Automatic Translation

for Consumers and Producers

Aarne Ranta
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
and
Digital Grammars AB

Chalmers Initiative Seminar
Digitalization - Opportunities and Challenges, 15-16 March 2017





The New York Times Magazine

The Great A.I. Awakening

How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

BY GIDEON LEWIS-KRAUS DEC. 14, 2016

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey,
Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser,
Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens,
George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa,
Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

8 October 2016

Five, perhaps three years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact.

Five, perhaps three years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact.

IBM press release 1954

http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html

MT Milestones

1940’s: cryptanalysis

1960’s: dictionaries, grammar rules

1990’s: SMT (Statistical Machine Translation)

- IBM 1989
- Google translate 2006

2010’s: NMT (Neural Machine Translation)

- Baidu 2015
- Google translate 2016

Consumers and producers

Consumers

- you and me, most of the time

Producers

- authorities
- companies
- love letter writers
- scammers

SMT

Collect statistics from parallel texts

Translate new texts by putting together pieces

This big car is yellow. *Denna stora bil är gul.*

This house is clean. *Detta hus är rent.*

This big house is yellow.

This big car is yellow.

Denna stora bil är gul.

This house is clean.

Detta hus är rent.

This big house is yellow.

Denna stora

This big car is yellow.

Denna stora bil är gul.

This house is clean.

Detta hus är rent.

This big house is yellow.

Denna stora hus är

This big car is yellow.

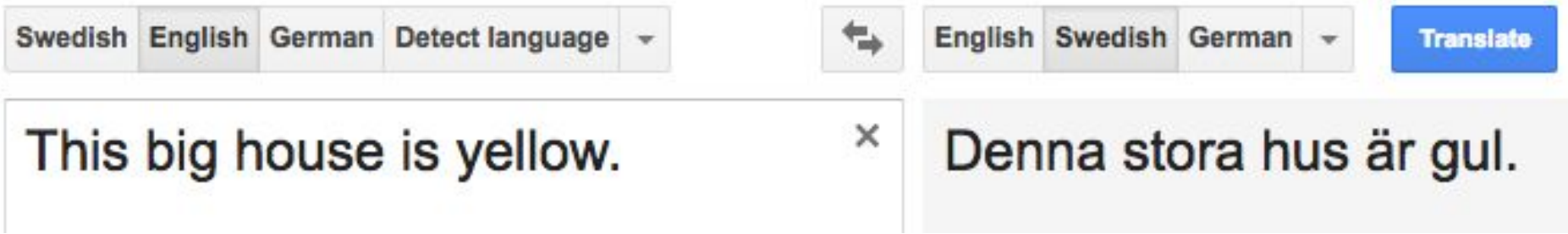
Denna stora bil är gul.

This house is clean.

Detta hus är rent.

This big house is yellow.

Denna stora hus är gul.



https://translate.google.com/ 13 March 2017

Data sparsity problem

We need hundreds of millions of words of text to cover

- word inflections
- word combinations

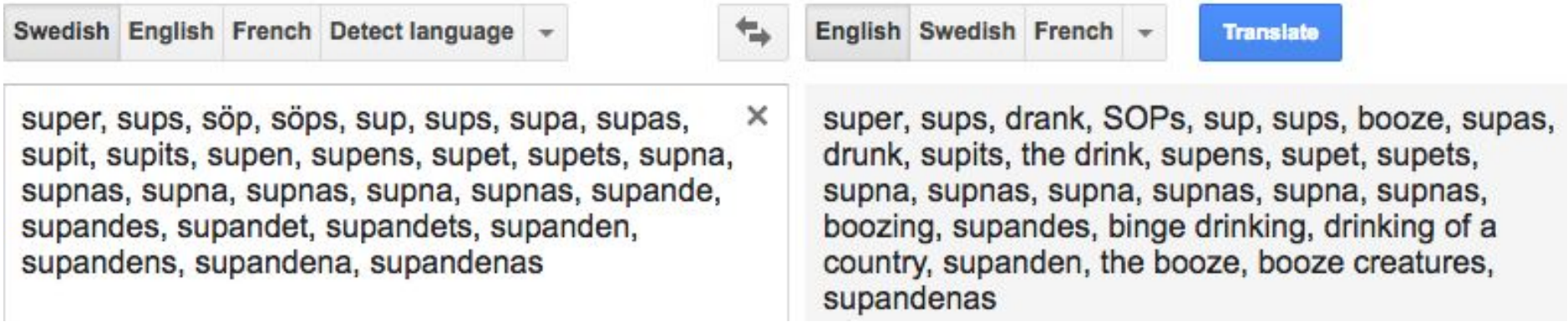
Swedish has

- ~ 10^6 word forms (+ unlimited compounds)
- ~ 10^{12} combinations of two words

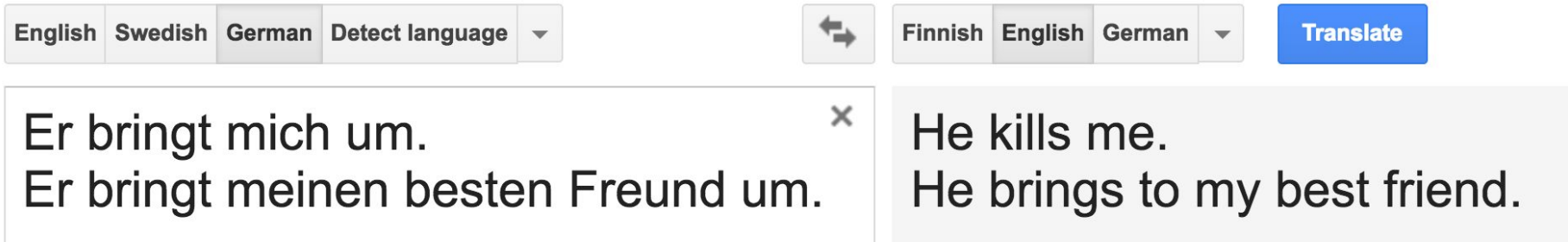
Finnish has

- ~ 10^8 word forms (+ unlimited compounds)
- ~ 10^{16} combinations of two words

