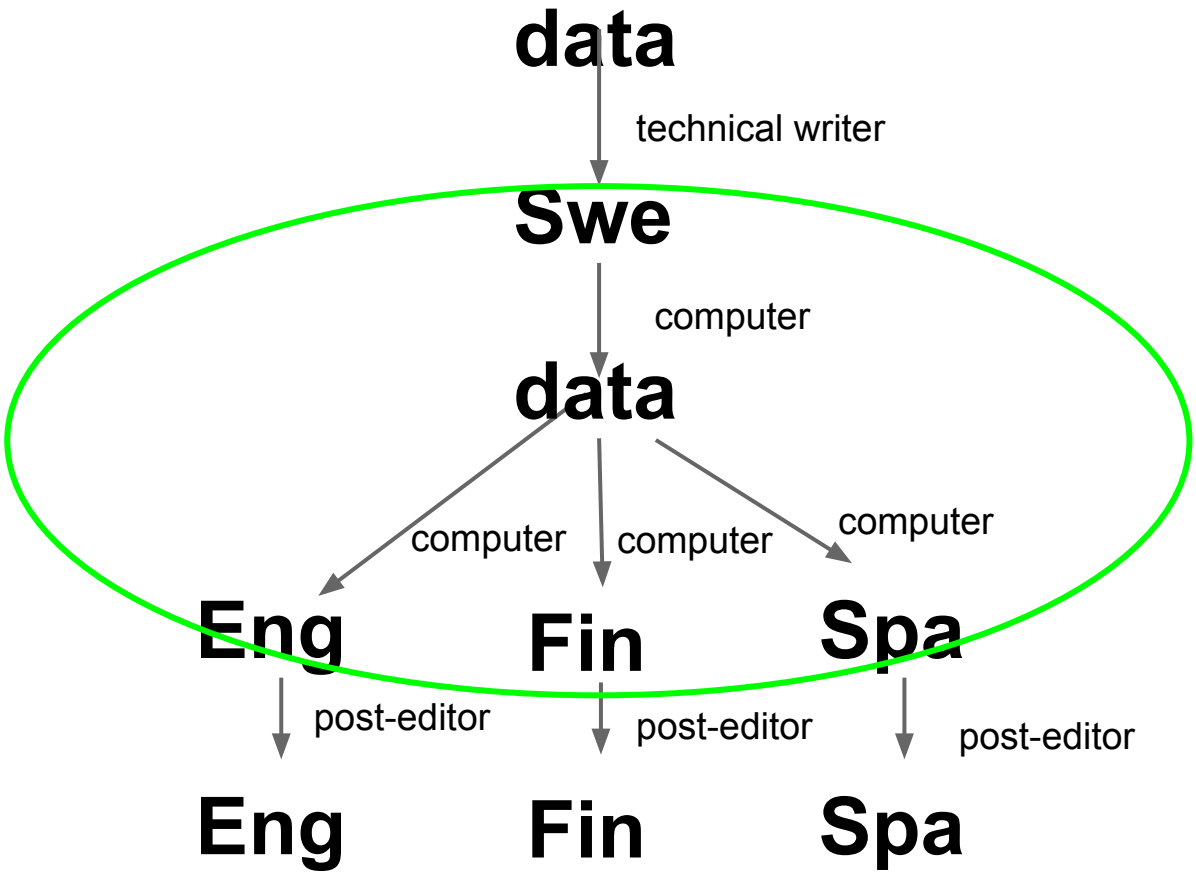
Documentation: Spa

El ancho libre de la puerta es de 121cm.
La zona peatonal se inclina 0.5% de lado

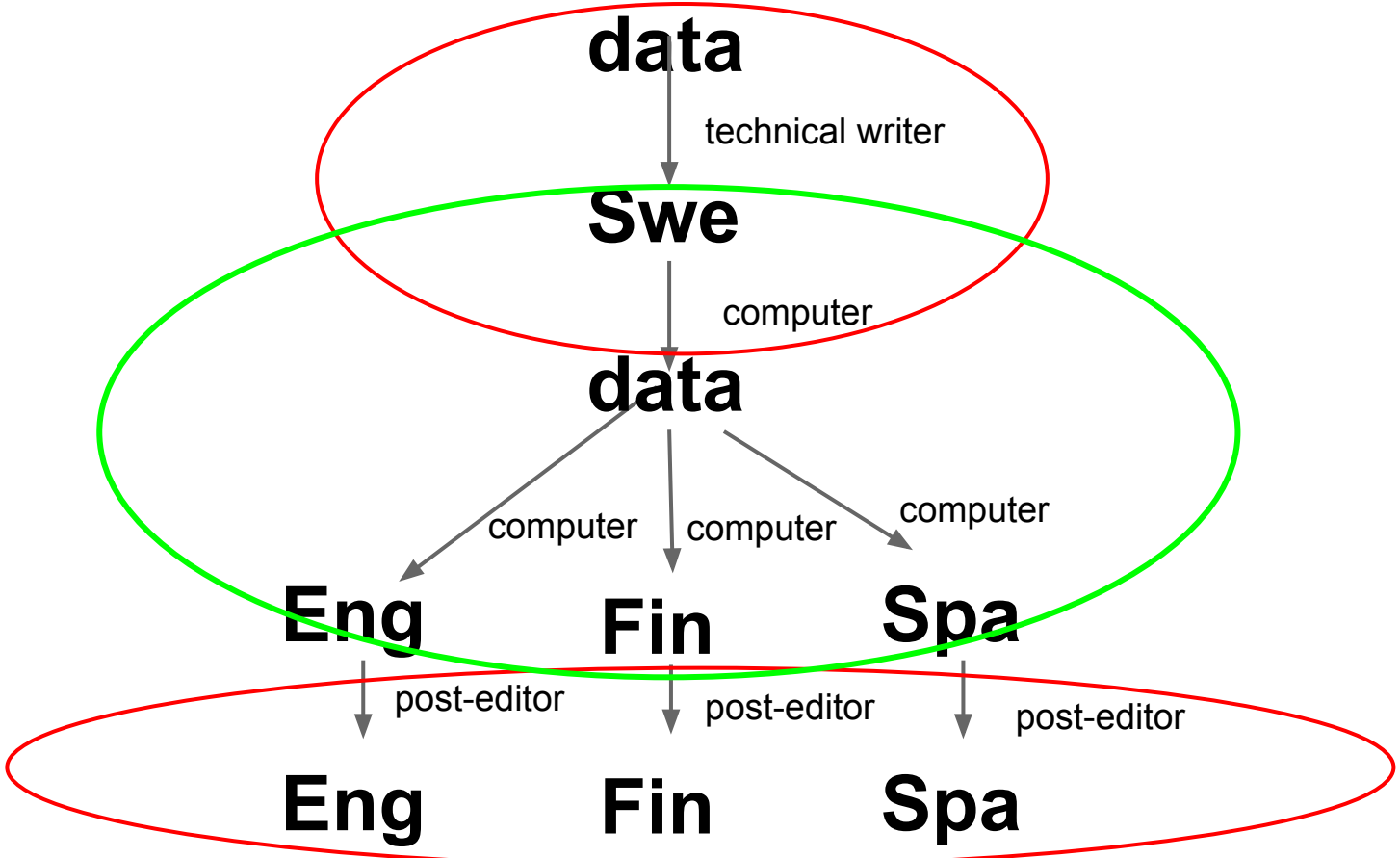
Traditional documentation



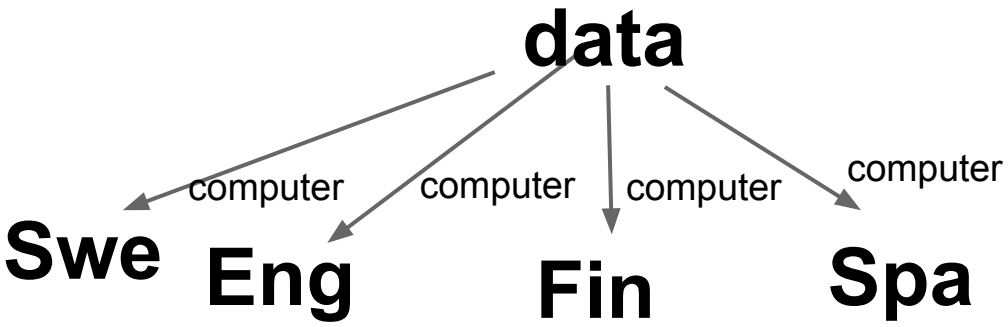
Introducing machine translation



To eliminate



Data-Driven Documentation



Advantages

Cheaper

Quicker

Better

More scalable

Cheaper

Initial cost: write the program

Later cost: mostly automatic

- post-editing at most 20% of human translation

Quicker

Translation in (almost) real time

The “almost” comes from

- new words
- post-editing need

Better

No accidental errors

Consistent terminology

More scalable

Adding new languages is easier:

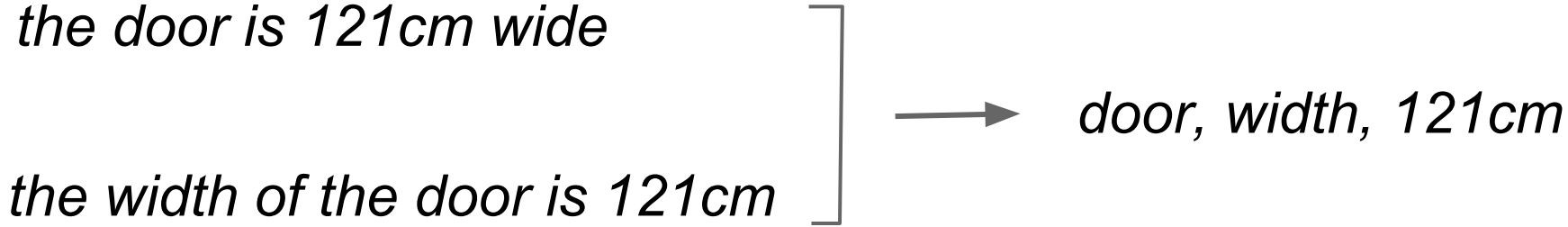
- data is common to all languages

Initial effort in vocabulary

- no work with the texts themselves

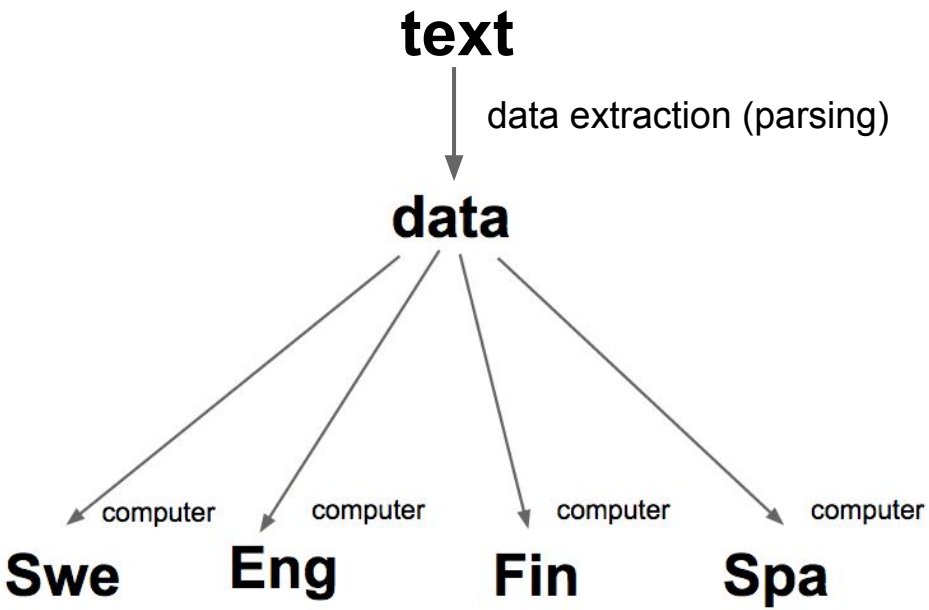
How to get there

1. Extract data from texts



2. Support input of new information as data

Translation =
Data Extraction +
Data-Driven Documentation



Technology:

GF = Grammatical Framework

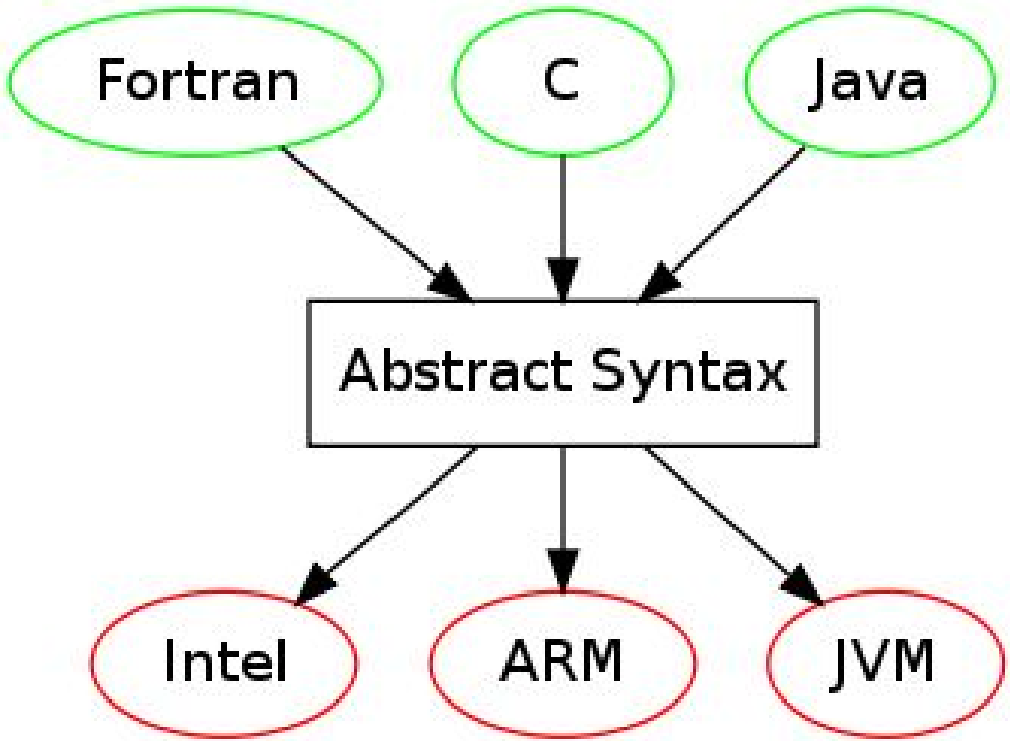
GF = Grammatical Framework

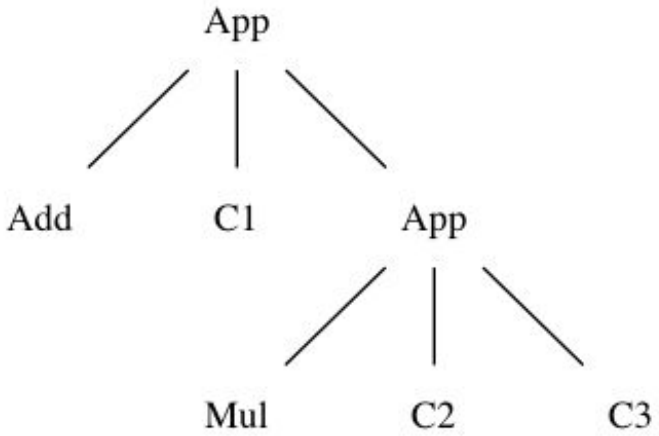
Xerox XRCE 1998, now open source

“Compiling natural language”