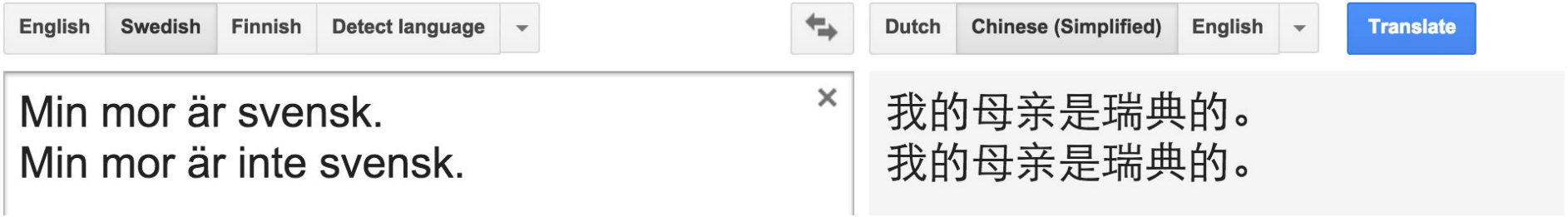
Word form coverage



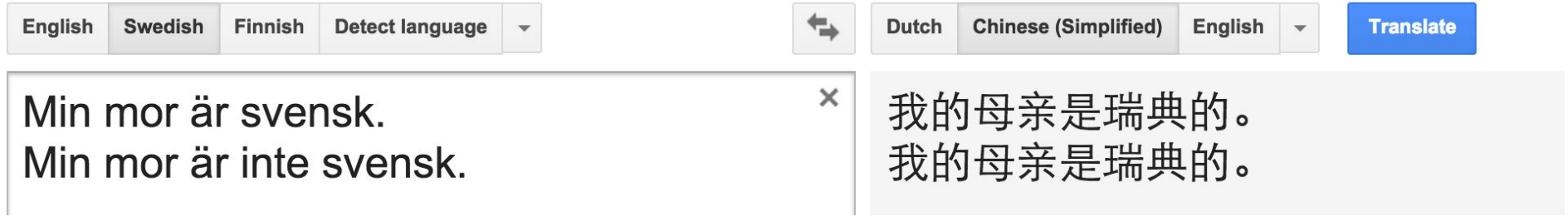
Long distance

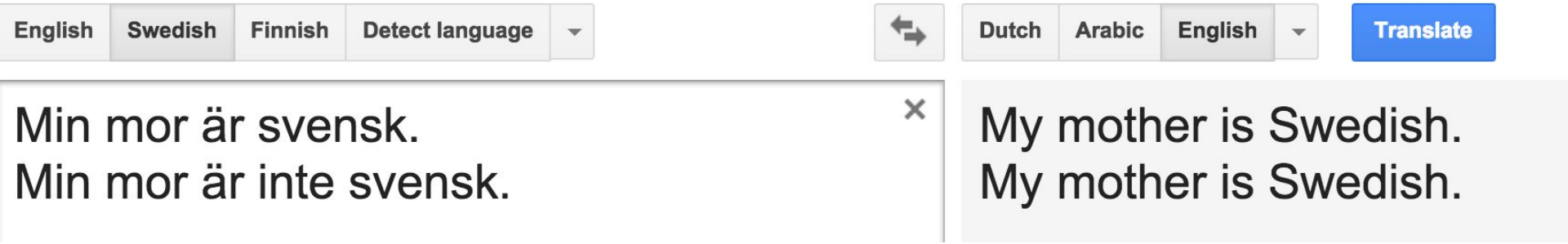


Reliability

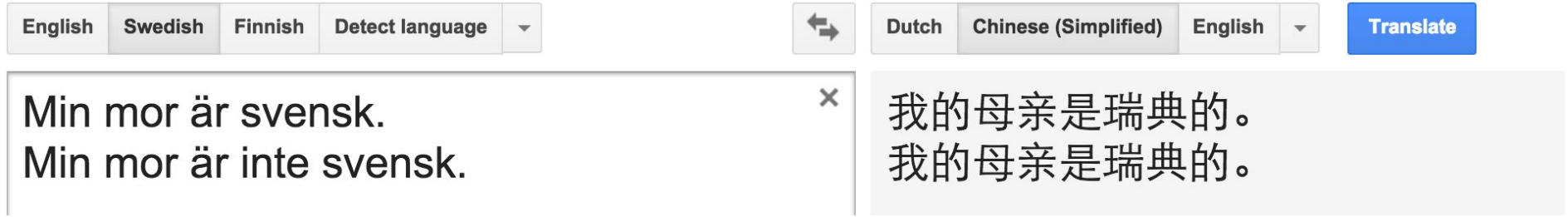


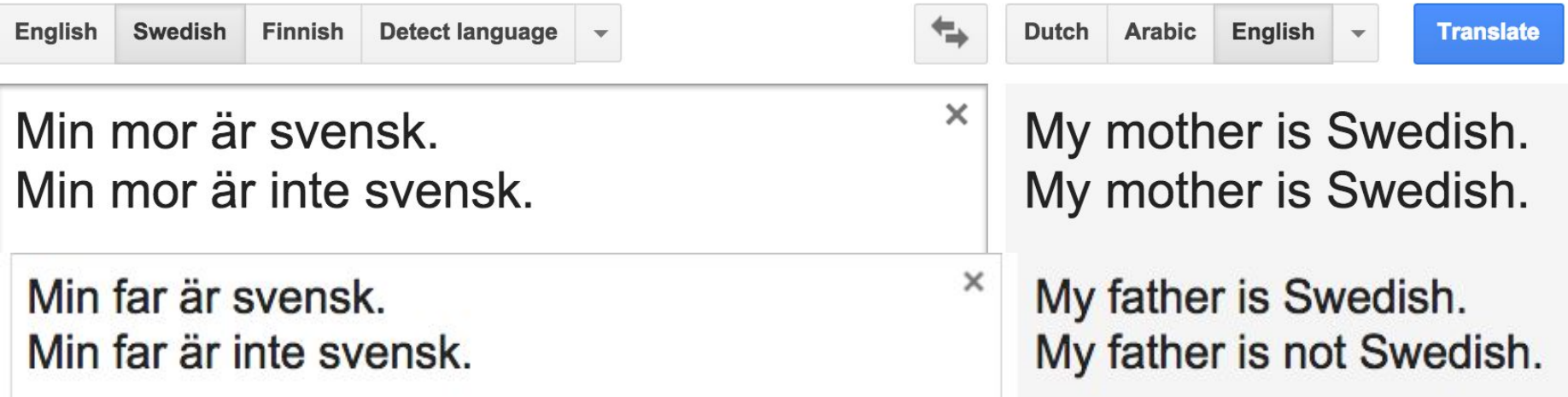
Reliability





Reliability





MT evaluation: the BLEU score

Similarity between MT result and a reference.

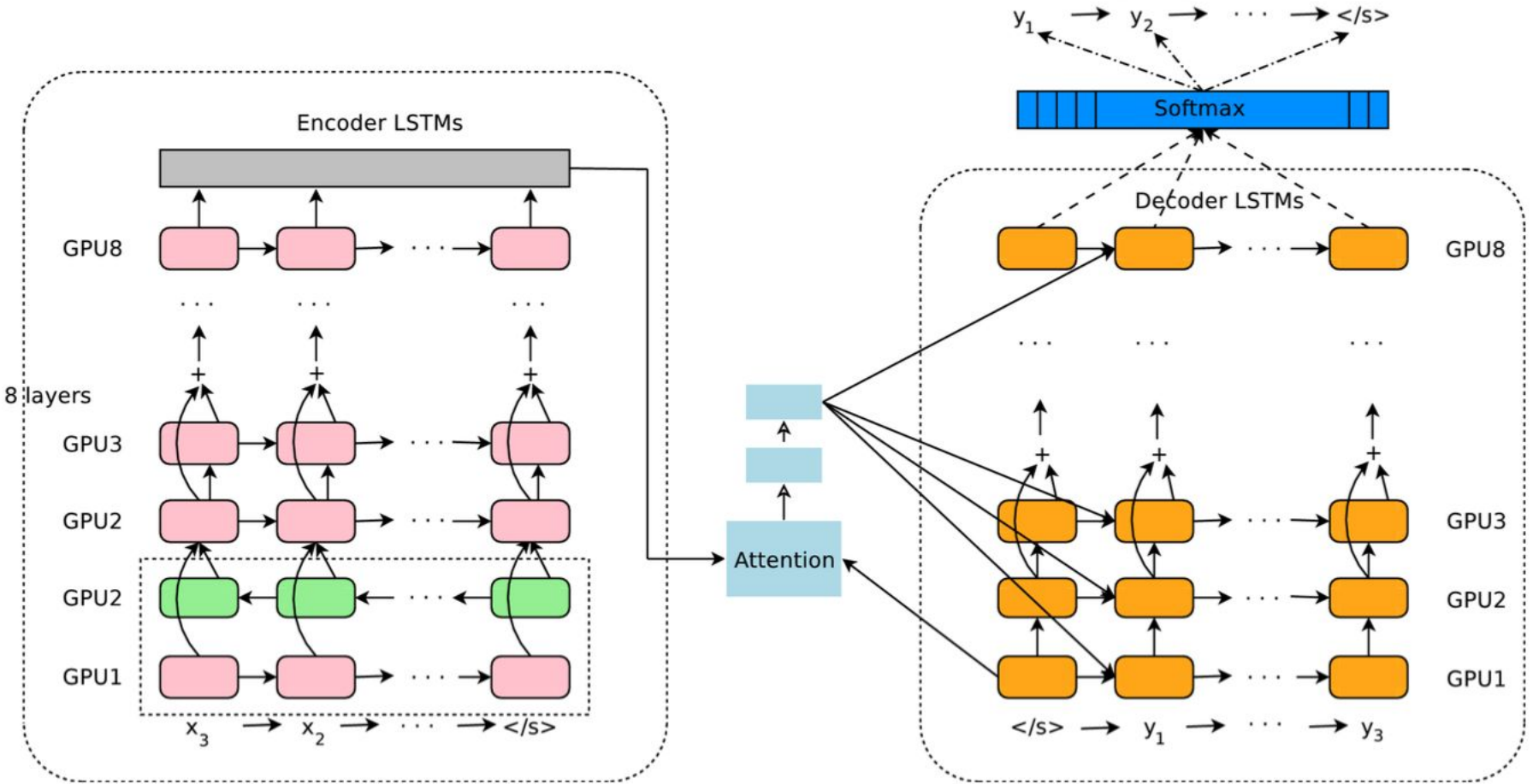
original	MT	reference	BLEU
Min mor är svensk.	*My mother is Swedish.*	*My mother is Swedish.*	1.0
Min mor är inte svensk.	*My mother is not Swedish.*	*My mother is Swedish.*	0.41

NMT

Neural network: a computational model of the brain

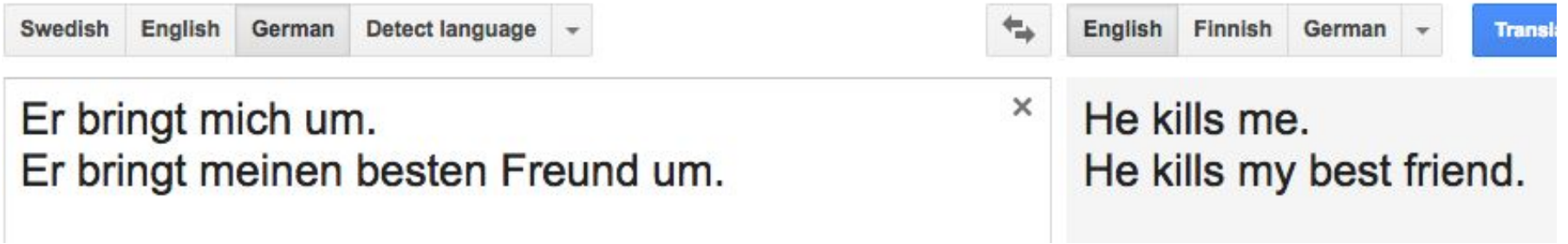
It can *learn* a translation function from parallel text

Improved performance with long distance


Encoder LSTMs
GPU8
8 layers
GPU3
GPU2
GPU2
GPU1
x3
x2
</s>
Attention
y1
y2
</s>
Softmax
Decoder LSTMs
GPU8
GPU3
GPU2
GPU1
</s>
y1
y3

Wu & al, 2016

NMT with long distance



https://translate.google.com/ 13 March 2017

Evaluation of NMT

BLEU scores (max 1.0)

SMT	NMT
0.37	0.41

Fluency (max 6.0)

SMT	NMT	human
3.87	4.44	4.82

English-French, from Wu & al, 2016

Consumers vs. producers

Consumer:

- foreign languages → my language
- get an idea
- no-one is responsible for the translation

Producer

- my language → foreign languages
- get the message through
- responsible for content and its translations

Can we ever reach producer quality

Bar-Hillel 1962: translation is **AI complete**

pen -> penna / lekhage

The pen is in the box. *Pennan är i lådan.*

The box is in the pen. *Lådan är i lekhagen.*

A recent real-life example

bestemor -> mormor

Should have been *farmor* !

The true method of translation?

SMT

- can’t be, and we know it

NMT

- can be, but we don’t know

The true method of translation?

SMT

- can’t be, and we know it

NMT

- can be, but we don’t know

GSM (Grammatical and Semantic Methods)

- can be, and we know what is happening

Grammars and semantics

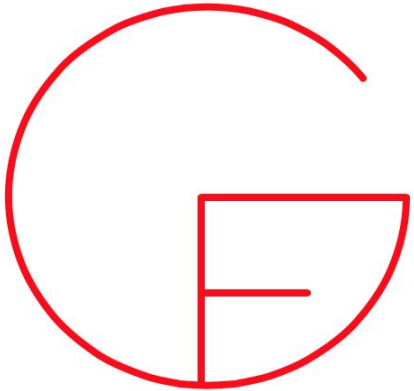
WordNet (Princeton)

FrameNet (Berkeley)

Universal Dependencies (Stanford, Google, Uppsala...)

Apertium (Alicante, Tromsø)

Grammatical Framework = GF (Xerox, Gothenburg)



The mission of GF is to formalize the grammars of the world and make them available for computer applications.

http://www.grammaticalframework.org

Since 1998

Global community of 200+ members

Nieminen, Camilleri 2013-

32 languages:

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

More in progress:

Latin Slovene Turkish Basque Hebrew Arabic Amharic Swahili ...

Abstract	PredVP (DetCN this_Det (ModCN (PositA big_A) (UseN house_N))) (CompAP (PositA yellow_A))
Afr	hierdie groot huis is geel
Bul	тази голяма къща е жълта
Cat	aquesta casa grosa és groga
Chi	这 间 大 房 子 是 黄 的 zh-TW zh-HK zh-CN zh-CN zh-HK zh-TW
Dan	dette store hus er gult da-DK
Dut	dit grote huis is geel nl-BE nl-NL nl-NL
Eng	this big house is yellow en-US
Est	see suur maja on kollane
Fin	tämä suuri talo on keltainen fi-FI
Fre	cette grande maison est jaune fr-CA fr-FR fr-FR
Ger	dieses große Haus ist gelb de-DE de-DE
Gre	αυτό το μεγάλο σπίτι είναι κίτρινο el-GR
Hin	यह बड़ा घर पीला है hi-IN hi-IN
Ice	þetta stóra hús er gult
Ita	questa grande casa è gialla it-IT it-IT it-IT
Jpn	この 大きな 家 は 黄色 だ ja-JP ja-JP
Lav	šī lielā māja ir dzeltena
Mlt	din id-dar kbira hija safra
Mon	энэ том байшин шар байна
Nep	यो ठुलो घर पहेलो छ
Nno	dette store huset er gult
Nor	dette store huset er gult nb-NO
Pes	این خانه ی بزرگ زرد است
Pnb	اے وڈا گھر کھٹا اے
Pol	ten duży dom jest żółty pl-PL pl-PL
Ron	această casă mare este galbenă ro-RO
Rus	этот большой дом жёлт ru-RU ru-RU ru-RU
Snd	هي وڏو گهر پيلو آهي
Spa	esta grande casa es amarilla es-AR es-ES es-MX es-ES es-MX es-ES es-US
Swe	det här stora huset är gult sv-SE
Tha	บ้าน ใหญ่ หลัง นี้ เหลือง th-TH
Urd	یہ بڑا گھر پیلا ہے