Library: 30 languages

Translation model: multi-source multi-target compiler



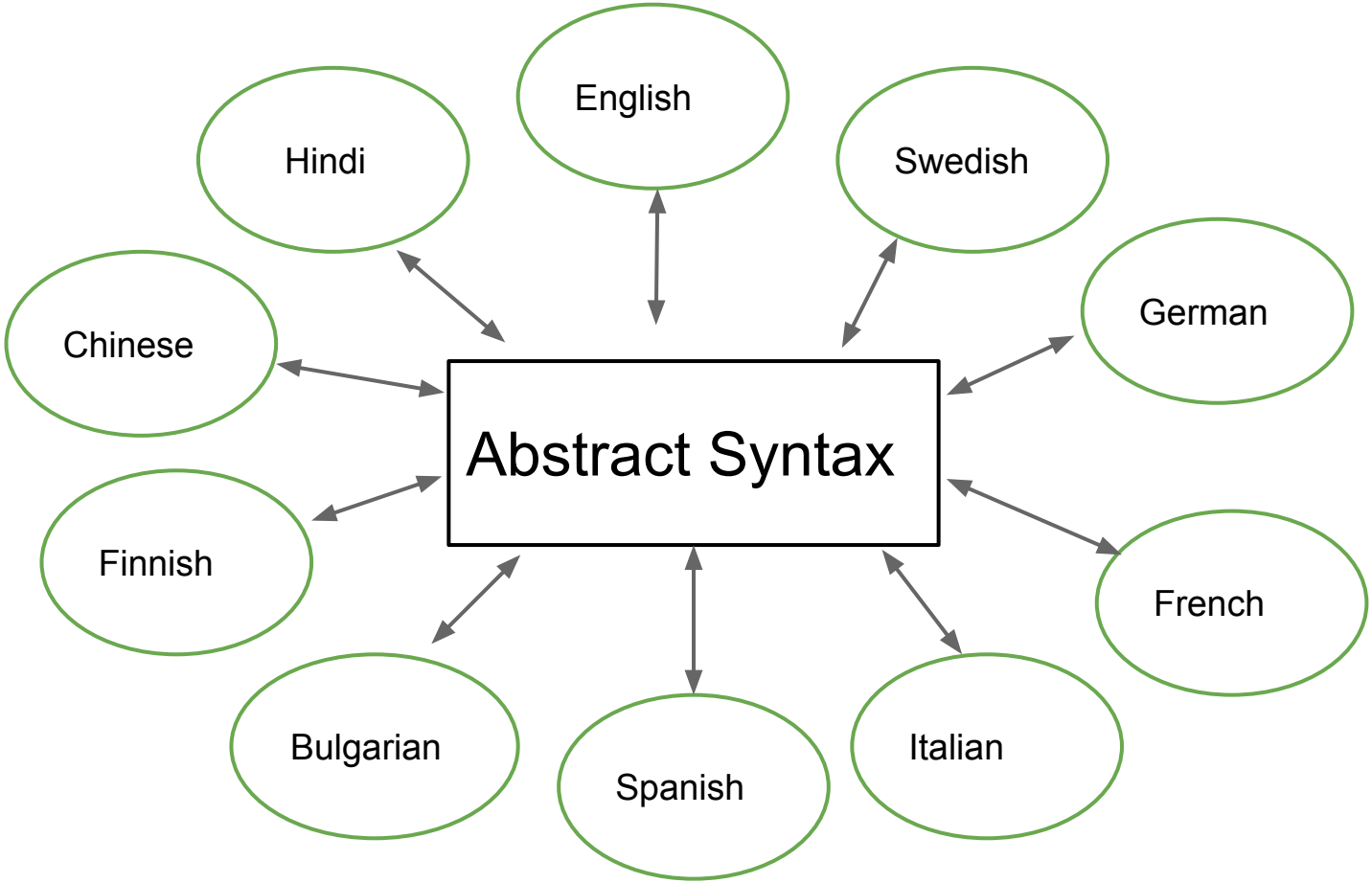


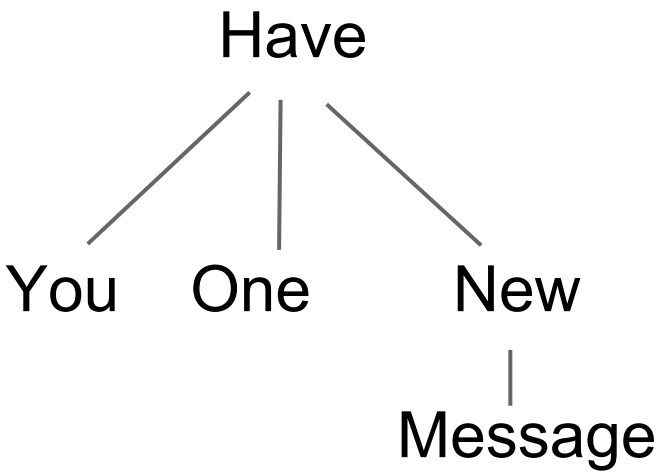
1 + 2 * 3

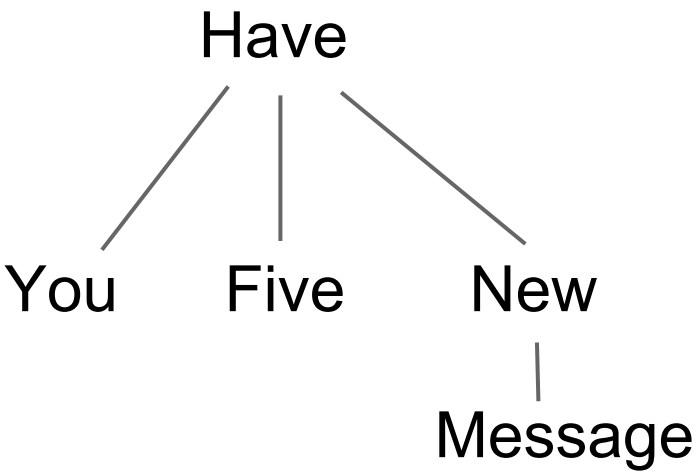
```
iconst_1
iconst_2
iconst_3
imul
iadd
```

```
(+ 1 (* 2 3))
```

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





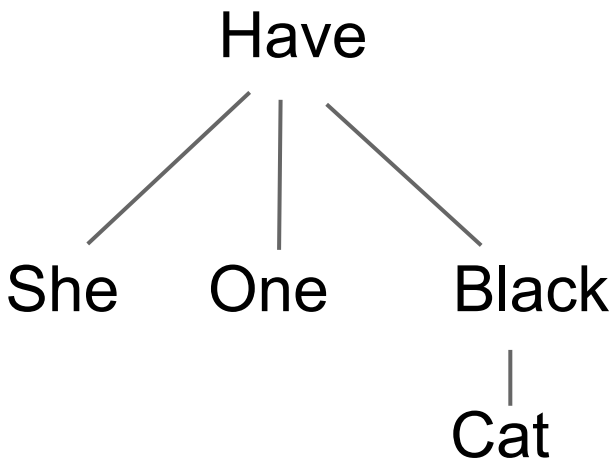


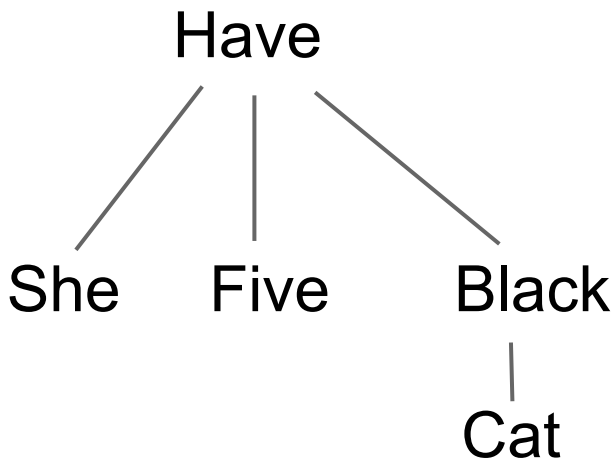
you have one new message

你 有 一 个 新 信 息

you have five new messages

你 有 五 个 新 信 息





she has one black cat
她 有 一 只 黑 猫

she has five black cats
她 有 五 只 黑 猫

Abstract and concrete syntax

Abstract syntax: semantic structure of data

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 ++ i
```

Abstract and concrete syntax

Abstract syntax

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

Concrete syntax, English

```
lin Have p n i = p ++ “have” ++ n ++ i
```

Concrete syntax, Chinese

```
lin Have p n i = p ++ “有” ++ n ++ i
```

Concrete syntax with parameters

Concrete syntax, English

```
lincat Number  = {s : Str ; n : Num}
lincat Item    = Num => Str
lincat Person  = {s : Str ; a : Agr}
lin Have p n i = p.s ++ have!p.a ++ n.s ++ i!n.n
```

Concrete syntax, Chinese

```
lincat Item    = {s : Str ; c : Str}
lin Have p n i = p ++ "有" ++ n ++ i.c ++ i.s
```

German inflection and word order

```
lin Have p n i =
  let
    subj = p.s ! Nom ;
    obj  = n.s ! i.g ! Acc ++ i.s ! n.n ! Acc ;
    verb = case p.a of {
      Ag Sg P1 => “habe” ;
      Ag Sg P2 => “hast” ;
      Ag Sg P3 => “hat” ;
      Ag Pl P2 => “habt” ;
      _ => “haben”
      }
   in
     case Ord of {
       Main => subj ++ verb ++ obj ;
       Sub  => subj ++ obj  ++ verb ;
       Inv  => obj  ++ subj ++ verb
       }
```

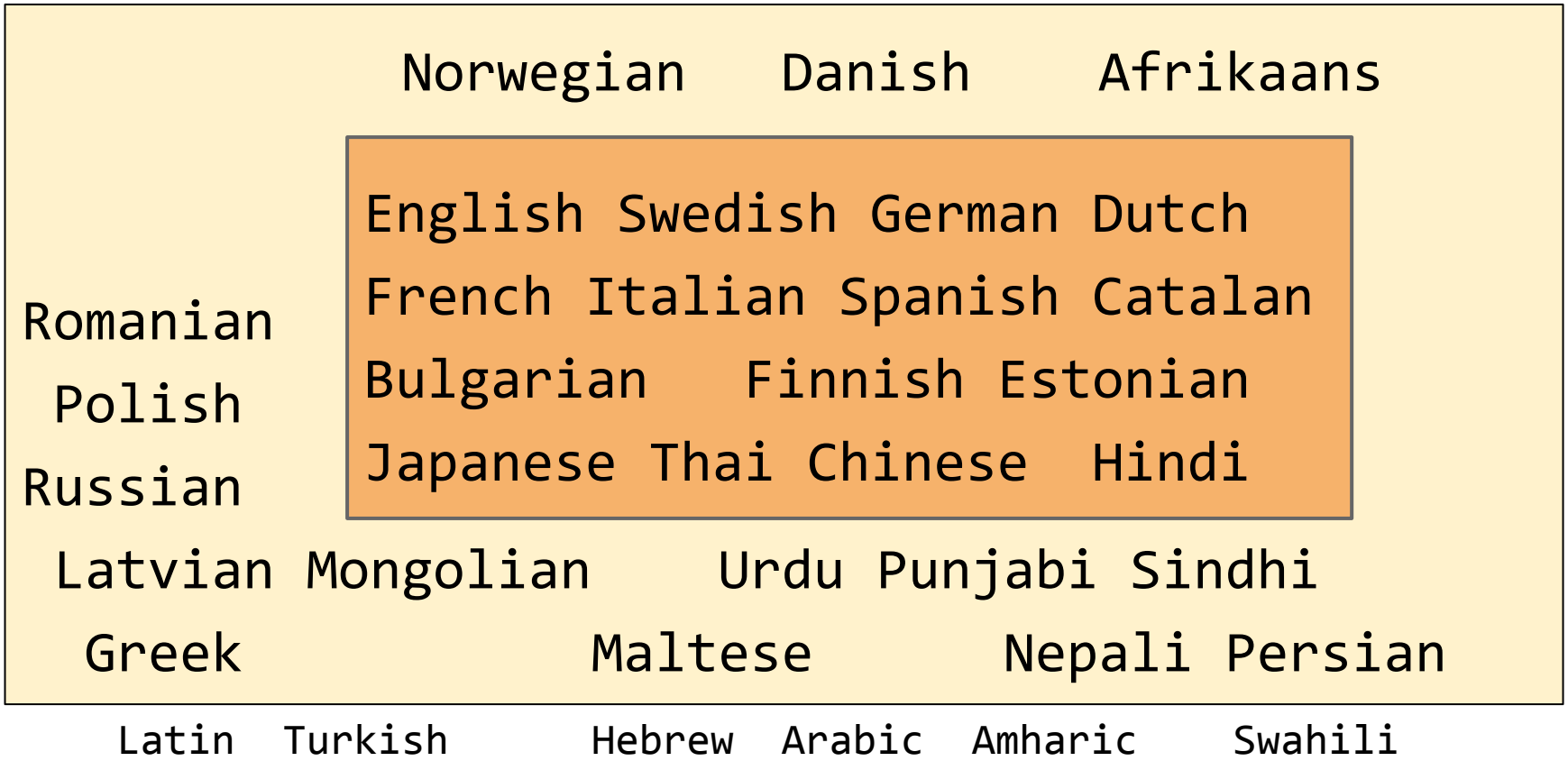
RGL = Resource Grammar Library

The standard library of GF

Takes care of linguistic details:

- morphology
- syntax

Makes GF productive and feasible


Norwegian Danish Afrikaans
English Swedish German Dutch
French Italian Spanish Catalan
Bulgarian Finnish Estonian
Japanese Thai Chinese Hindi
Romanian
Polish
Russian
Latvian Mongolian Urdu Punjabi Sindhi
Greek Maltese Nepali Persian
Latin Turkish Hebrew Arabic Amharic Swahili

The English rules with RGL

```
lin
 Have p n i = mkCl p have_V2 (mkNP n i)
 Message = mkN "message"
```

The Chinese rules with RGL

```
lin
 Have p n i = mkCl p have_V2 (mkNP n i)
 Message = mkN “信 息”
```

The German rules with RGL

```
lin
 Have p n i = mkCl p have_V2 (mkNP n i)
 Message = mkN "Nachricht" "Nachrichten" Fem
```

What is data?

Anything that can be represented as an abstract syntax in GF!

- relational data
- Semantic Web data (OWL, RDF)
- algebraic datatypes
- logical formulas
- dependent types and lambda calculus
- Constructive Type Theory

Some applications

Mathematical teaching material (WebALT)

Tourist phrasebook (MOLTO)

Formal specifications (Galois)

Patent query language (Ontotext)

Museum query language and texts (Ontotext)

Business models (Be Informed)

Medical examination journals (Lingsoft)

Speech commands in cars (Talkamatic)

Accessibilty database (Digital Grammars/TD)

2010-2013: MOLTO

Adam and Eve was painted by Albrecht Dürer in 1507. It measures 81 by 209 cm. This work is displayed at the Museo del Prado.

Adam and Eve a été peint par Albrecht Dürer en 1507. Il est de 81 sur 209 cm. Cette oeuvre est exposée au Musée du Prado.

Knowledge Base Results for "show everything about all paintings that are painted on canvas" (100 of many)

```
PREFIX painting: <http://spraakbanken.gu.se/rdf/owl/painting.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT distinct ?painting ?title ?author ?year ?length ?height ?museum
WHERE
{ ?painting rdf:type painting:Painting ;
rdfs:label ?title ;
painting:hasCurrentLocation ?museum;
painting:hasCreationDate ?date;
painting:hasDimension ?dim ;

painting:createdBy ?author . ?author rdfs:label ?painter .
?date painting:toTimePeriodValue ?year . ?dim painting:lengthValue ?length ;
painting:heightValue ?height . ?museum rdfs:label ?loc .