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

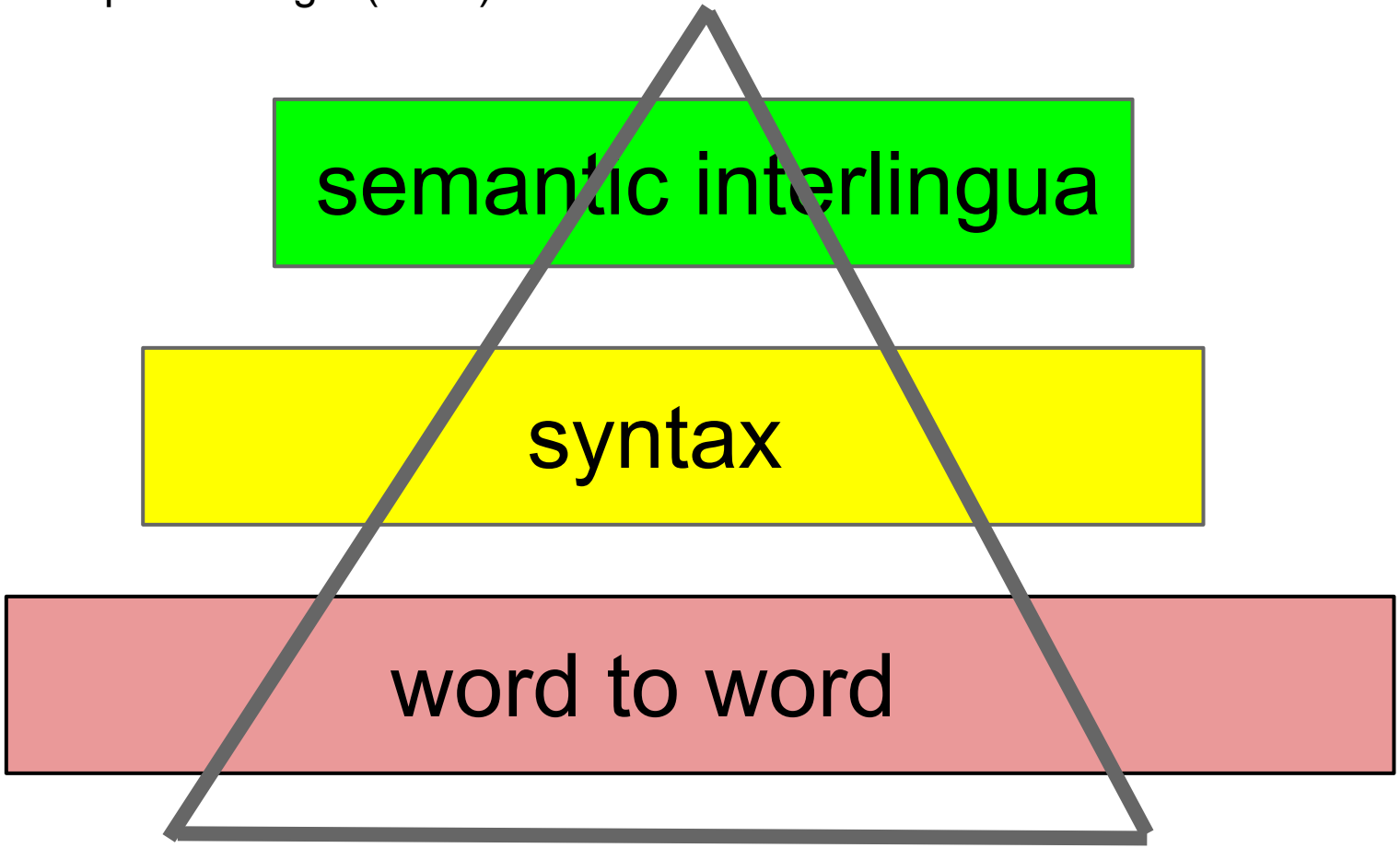
GF Offline Translator

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

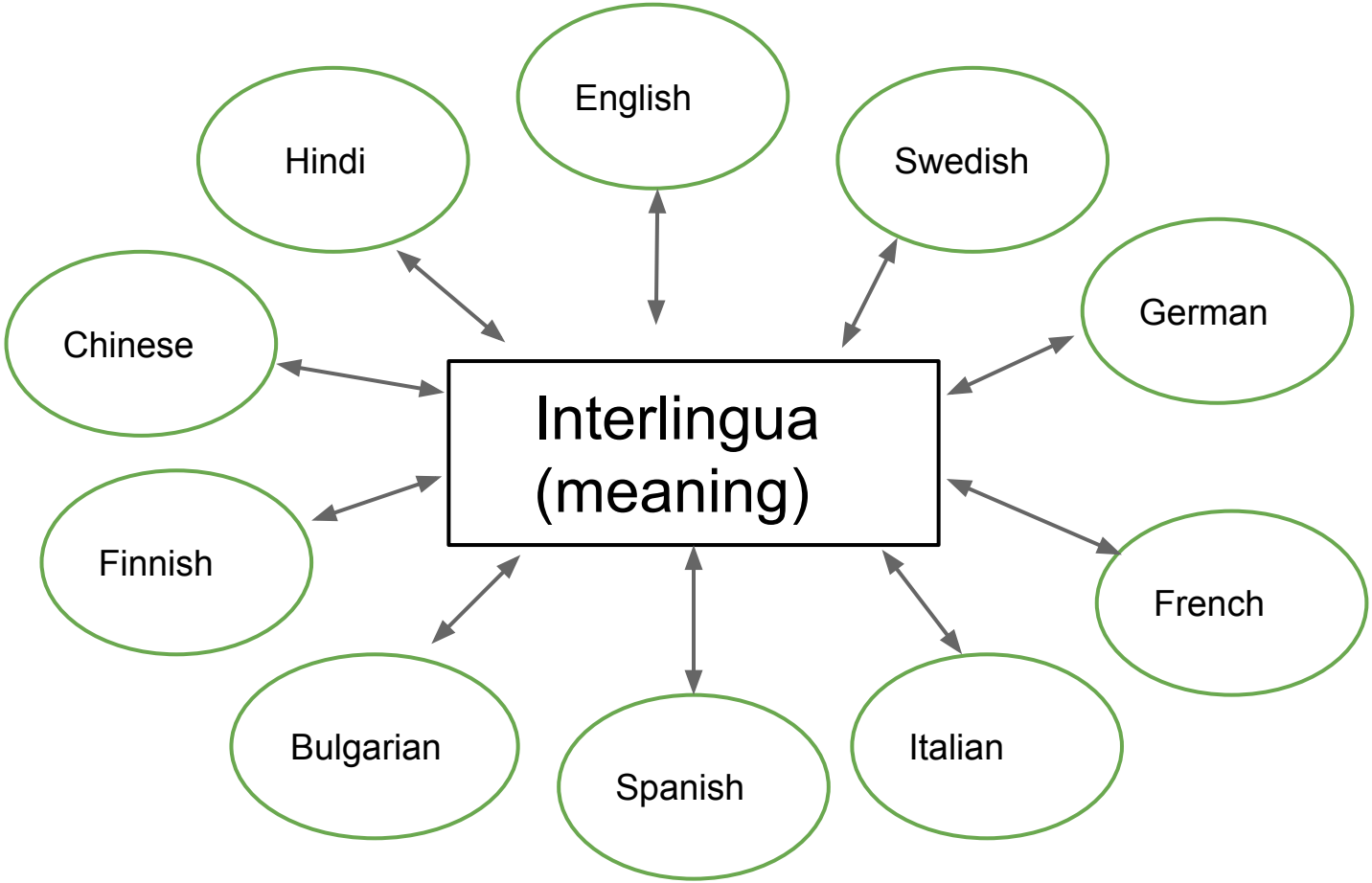
https://itunes.apple.com/us/app/gf-offline-translator/id1023328422?mt=8

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

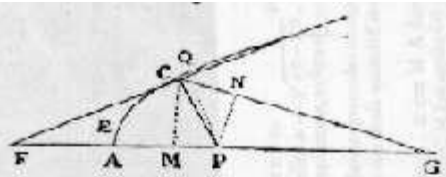
The Vauquois triangle (1968)



Interlingual translation



Facon generale pour trouuer des lignes droites, qui coupent les courbes données, ou leurs contingentes, à angles droits.



Soit C E la ligne courbe, & qu'il faille tirer vne ligne droite par le point C, qui face auec elle des angles droits. Ie suppose la chose desia faite, & que la ligne cherchée est C P, laquelle ie prolonge iusques au point P, ou elle rencontre la ligne droite G A, que ie suppose estre celle aux poins de laquelle on rapporte tous ceux de la ligne C E: en sorte que faisant M A ou C B $\propto y$, & C M, ou B A $\propto x$, iay quelque equation, qui explique le rapport, qui est entre x & y. Puis ie fais P C $\propto s$, & P A $\propto v$, ou P M $\propto v - y$, & a cause du triangle rectangle P M C iay ss, qui est le quarré de la baze esgal à $xx + vv - 2vy + yy$, qui sont les quarrés des deux costés. c'est a dire iay $x \propto \sqrt{ss - vv + 2vy - yy}$, oubien $y \propto v + \sqrt{ss - xx}$, & par le moyen de cete equation, i'oste de l'autre equation qui m'explique le rapport qu'ont tous les poins de la courbe C E a ceux de la droite G A, l'vne des deux quantités indeterminées x ou y. ce qui est aysé a faire en mettant partout $\sqrt{ss - vv + 2vy - yy}$ au lieu d'x, & le quarré de cete somme au lieu d'xx, & son cube au lieu d'x^3, & ainsi des autres, si c'est x que ie veuille oster; ou-

bien

if he put into his dictionary a single symbol corresponding to *aimer*, *amare*, *philein* and each of the synonyms, **a book written in such symbols could be translated by all who possessed the dictionary**.