}
```

implies (mkProp (subset (Var2Set A) (Var2Set B))) (mkProp (notprsubset (V

- ако A е подмножество на B тогава B не е грозно подмножество на D
- si A és un subconjunt de B llavors B no és un subconjunt propi de D
- if A is a subset of B then B is not a proper subset of D en-US
- jos A on B:n osajoukko niin B ei ole D:n aito osajoukko
- si A est un sous-ensemble de B alors B n' est pas un sous-ensemble propre de
- wenn A eine Teilmenge von B ist dann ist B nicht eine echte Teilmenge von
- अगर A एक B का sub समुच्चय है तब B एक D का उचित sub समुच्चय नहीं है
- se A è un sottoinsieme di B quindi B non è un sottoinsieme proprio di D it-IT it-IT
- A \subseteq B \Rightarrow B \not\subset D

MOLTO Multilingual Online Translation

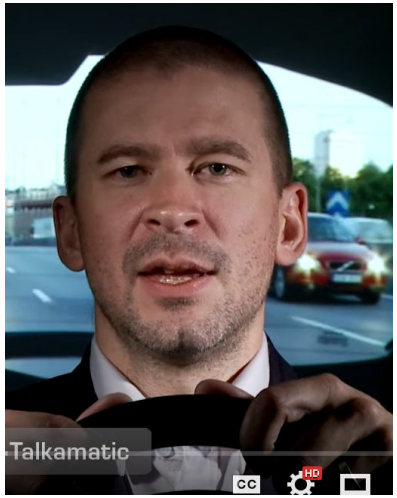
Non multa, sed multum not quantity but quality

FP7-ICT-247914

Talkamatic
FREE DIALOGUE

digital Grammars
Language technology to rely on.

GUO YUAN ZHONG CHUANG



next_membership_level_sys_answer silver (next_membership_points_sys_answer integer0_99_50)
test_mockup_travelChi: 您有五十个常旅客点符合会员条件，您现在是在伦敦.
test_mockup_travelDut: je hebt vijftig punten nodig om het zilveren niveau te bereiken
test_mockup_travelEng: you need fifty points to reach silver level
test_mockup_travelFin: sinä tarvitset viisikymmentä pistettä päästäksesi hopeatasolle
test_mockup_travelFre: tu as besoin de cinquante points pour atteindre le niveau argent
test_mockup_travelGer: Sie brauchen fünfzig Punkte um das Silberniveau zu erreichen
test_mockup_travelIta: avete bisogno di cinquanta punti per raggiungere il livello argento
test_mockup_travelSpa: necesitas cincuenta puntos para llegar al nivel plata

TitleParagraph DefinitionTitle
DefPredParagraph type_Sort A_Var contractible_Pred (ExistCalledProp a_Var (ExpSort (VarExp A_Var)) (FunInd centre_of_contraction_Fun) (ForAllProp (BaseVar x_Var) (ExpSort (VarExp A_Var)) (ExpProp (equalExp (VarExp a_Var) (VarExp x_Var)))))