In a single day one can learn to name every one of the infinite series of numbers, and thus to **write infinitely many different words in an unknown language**. The same could be done for all the other words necessary to express all the other things which fall within the purview of the human mind.

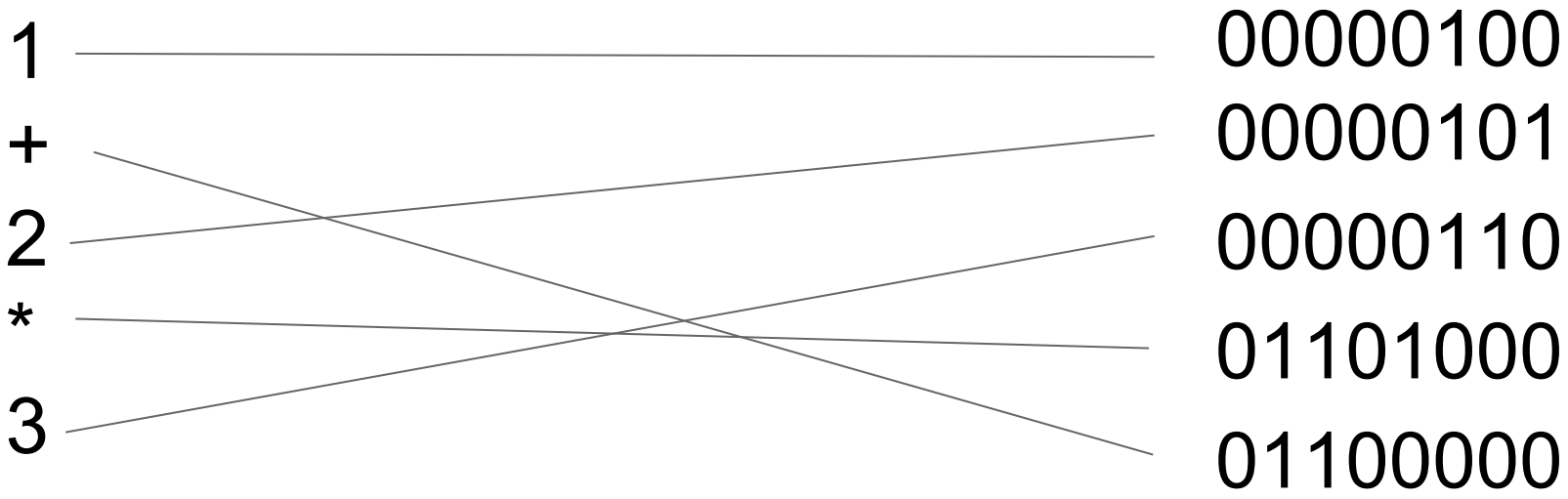
Descartes, letter to Mersenne, 1629

Compilers: Java to JVM

1 + 2 * 3

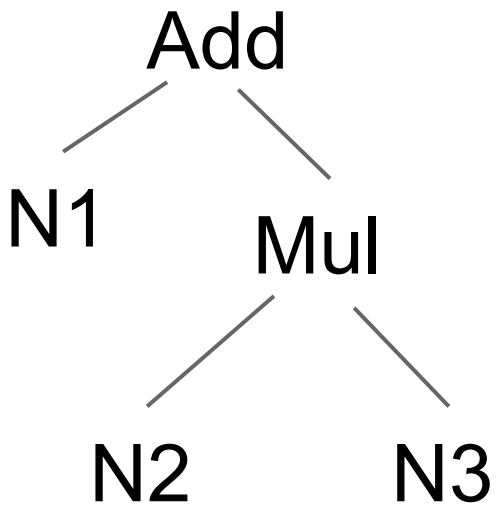
00000100
00000101
00000110
01101000
01100000

Task: from infix to postfix



Interlingua: abstract syntax

1 + 2 * 3



00000100
00000101
00000110
01101000
01100000

A compiler grammar

interlingua

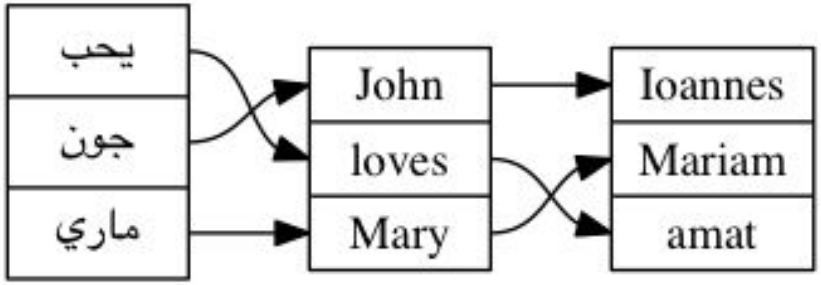
Exp Add(Exp,Exp)

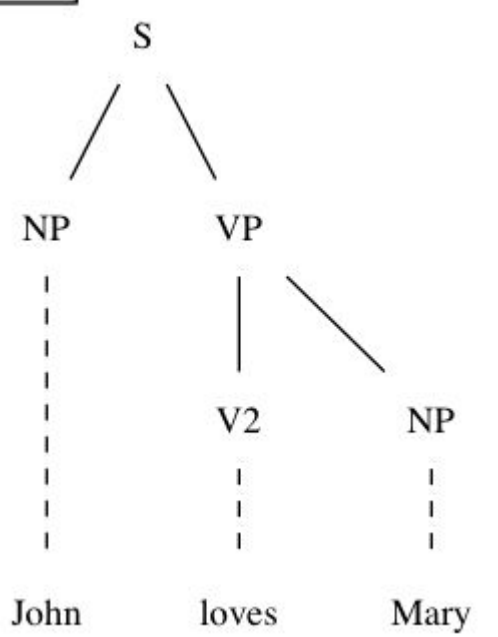
Java

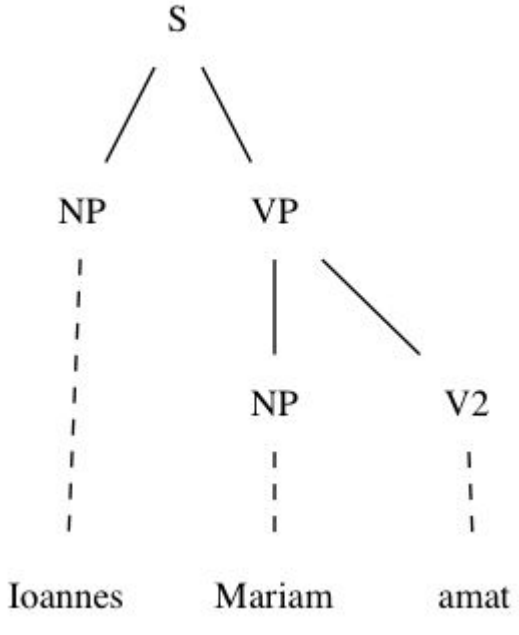
```
Add X Y    →    X “+” Y
```

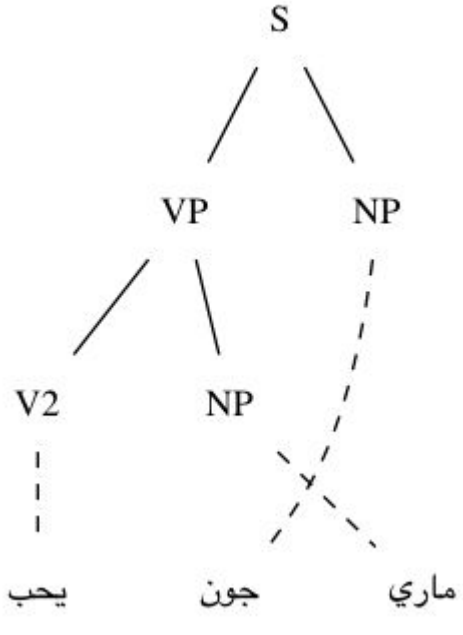
JVM

```
Add X Y    →    X Y “01100000”
```









English: SVO

Latin: SOV

Arabic: VSO

Compiling natural language

```
Interlingua:  S Pred(NP,V2,NP)

SVO: Pred S V O  →  S V O

SOV: Pred S V O  →  S O V

VSO: Pred S V O  →  V S O
```

Grammars

PhD in linguistics

5 years of work

Statistics

BSc in computer science

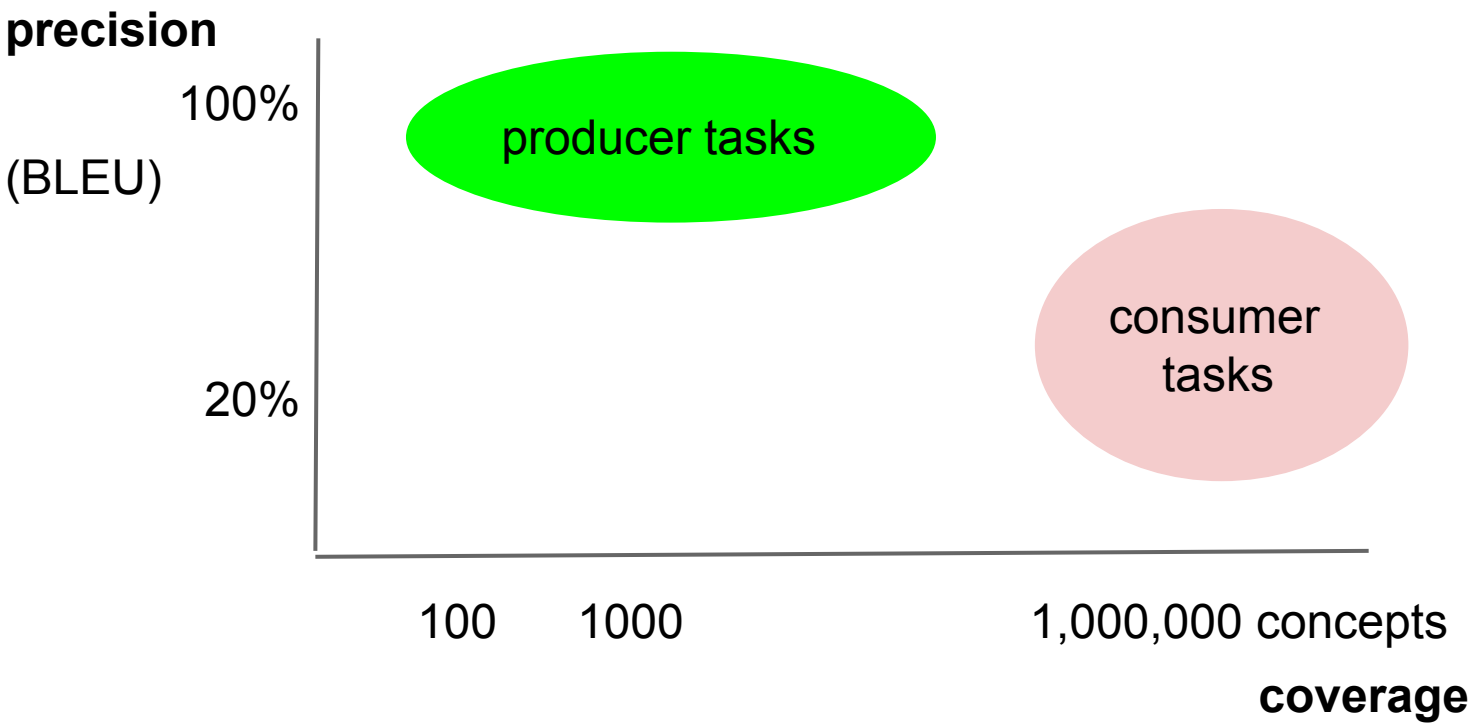
5 weeks of waiting

Two ways out

1. Focus on producer tasks

2. Improve grammar engineering

Producer vs. consumer task



TD

digital Grammars
Language technology to rely on.

- Svängrumsytan utanför dörren lutar 1% i sidled.
- The turning space outside the gate tilts 1% sideways.
- Kääntymätila oven ulkopuolella kallistuu 1% sivusuunnassa.
- Der Schwenkbereich außerhalb der Tür neigt sich um 1% seitlich

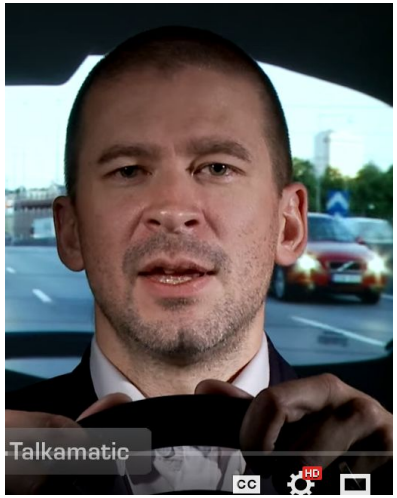
UttSTD (PredUttTD (AdvNPTD (DetCNTD (DetQuant DefArt NumSg) (UseNTD svängrumsyta_NTD)) (PrepNPTD utanför_Prep (DetCNTD (DetQuant DefArt NumSg) (UseNTD dörr_NTD)))) (AdvVPTD (luta_VPTD (procentMeasure 1)) i_sidled_AdvTD))