FormatParagraph EmptyLineFormat
TitleParagraph DefinitionTitle
DefPredParagraph (mapSort (mapExp (VarExp A_Var) (VarExp B_Var))) f_Var equivalence_Pred (ForAllProp (BaseVar y_Var) (ExpSort (VarExp B_Var)) (PredProp contractible_Pred (AliasInd (AppFunItInd fiber_Fun) (FunInd (ExpFun (ComprehensionExp x_Var (VarExp A_Var) (equalExp (AppExp f_Var (VarExp x_Var)) (VarExp y_Var))))))))
DefPropParagraph (ExpProp (equivalenceExp (VarExp A_Var) (VarExp B_Var))) (ExistSortProp (equivalenceSort (mapExp (VarExp A_Var) (VarExp B_Var))))
FormatParagraph EmptyLineFormat
TitleParagraph LemmaTitle
TheoremParagraph (ForAllProp (BaseVar A_Var) type_Sort (PredProp equivalence_Pred (AliasInd (FunInd identity_map_Fun) (FunInd (ExpFun (DefExp (identityMapExp (VarExp A_Var)) (TypedExp (BaseExp (lambdaExp x_Var (VarExp A_Var) (VarExp x_Var))) (mapExp (VarExp A_Var) (VarExp A_Var)))))))))
FormatParagraph EmptyLineFormat
TitleParagraph ProofTitle
AssumptionParagraph (ConsAssumption (ForAssumption y_Var (ExpSort (VarExp A_Var)) (LetAssumption (FunInd (ExpFun (DefExp (fiberExp (VarExp y_Var) (VarExp A_Var)) (ComprehensionExp x_Var (VarExp A_Var) (equalExp (VarExp x_Var) (VarExp y_Var)))))) (AppFunItInd (fiberWrt_Fun (FunInd (ExpFun (identityMapExp (VarExp A_Var))))))) (BaseAssumption (LetExpAssumption (barExp (VarExp y_Var)) (TypedExp (BaseExp (pairExp (VarExp y_Var) (reflexivityExp (VarExp A_Var) (VarExp y_Var)))) (fiberExp (VarExp y_Var) (VarExp A_Var)))))
ConclusionParagraph (AsConclusion (ForAllProp (BaseVar y_Var) (ExpSort (VarExp A_Var)) (ExpProp (equalExp (pairExp (VarExp y_Var) (reflexivityExp (VarExp A_Var) (VarExp y_Var))) (VarExp y_Var)))) (ApplyLabelConclusion id_induction_Label (ConsInd (FunInd (ExpFun (VarExp y_Var))) (ConsInd (FunInd (ExpFun (TypedExp (BaseExp (VarExp x_Var)) (VarExp A_Var)))) (ConsInd (FunInd (ExpFun (TypedExp (BaseExp (VarExp z_Var)) (idPropExp (VarExp x_Var) (VarExp y_Var))))) BaseInd)))