http://www.t-d.se/sv/TD2/

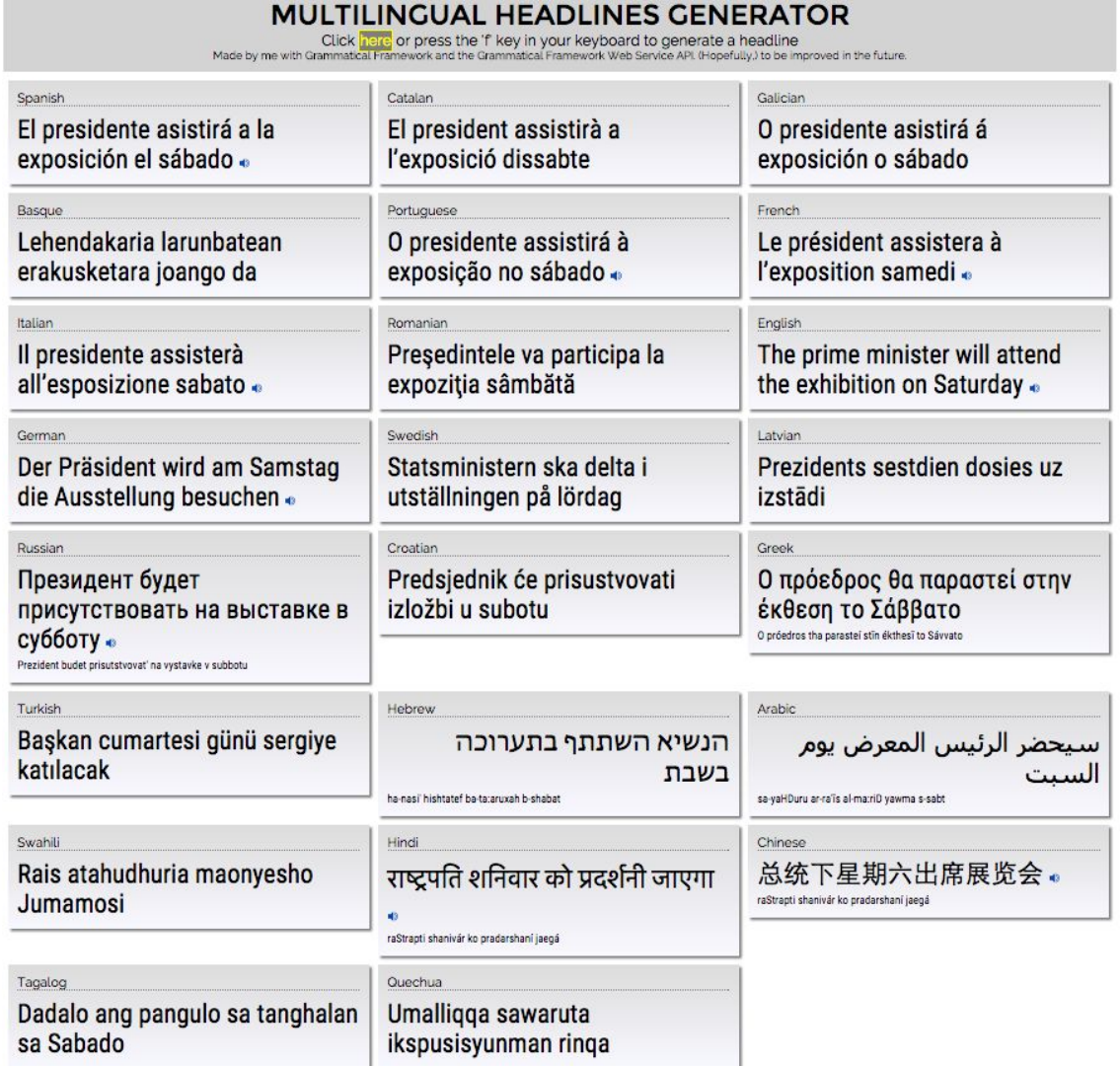
Talkamatic
FREE DIALOGUE

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


MULTILINGUAL HEADLINES GENERATOR
Click here or press the 'f' key in your keyboard to generate a headline
Made by me with Grammatical Framework and the Grammatical Framework Web Service API. (Hopefully.) to be improved in the future.
Spanish
El presidente asistirá a la exposición el sábado
Catalan
El president assistirà a l'exposició dissabte
Galician
O presidente asistirá á exposición o sábado
Basque
Lehendakaria larunbatean erakusketara joango da
Portuguese
O presidente assistirá à exposição no sábado
French
Le président assistera à l'exposition samedi
Italian
Il presidente assisterà all'esposizione sabato
Romanian
Preşedintele va participa la expoziţia sâmbătă
English
The prime minister will attend the exhibition on Saturday
German
Der Präsident wird am Samstag die Ausstellung besuchen
Swedish
Statsministern ska delta i utställningen på lördag
Latvian
Prezidents sestdien dosies uz izstādi
Russian
Президент будет присутствовать на выставке в субботу
Prezident budet prisutstvovat' na vystavke v subbotu
Croatian
Predsjednik će prisustvovati izložbi u subotu
Greek
Ο πρόεδρος θα παραστεί στην έκθεση το Σάββατο
O próedros tha parastetí stin ékthesí to Sávvato
Turkish
Başkan cumartesi günü sergiye katılacak
Hebrew
הנשיא השתתף בתערוכה בשבת
ha-nasí' hishtatef ba-ta:aruxah b-shabat
Arabic
سيحضر الرئيس المعرض يوم السبت
sa-yaHDuru ar-ra'Is al-ma:riD yawma s-sabt
Swahili
Rais atahudhuria maonyesho Jumamosi
Hindi
राष्ट्रपति शनिवार को प्रदर्शनी जाएगा
raStrapti shanivár ko pradarshaní jaegá
Chinese
总统下星期六出席展览会
raStrapti shanivár ko pradarshaní jaegá
Tagalog
Dadalo ang pangulo sa tanghalan sa Sabado
Quechua
Umalliqqa sawaruta ikspusisyunman rinqa

Moreno 2016

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)) (Ap
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 ur
de contraction, tel que pour tous les $x : A$, $a = x$.

Définition: Une application $f : A \to B$ est une éc
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 $\bar{y} := (y, r_A y) : \{y\}_A$. Comr
$(y, r_A y) = y$, 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

Grammar engineering: libraries for inflection

```
mkV “supa” “söp” “supit”
```

	active	passive
infitiv	supa	supas
presens	super	sups
preteritum	söp	söps
supinum	supit	supits
imperativ	sup	sups

Particip Presens

		obest	best
nom	**sg**	supande	supandet
	pl	supanden	supandena
gen	**sg**	supandes	supandets
	pl	supandens	supandenas

Particip Perfekt

		nom	gen
obest	**utr**	supen	supens
	neut	supet	supets
	pl	supna	supnas
best	**sg**	supna	supnas
	pl	supna	supnas

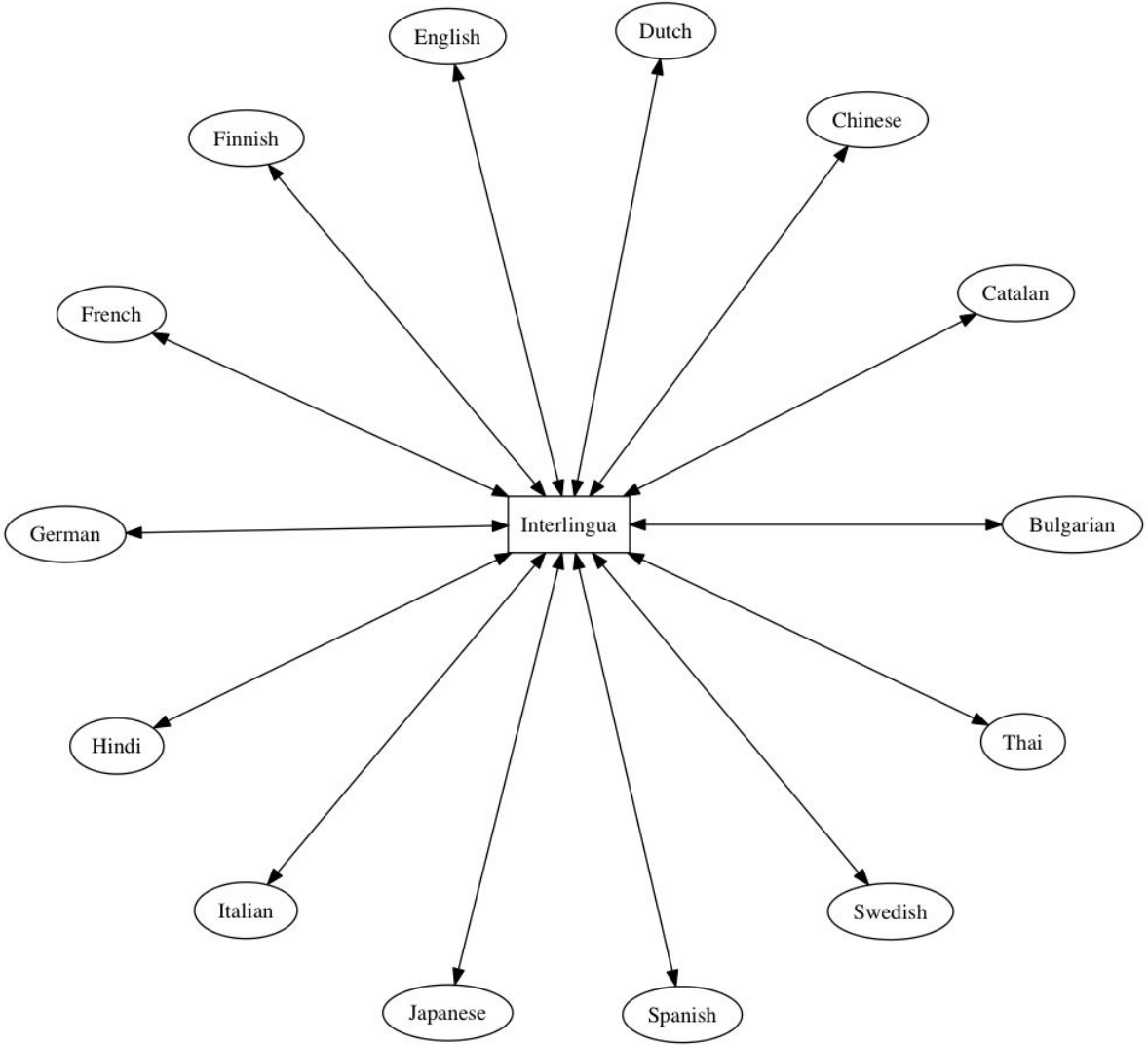
Grammar engineering: libraries for syntax