(DisplayExpProp (equalExp (pairExp (VarExp x_Var) (VarExp z_Var)) (VarExp y_Var)))))
ConclusionSoThatParagraph (ForConclusion (BaseVar y_Var) (ExpSort (VarExp A_Var)) (A
BaseInd) (ExpProp (equalExp (VarExp u_Var) (VarExp y_Var))))) (PredProp contractible_Pr
ConclusionParagraph (PropConclusion (PredProp equivalence_Pred (FunInd (ExpFun (Type
QEDParagraph

Définition: Un type A est contractible, s'il existe u
de contraction, tel que pour tous les $x : A$, $a = x$.

Définition: Une application $f : A \to B$ est une éq
les $y : B$, sa fibre, $\{x : A \mid fx = y\}$, est contractible. N
existe une équivalence $A \to B$.

Lemme: Pour tout type A, l'identité, $1_A := \lambda_{x:}$
équivalence.

Démonstration: Pour tout $y : A$, soit $\{y\}_A := \{$
par rapport de 1_A et soit $\bar{y} := (y, r_A y) : \{y\}_A$. Comm
$(y, r_A y) = y$, nous pouvons appliquer Id-induction sur
pour obtenir que

$$(x, z) = y$$

. Donc, pour les $y : A$, nous pouvons appliquer Σ -élimination sur $u : \{y\}_A$ pour obtenir que $u = y$, de façon que $\{y\}_A$ soit contractible. Alors, $1_A : A \to A$ est une équivalence. □

Definition: A type A is contractible, if there is $a : A$, called the center of contraction, such that for all $x : A$, $a = x$.

Definition: A map $f : A \to B$ is an equivalence, if for all $y : B$, its fiber, $\{x : A \mid fx = y\}$, is contractible. We write $A \simeq B$, if there is an equivalence $A \to B$.

Lemma: For each type A, the identity map, $1_A := \lambda_{x:A} x : A \to A$, is an equivalence.

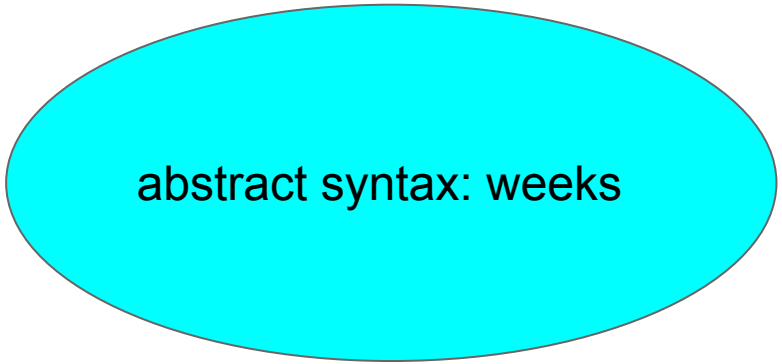
Proof: For each $y : A$, let $\{y\}_A := \{x : A \mid x = y\}$ be its fiber with respect to 1_A and let $\bar{y} := (y, r_A y) : \{y\}_A$. As for all $y : A$, $(y, r_A y) = y$, we may apply Id-induction on y, $x : A$ and $z : (x = y)$ to get that

$$(x, z) = y$$

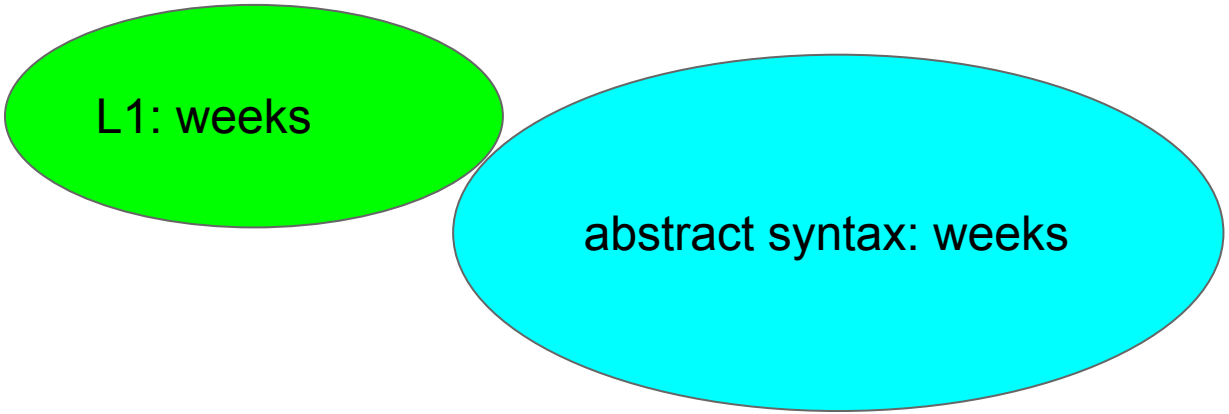
. Hence, for $y : A$, we may apply Σ -elimination on $u : \{y\}_A$ to get that $u = y$, so that $\{y\}_A$ is contractible. Thus, $1_A : A \to A$ is an equivalence. □

https://github.com/GrammaticalFramework/gf-contrib/tree/master/homotopy-typetheory

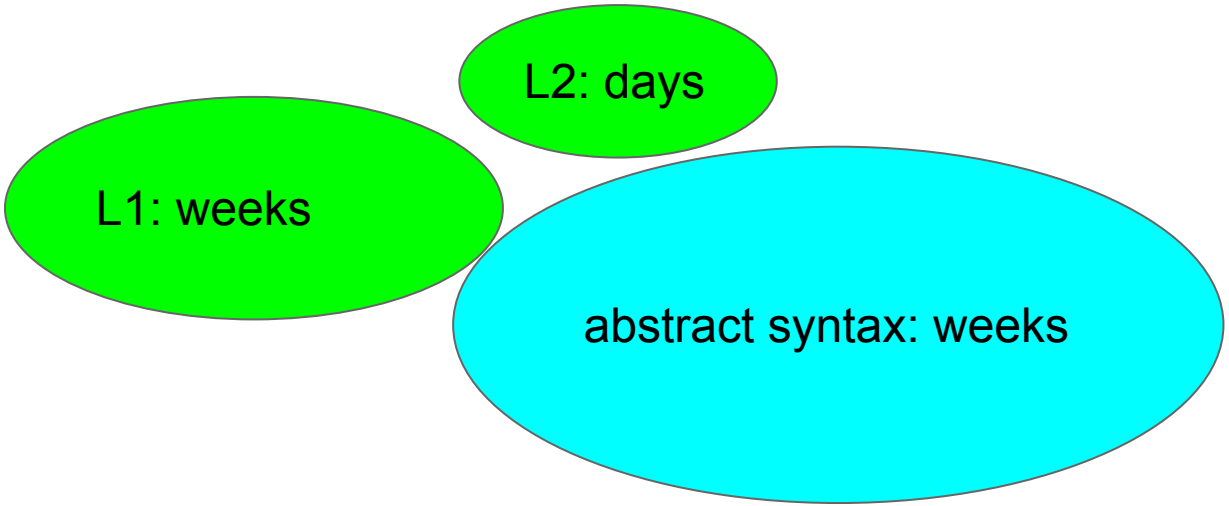
GF grammar building effort



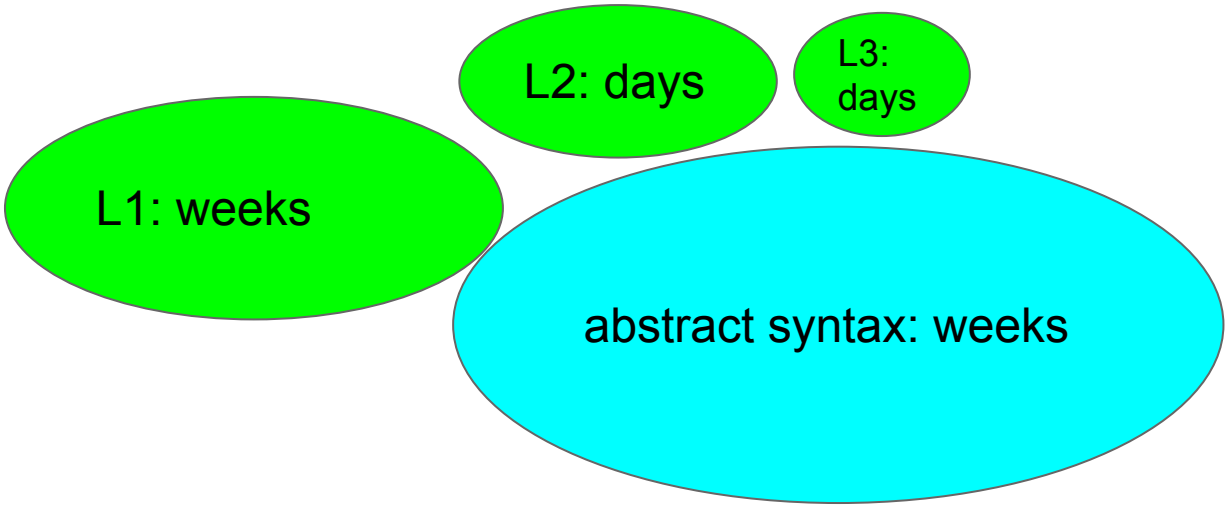
GF grammar building effort



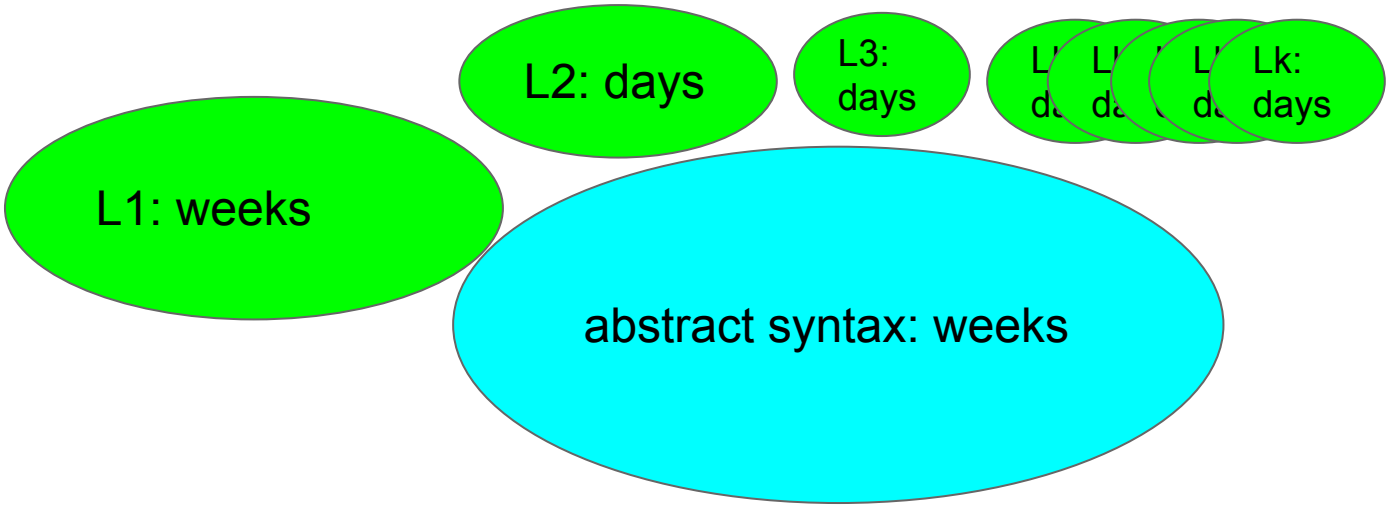
GF grammar building effort



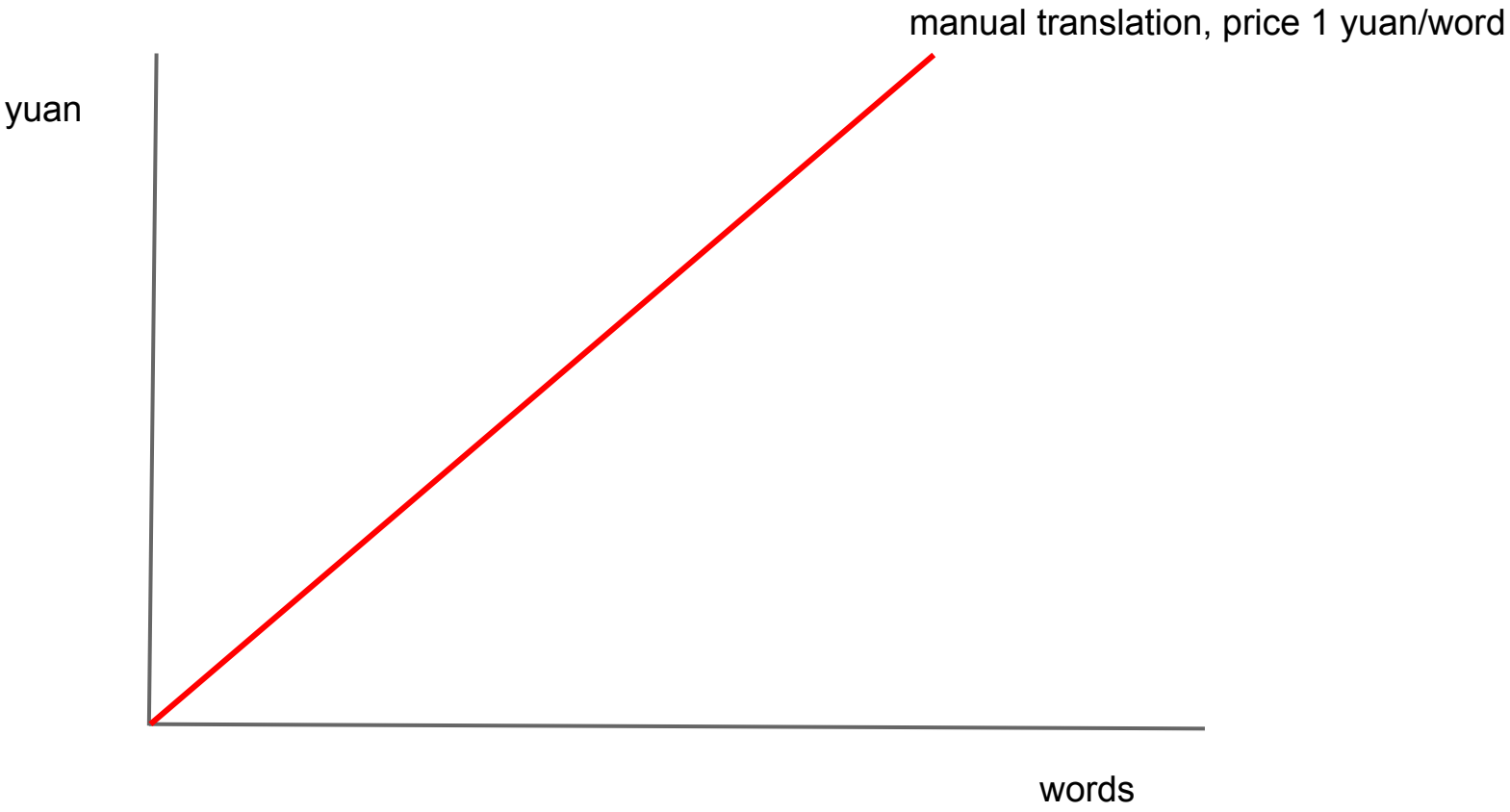
GF grammar building effort



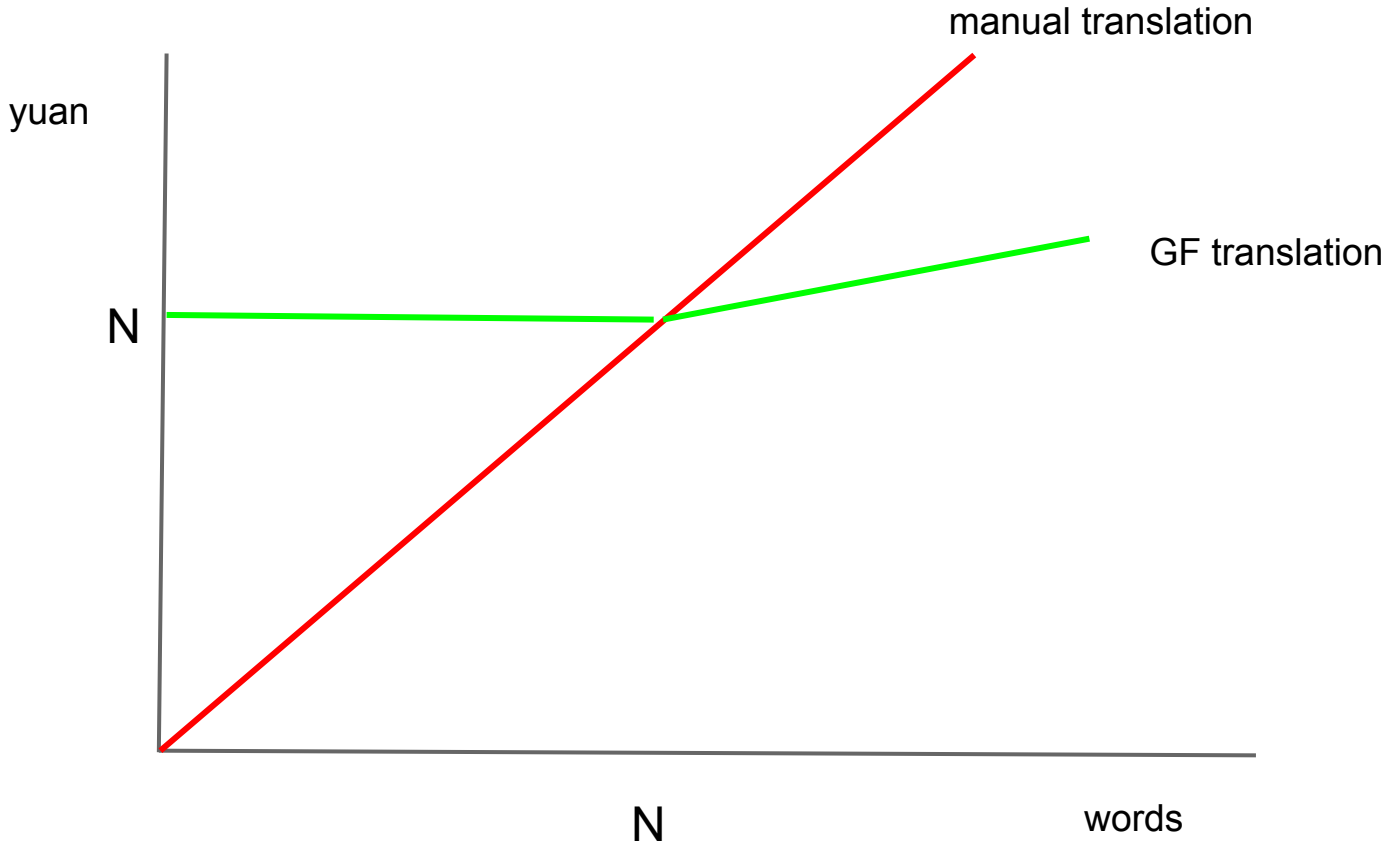
GF grammar building effort



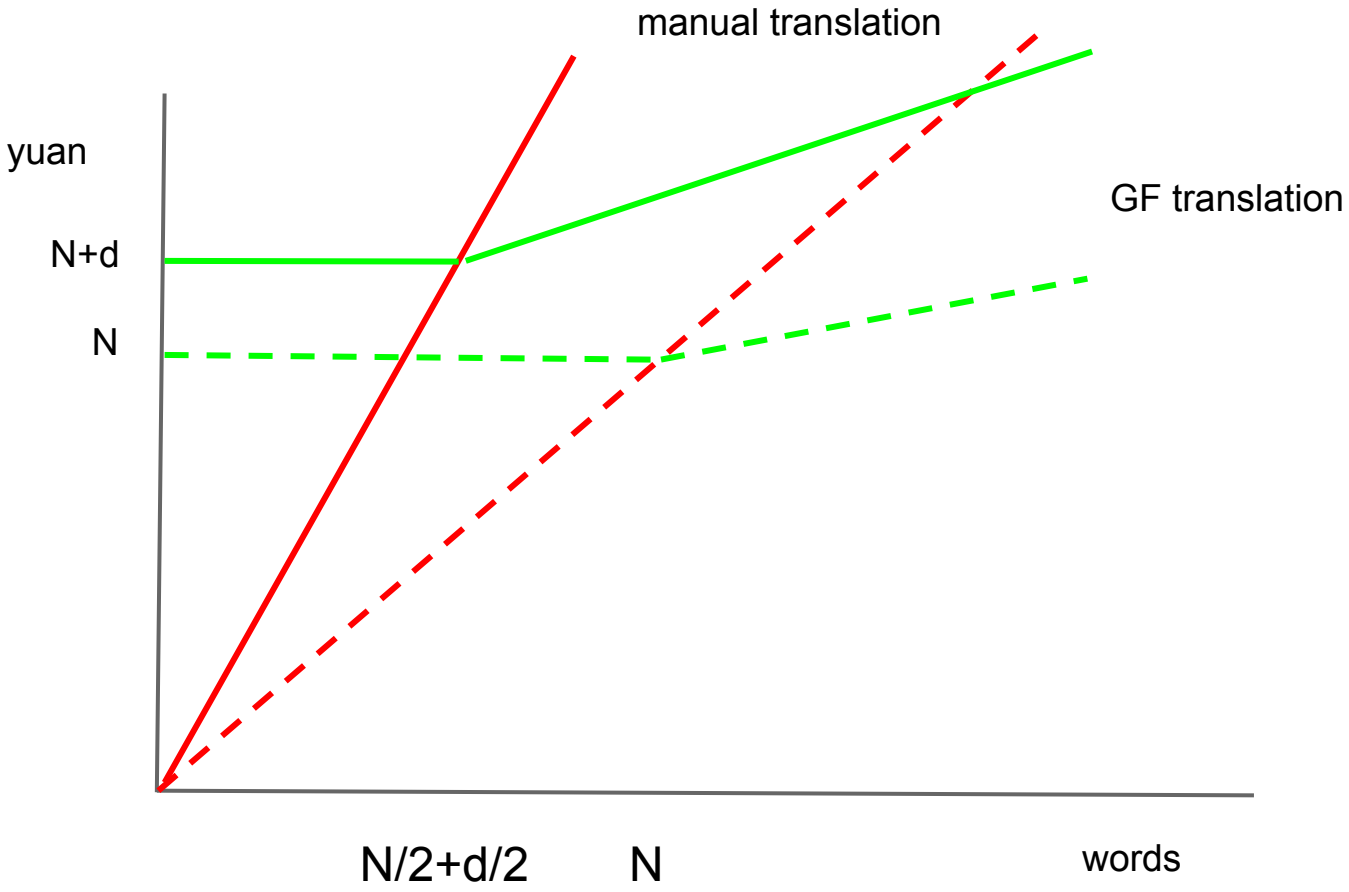
Price of translation, 1 target language



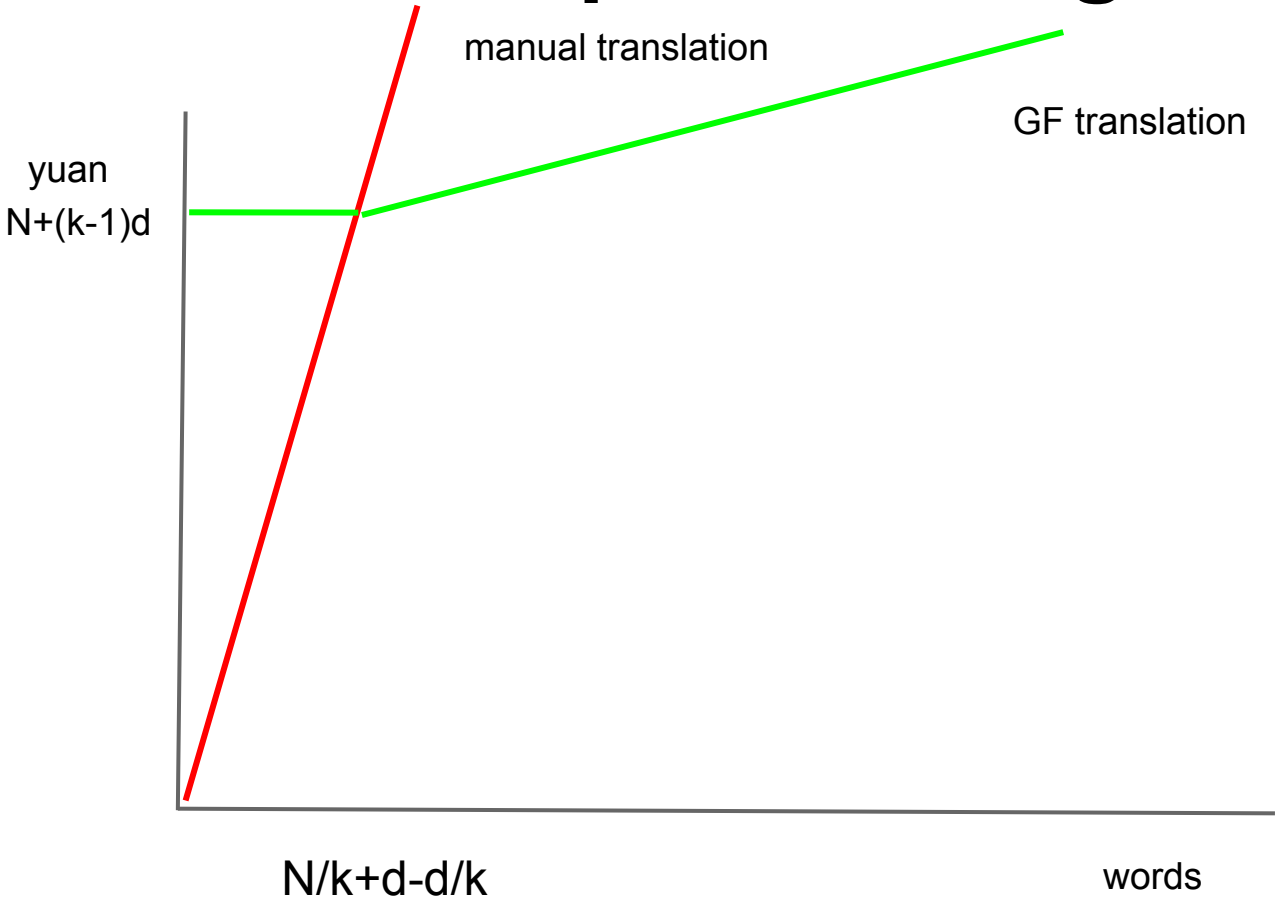
Break-even point, 1 target language

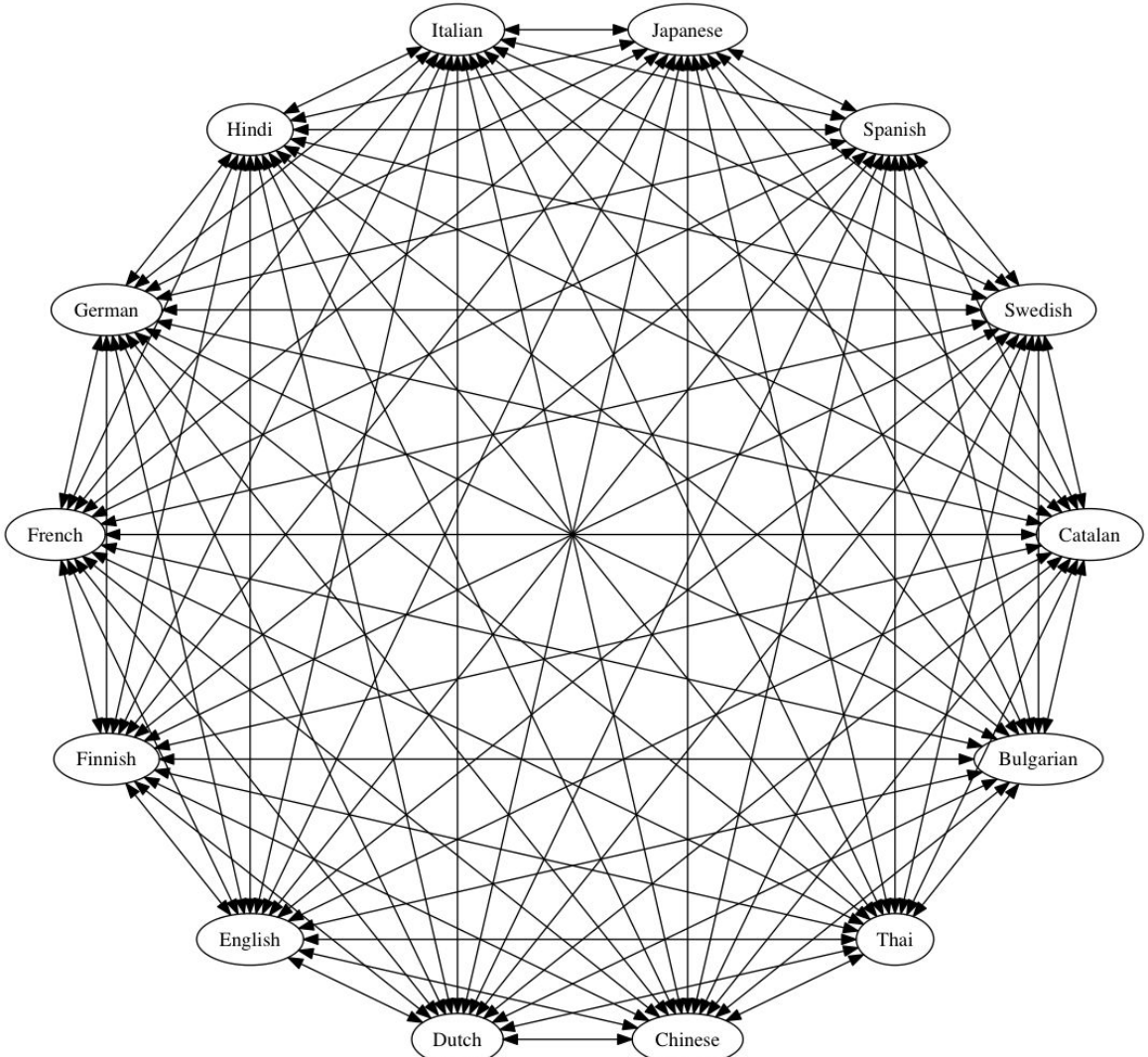


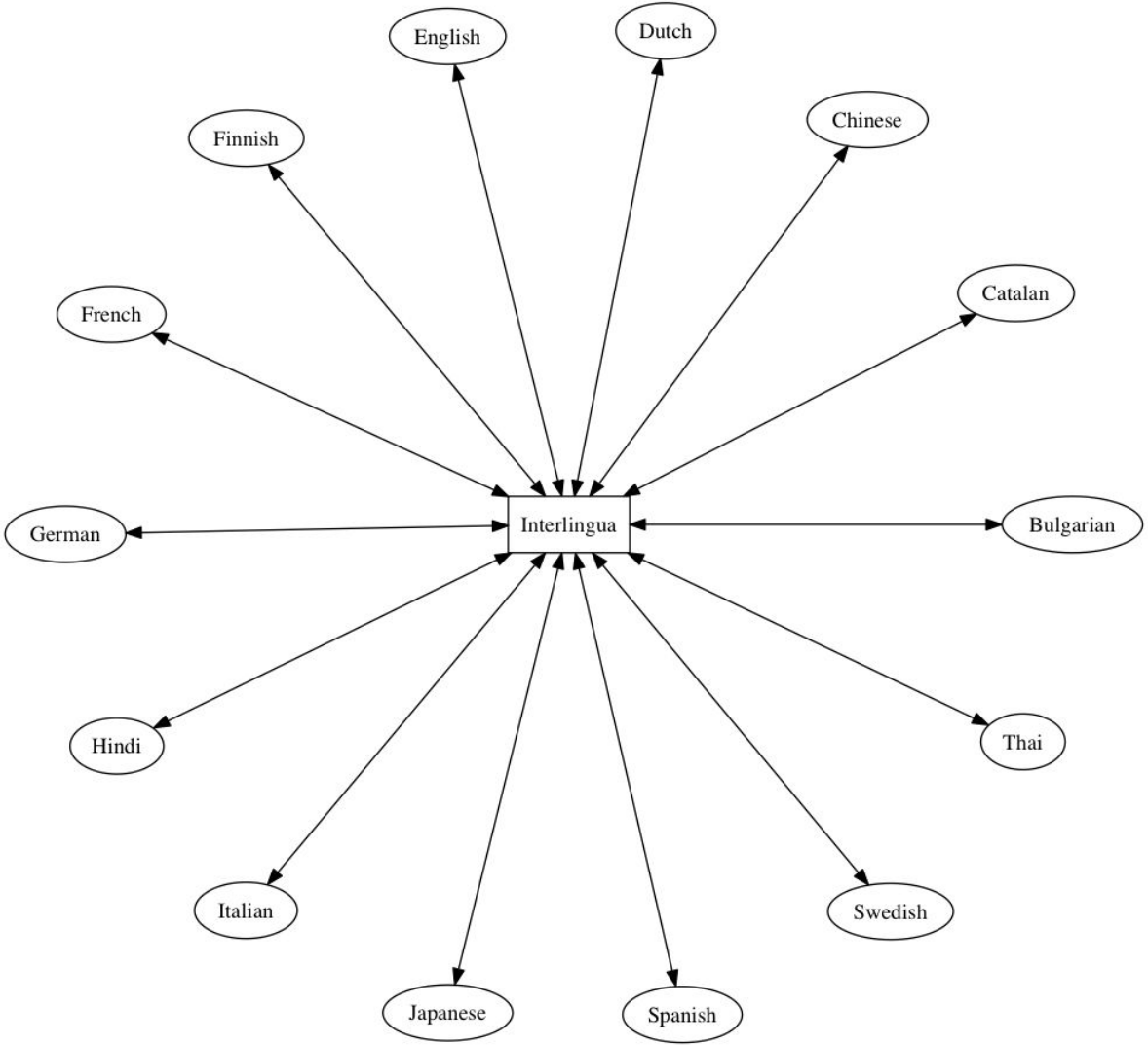
Break-even point, 2 target languages



Break-even point, k target languages




Italian
Japanese
Hindi
Spanish
German
Swedish
French
Catalan
Finnish
Bulgarian
English
Thai
Dutch
Chinese

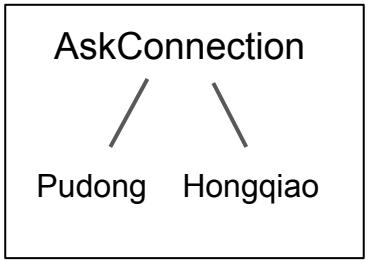

English
Dutch
Finnish
Chinese
French
Catalan
German
Interlingua
Bulgarian
Hindi
Thai
Italian
Swedish
Japanese
Spanish

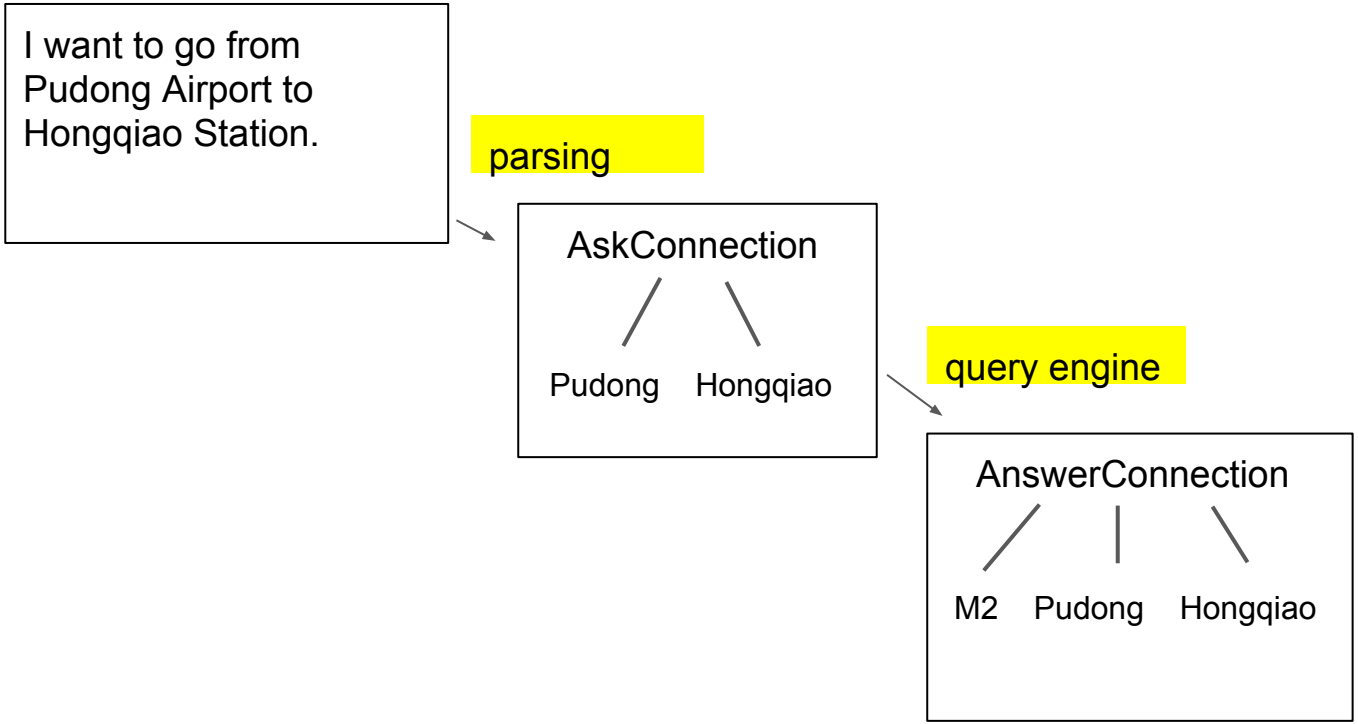
Data-Driven Question Answering

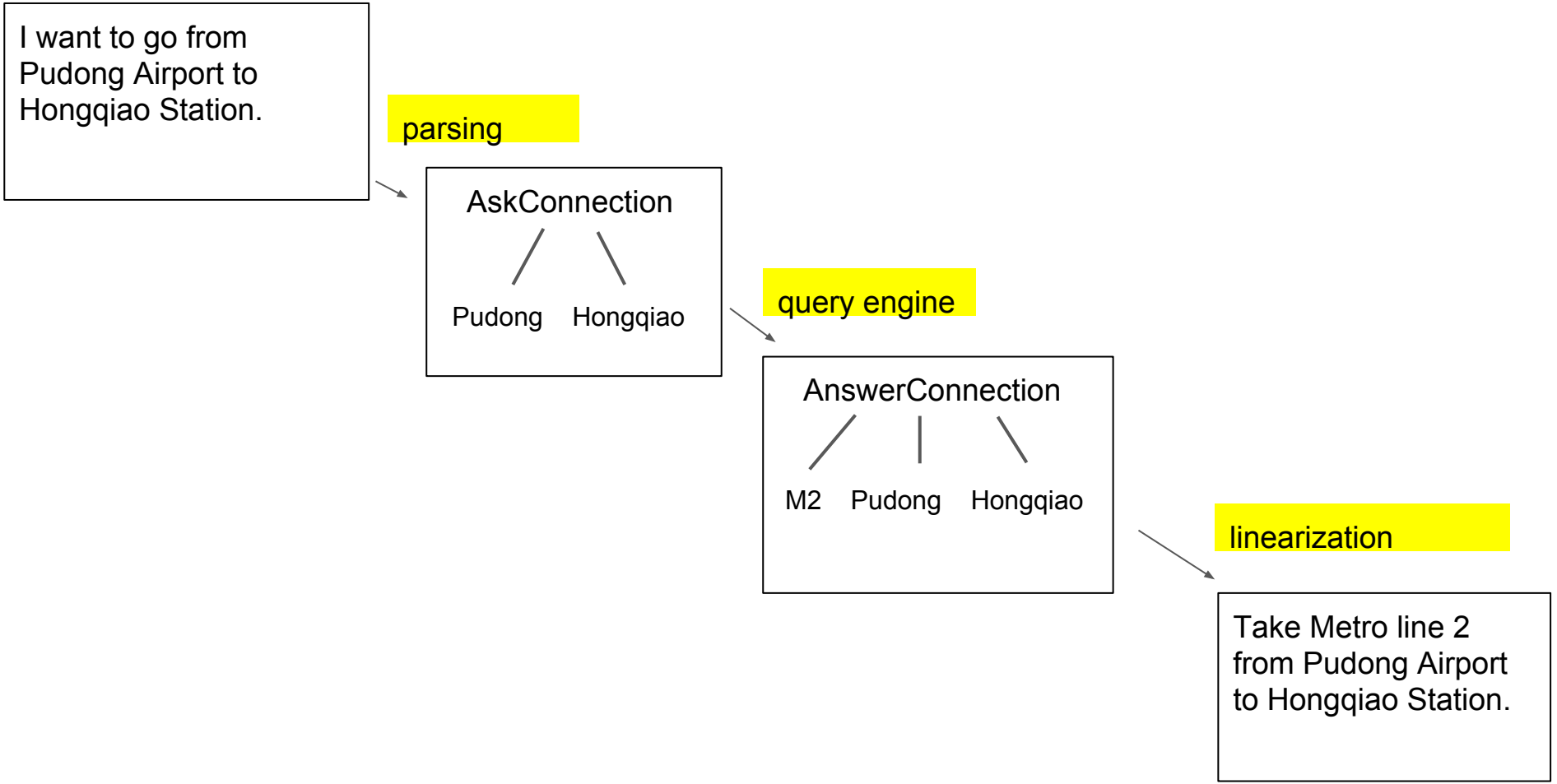
I want to go from Pudong Airport to Hongqiao Station.

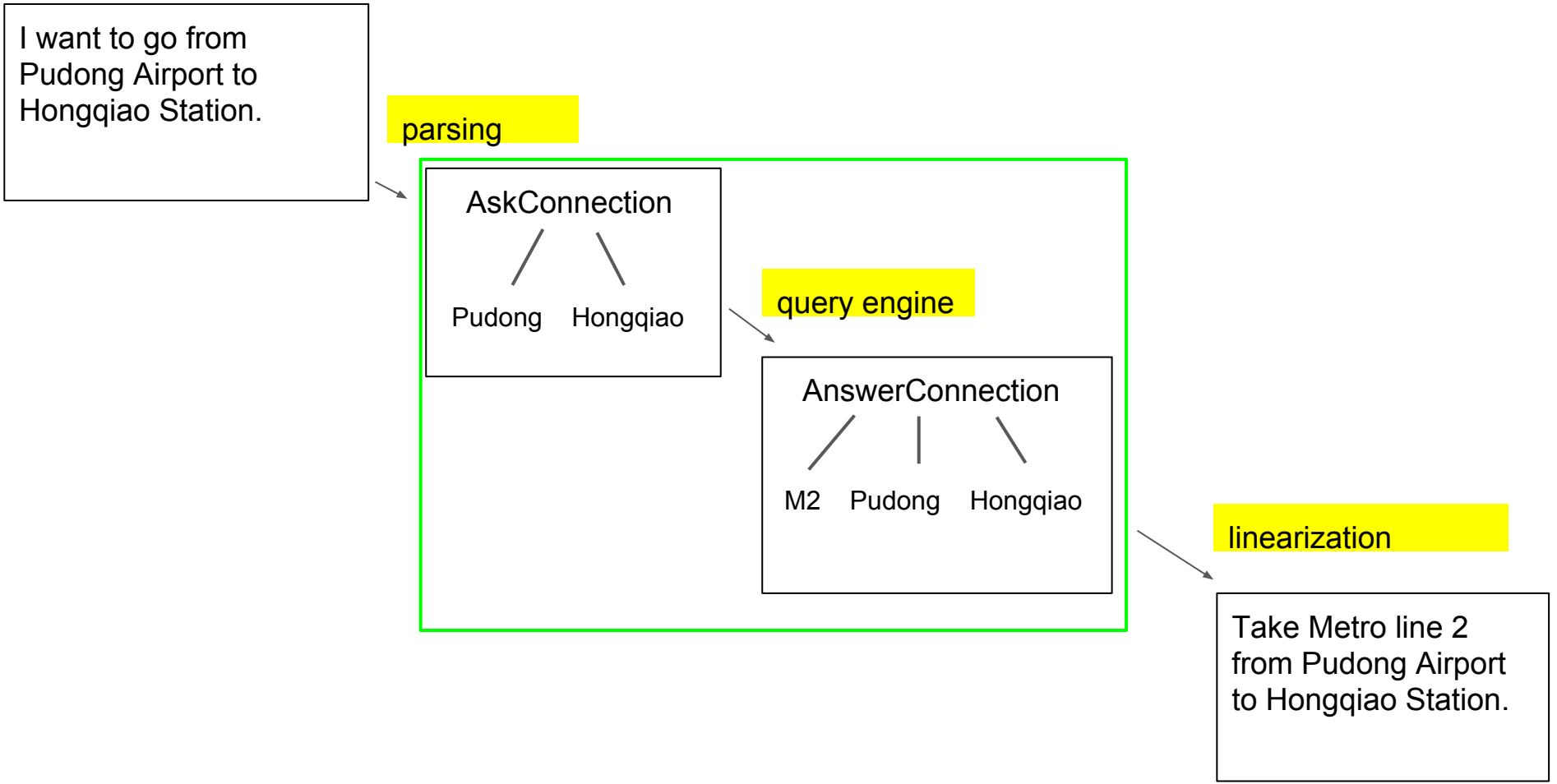
I want to go from Pudong Airport to Hongqiao Station.

parsing

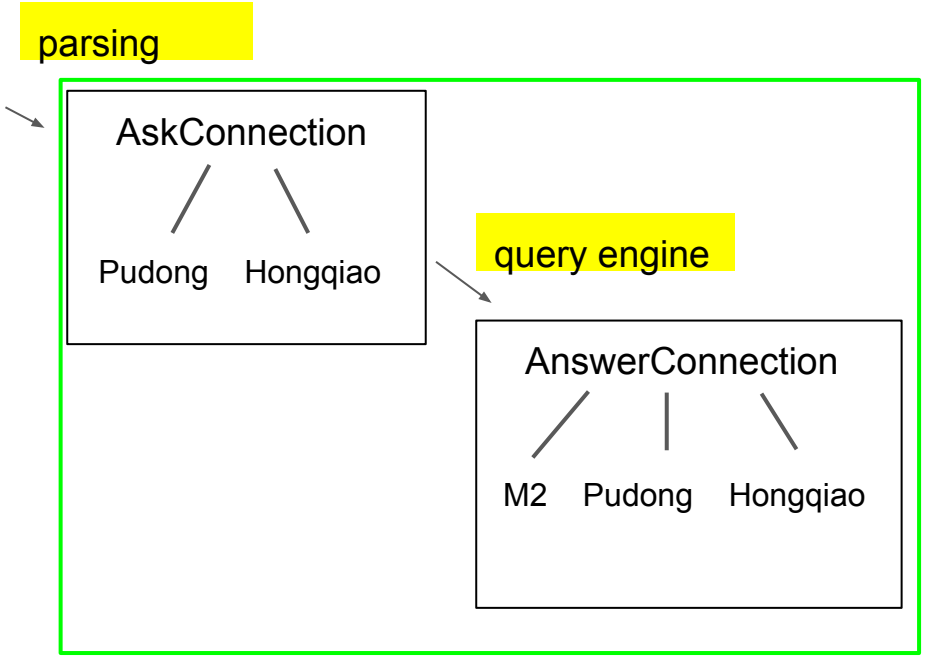



I want to go from Pudong Airport to Hongqiao Station.
parsing
AskConnection
Pudong
Hongqiao
query engine
AnswerConnection
M2
Pudong
Hongqiao


I want to go from Pudong Airport to Hongqiao Station.
parsing
AskConnection
Pudong
Hongqiao
query engine
AnswerConnection
M2
Pudong
Hongqiao
linearization
Take Metro line 2 from Pudong Airport to Hongqiao Station.


I want to go from Pudong Airport to Hongqiao Station.
parsing
AskConnection
Pudong
Hongqiao
query engine
AnswerConnection
M2
Pudong
Hongqiao
linearization
Take Metro line 2 from Pudong Airport to Hongqiao Station.

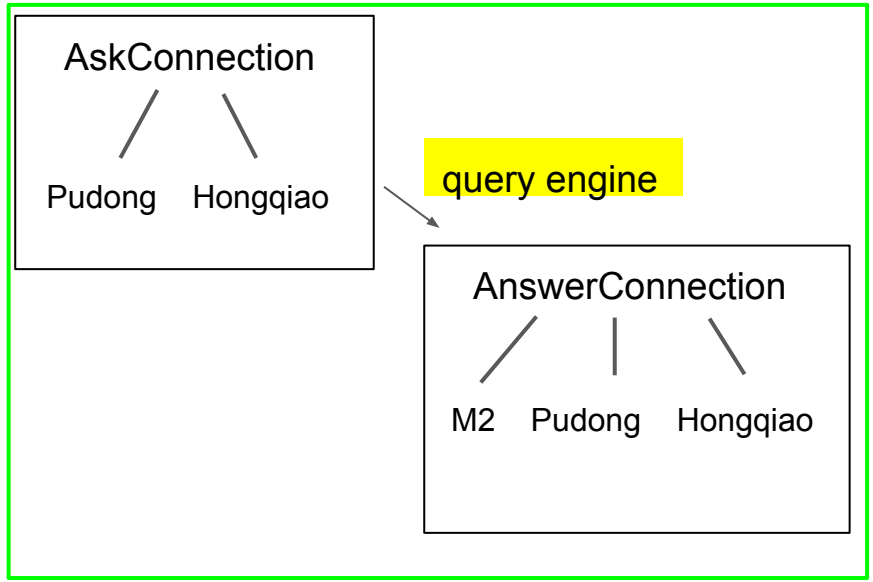
从 浦 东 机 场 到 虹 桥 站 怎 么 走 ？



在 浦 东 坐 2 号 地 铁 到 虹 桥 站

Kuinka pääsee Pudongin lentokentältä Hongqiao-asemalle?

parsing



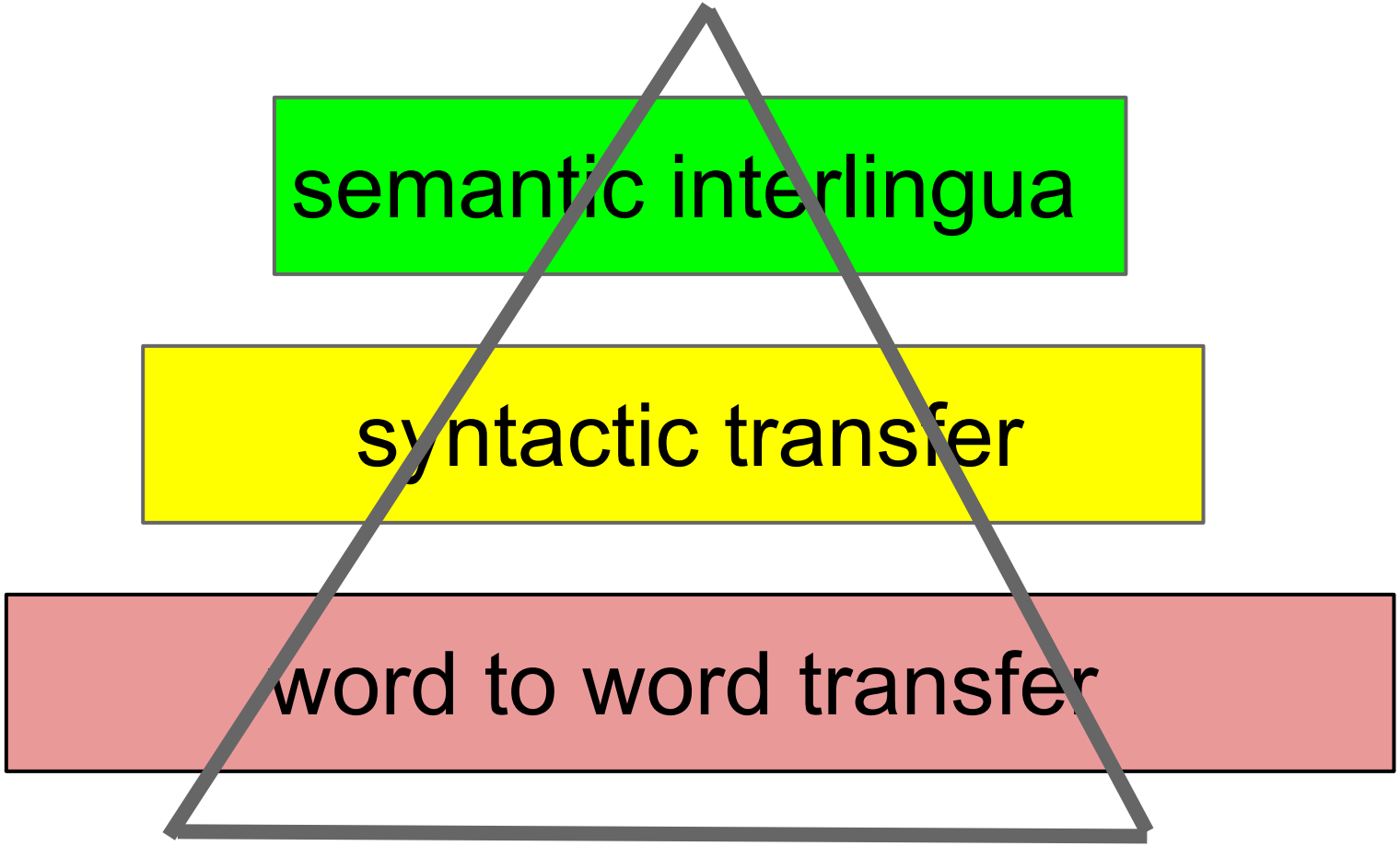
linearization

Mene metrolla 2 Pudongin lentokentältä Hongqiao-asemalle.

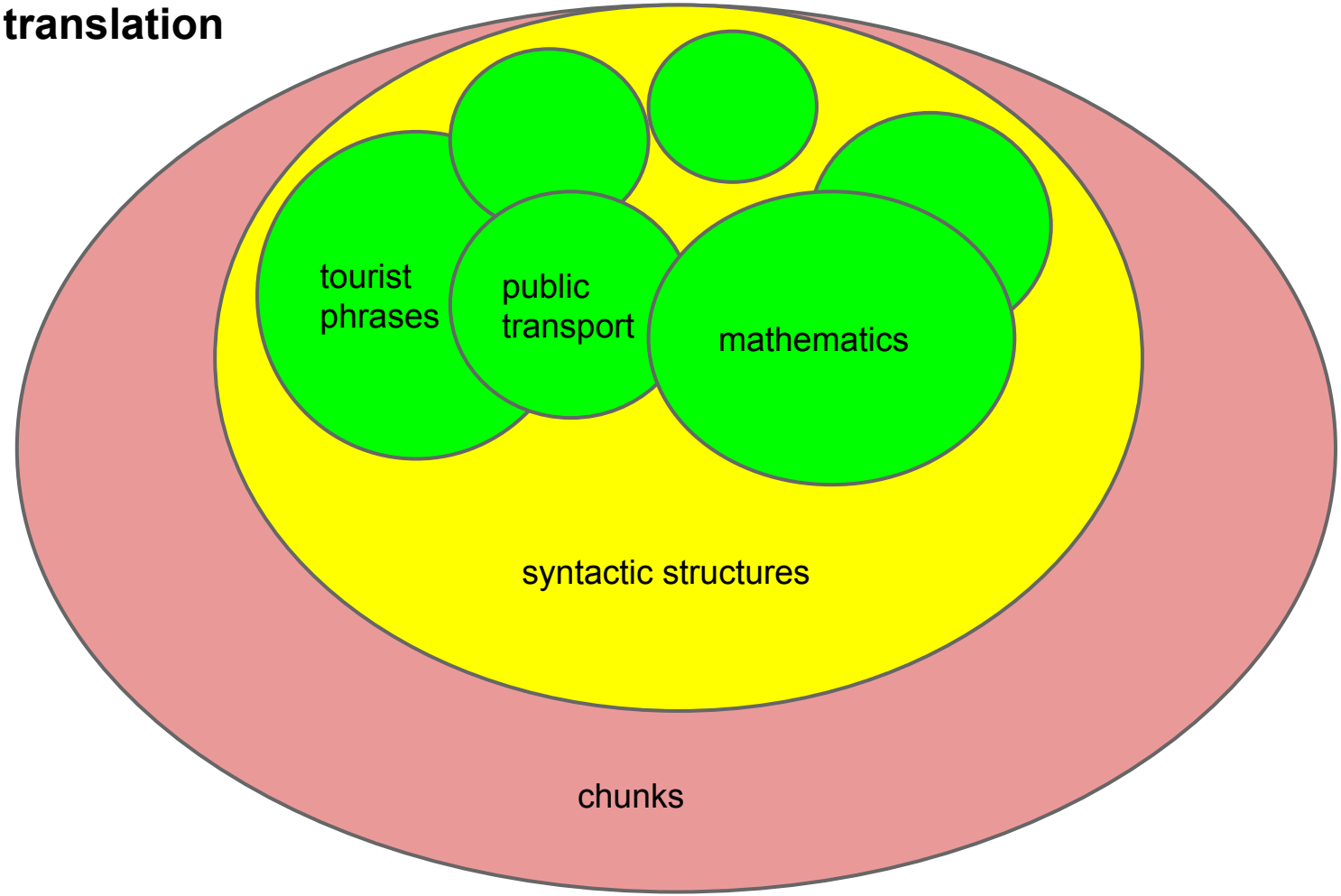
Scaling up

approaching wide-coverage translation

The Vauquois triangle



Layered translation



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.

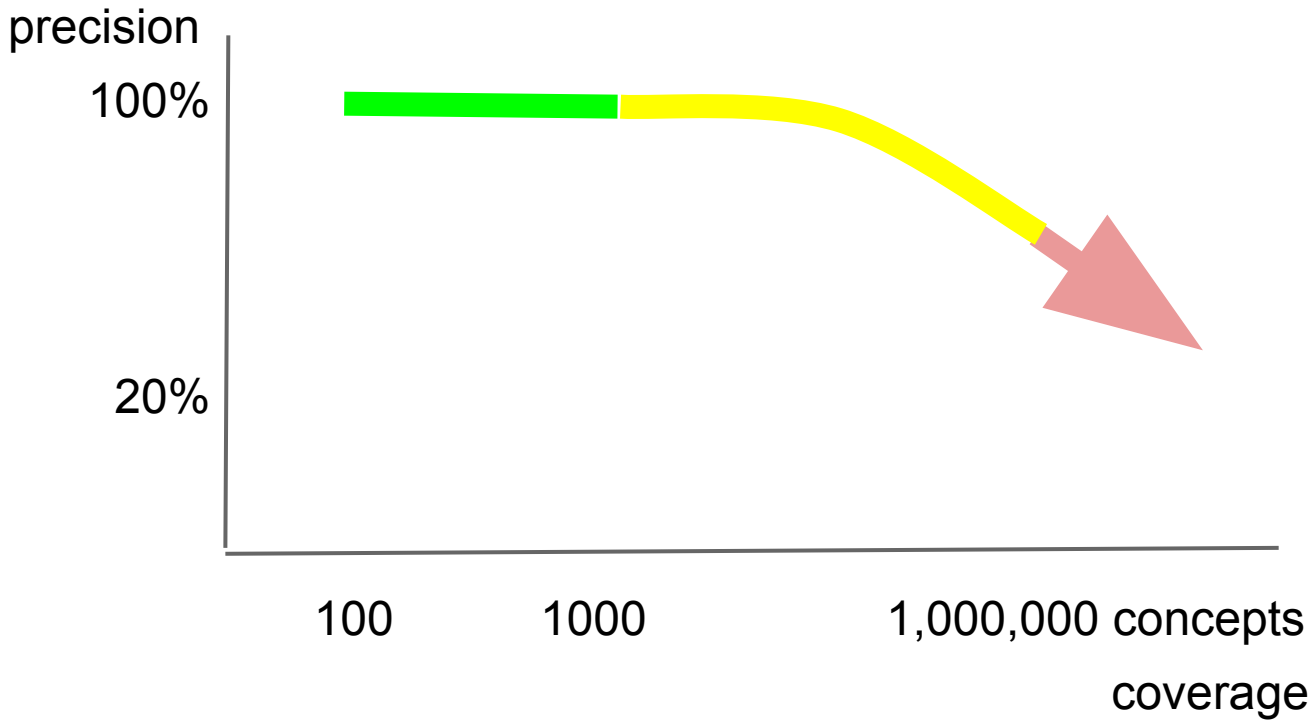
syntax

Little boy eat big snake.

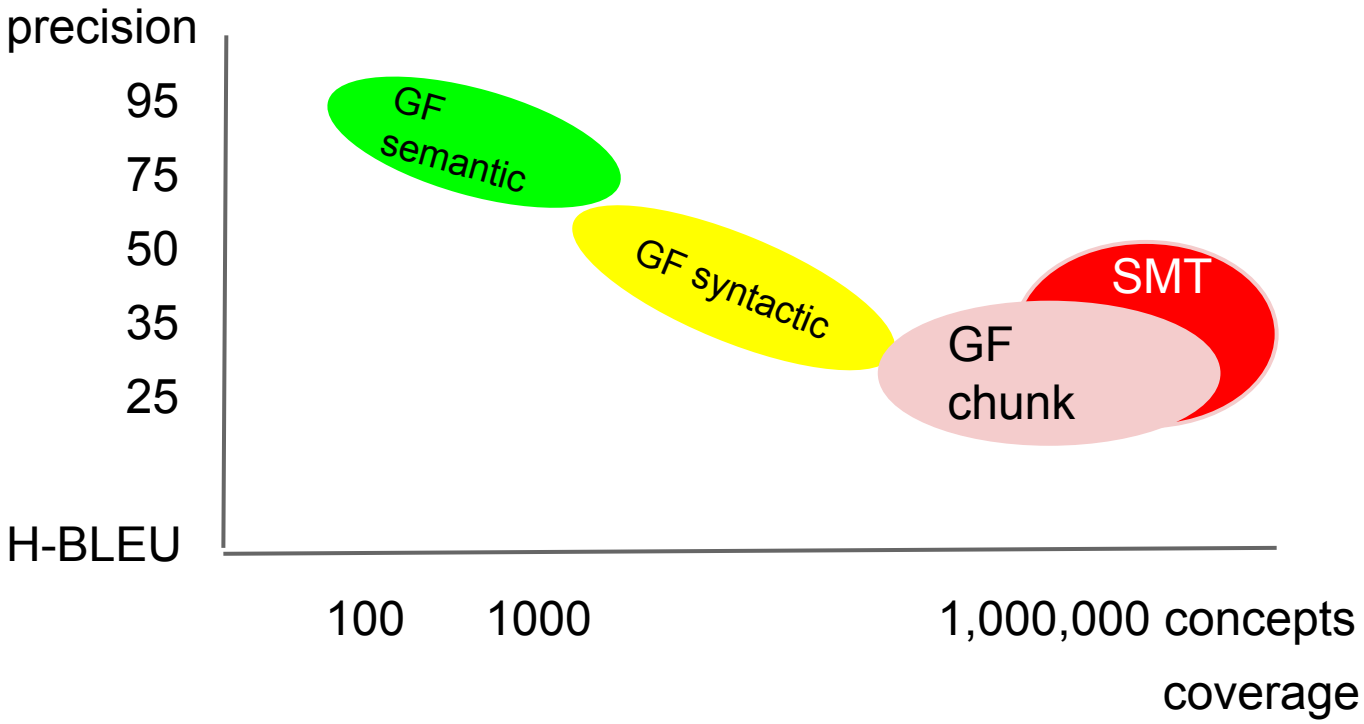
小 男 孩 吃 大 蛇.

chunks

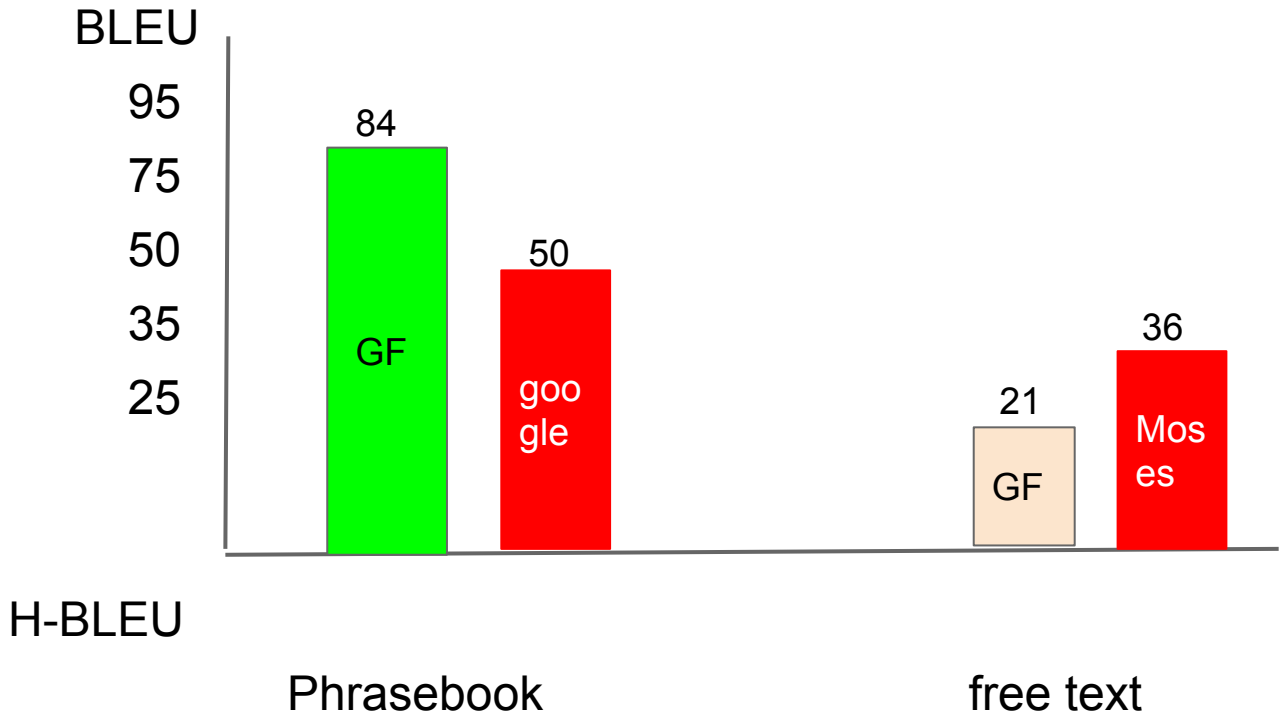
Graceful degradation



Where we are now



Some English-Chinese scores



GF vs. Statistical Translation

+ grammatical correctness
+ feedback: trees, colours
+ less dependent on language data
+ compact size of multilingual systems

GF vs. Statistical Translation

+ grammatical correctness
+ feedback: trees, colours
+ less dependent on language data
+ compact size of multilingual systems
- non-compositional idioms
- contextual disambiguation

GF vs. Statistical Translation

+ grammatical correctness
+ feedback: trees, colours
+ less dependent on language data
+ compact size of multilingual systems
- non-compositional idioms
- contextual disambiguation

- cases for hybrid methods

Size of mobile app

For 15 languages, 210 language pairs

- 16 modules, 40 MB in total
- Google translate offline: 210 modules, 150 MB each
- Baidu translate offline: 30 MB each

Resources

www.grammaticalframework.org

Grammatical Framework

A programming language for multilingual grammar applications

Use GF

- GF Cloud
- Android app
- Other Demos

- Download GF
- GF Eclipse Plugin
- GF Editor Modes

- User Group
- Bug Reports
- Blog

Learn GF

- QuickStart
- QuickRefCard
- GF Shell Reference
- **GF Summer School**.

- The GF Book
- GF Tutorial
- Reference Manual
- Best Practices [PDF]

- Library Synopsis
- Library Tutorial [PDF]
- Coverage Map

Develop GF

- build passing
- GF Developers Guide
- GitHub mirror
- Wiki
- Browse Source Code
- Authors

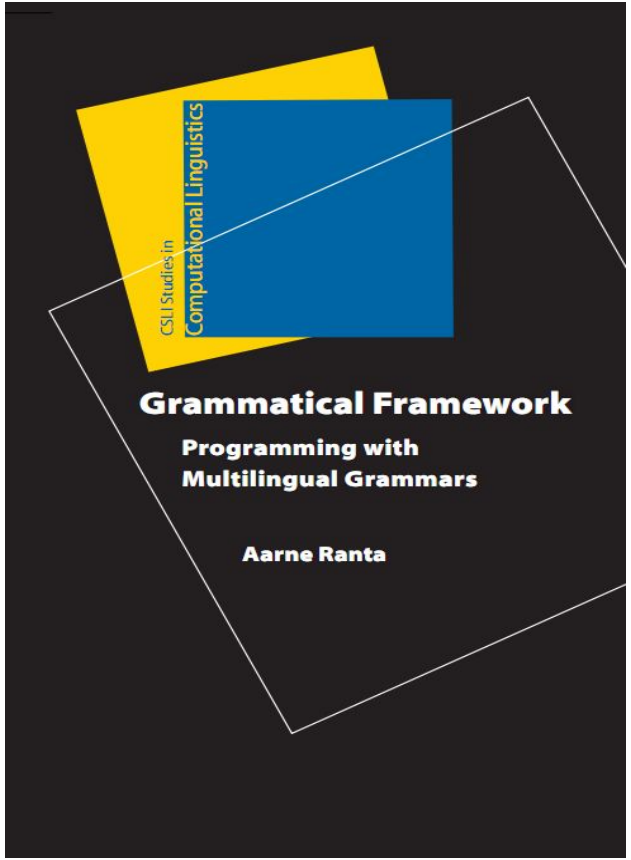
Develop Applications

- PGF library API (Haskell)
- PGF library API (Python)
- GF on Android (new)
- GF on Android (old)

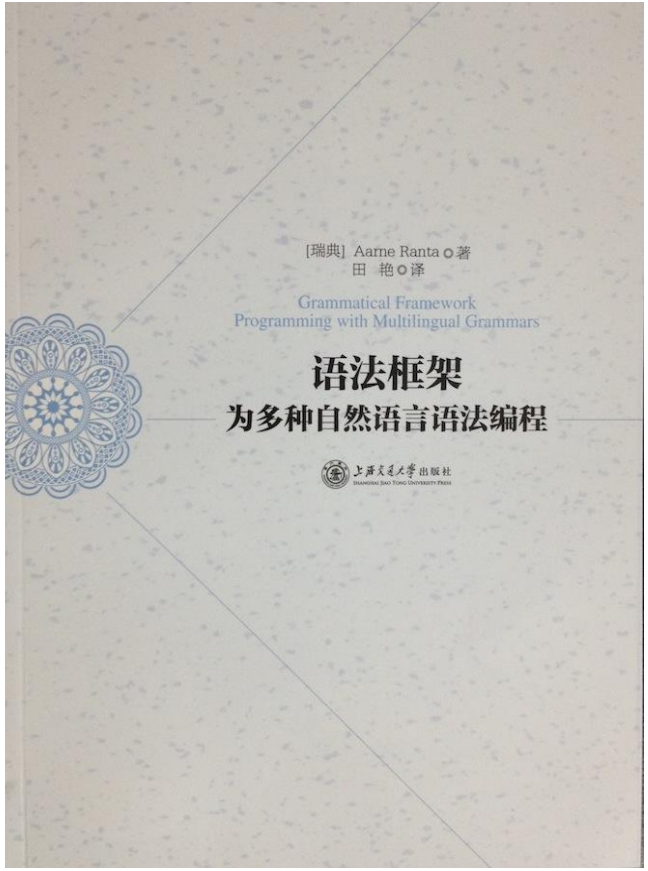
Related to GF

- Publications
- GF Summer Schools
- The REMU Project
- The MOLTO Project
- GF on Wikipedia

- Digital Grammars AB



CSLI, Stanford, 2011



Shanghai Jiao Tong University press, 2014

GF languages and contributors

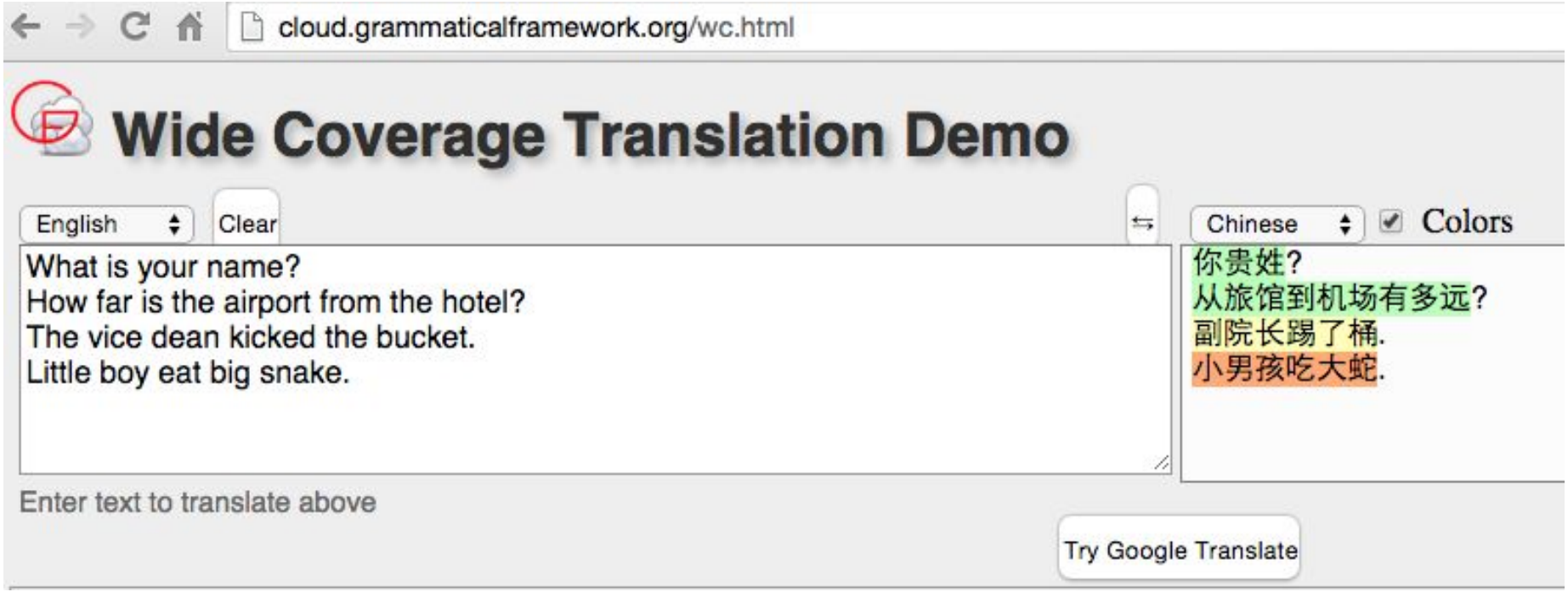
GF Offline Translator

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

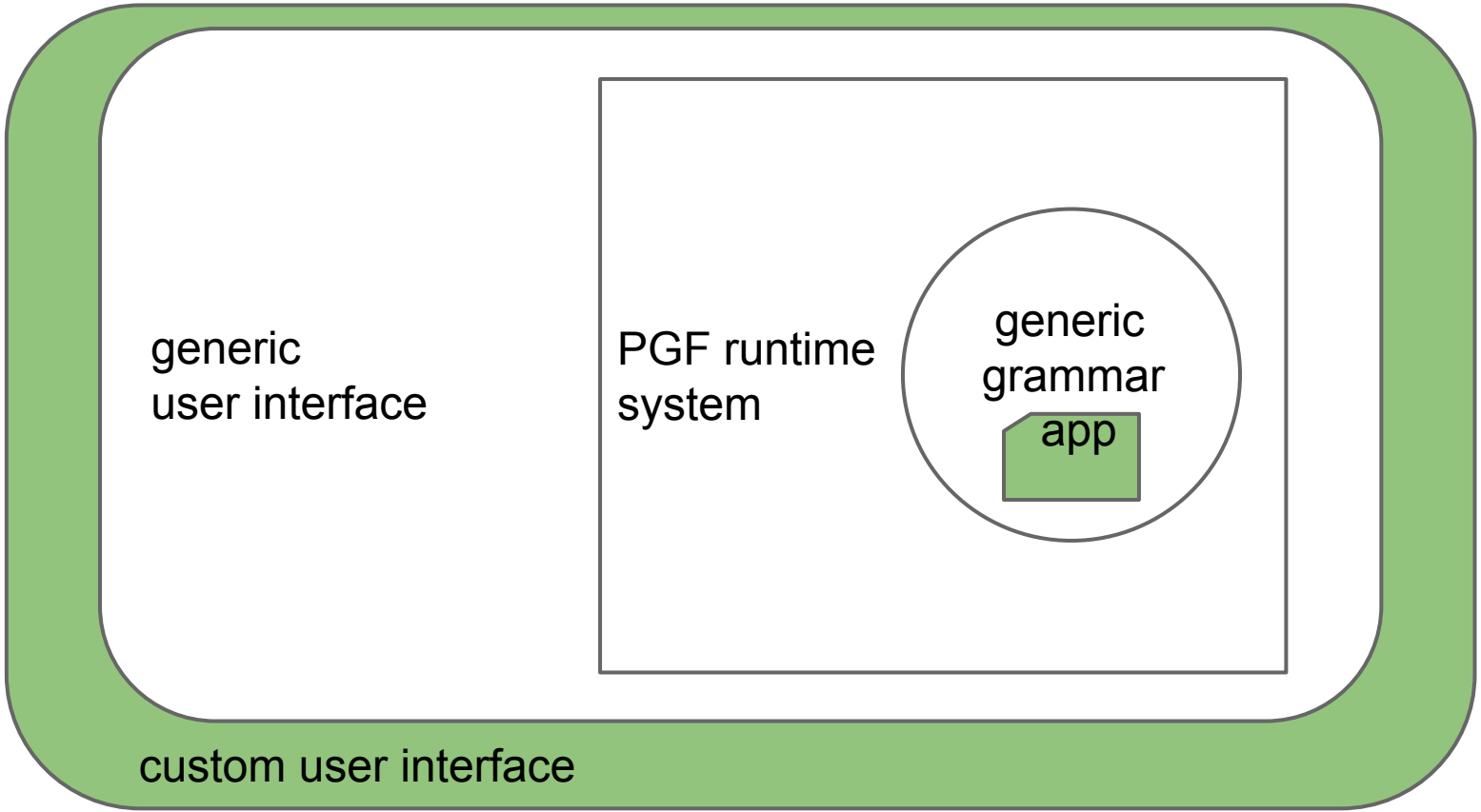
https://itunes.apple.com/us/app/gf-offline-translator/id1023328422?mt=8

K. Angelov, B. Bringert & A. Ranta, Speech-enabled hybrid multilingual translation for mobile devices, EACL 2014.

GF Cloud Translator



White: free, open-source (BSD) **Green**: can be sold



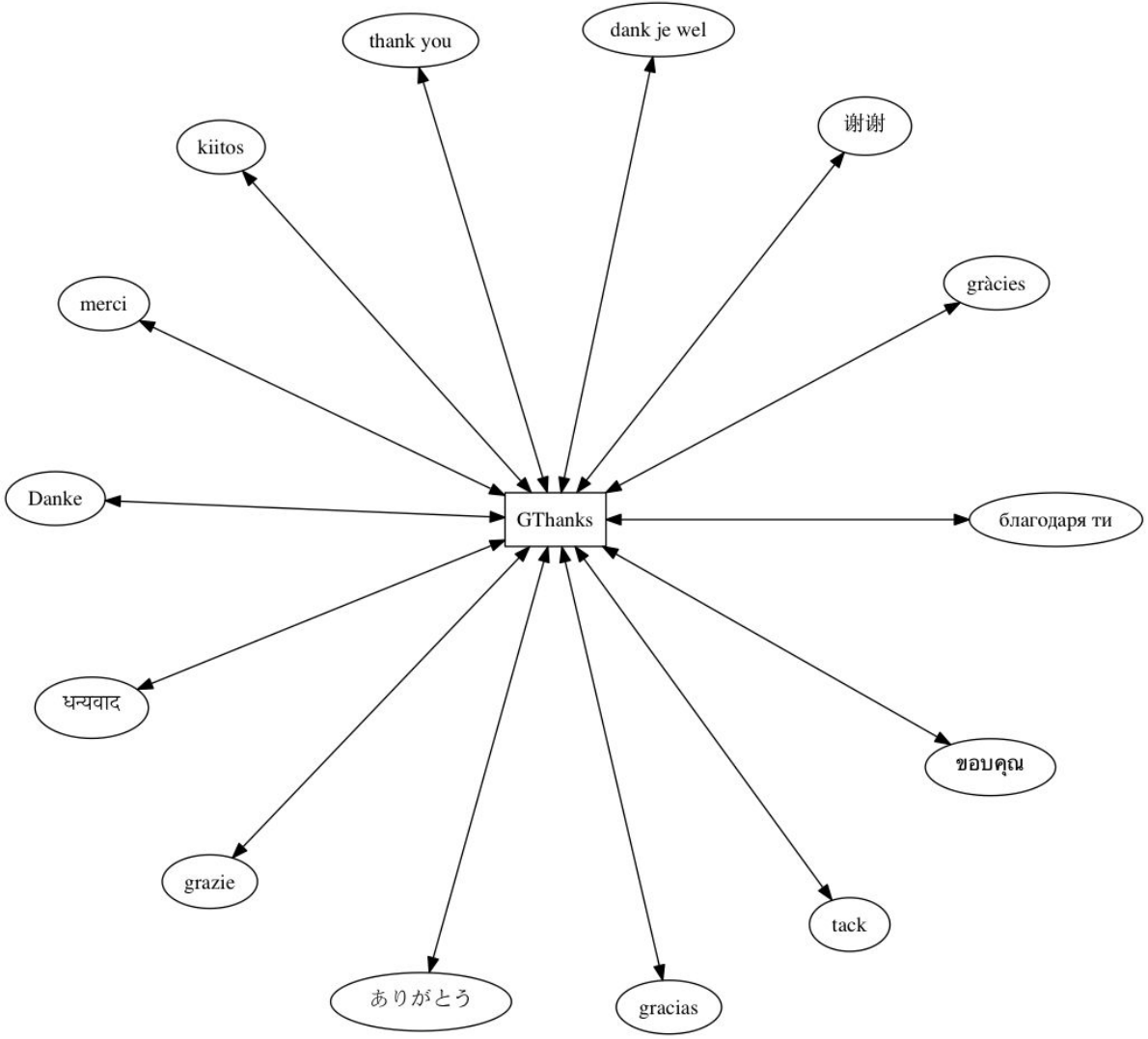
Take home points

Data-Driven Documentation

- abstract syntax as data representation
- translation interlingua
- multilingual question answering

GF = Grammatical Framework

- 30 languages
- scalable with confidence levels
- open source + commercial applications


thank you
dank je wel
kiitos
谢谢
merci
gràcies
Danke
GThanks
благодаря ти
धन्यवाद
ขอบคุณ
grazie
tack
ありがとう
gracias