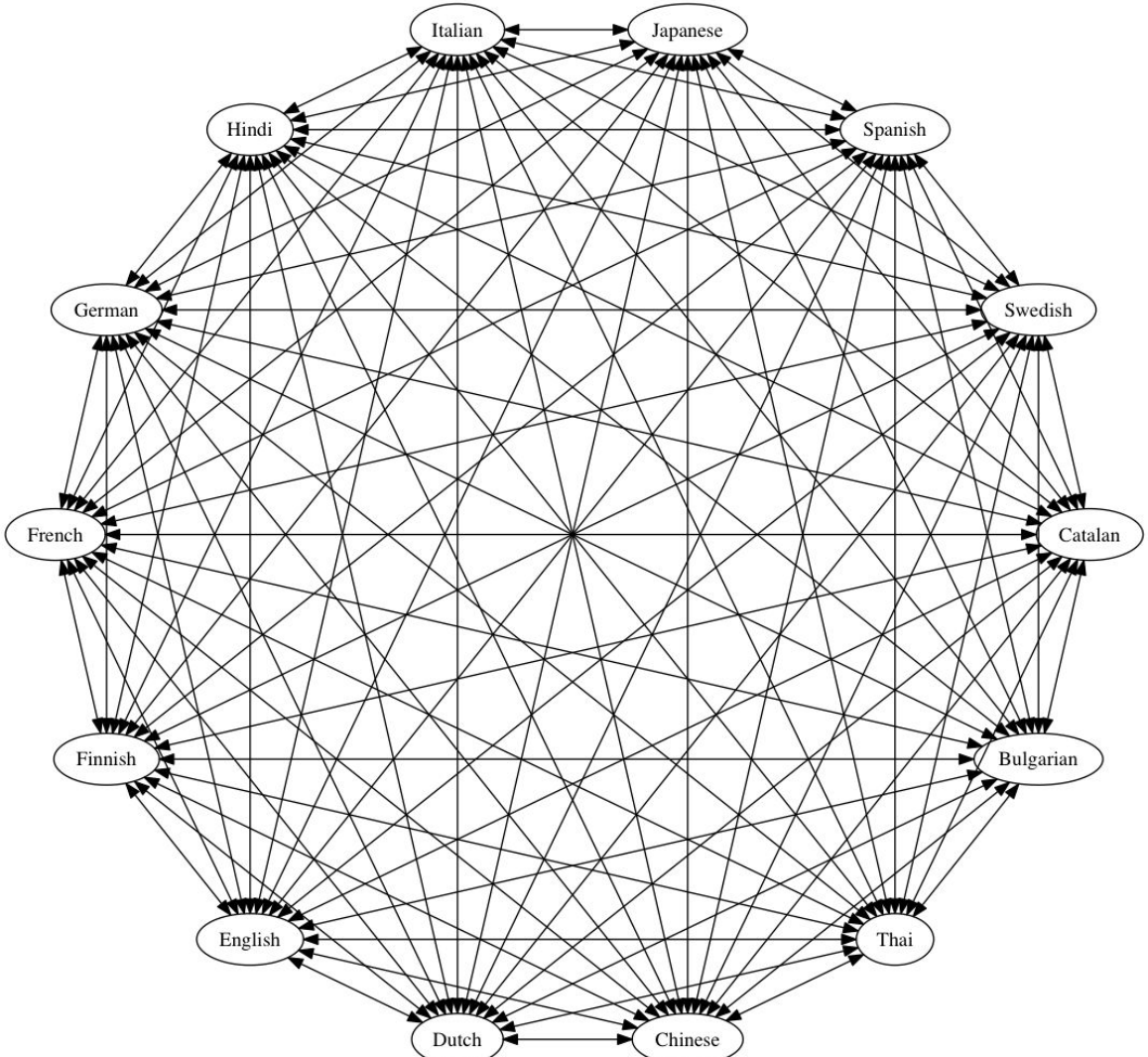
```
Pred
  i_NP
  love_V2
  you_NP
```

Pres Pos	*I love you*	*je t'aime*
Pres Neg	*I don't love you*	*je ne t'aime pas*
Pres Quest Pos	*do I love you*	*t'aime-je*
Pres Quest Neg	*don't I love you*	*ne t'aime-je pas*
Past Pos	*I loved you*	*je t'aimais*
Past Anter Pos	*I had loved you*	*je t'avais aimée*
Cond Anter Quest Neg	*wouldn't I have loved you*	*ne t'aurais-je pas aimée*
...	...	...

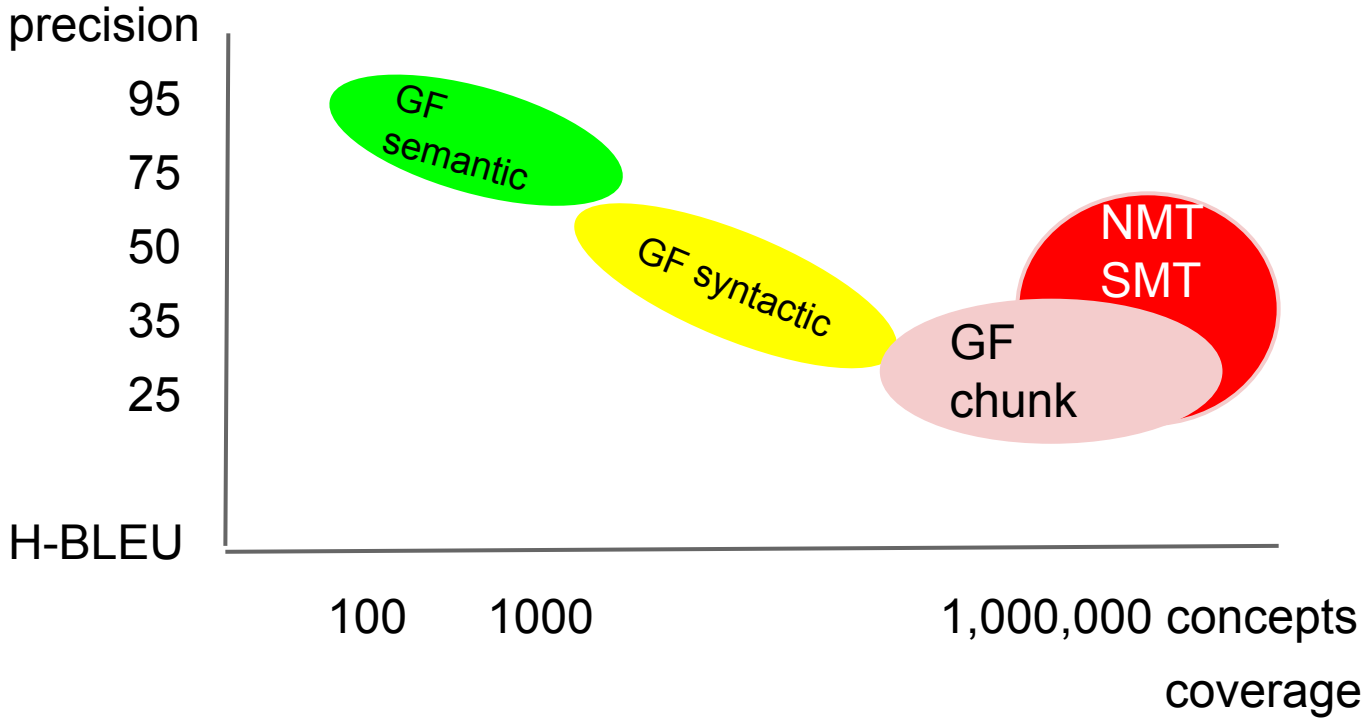
Grammar engineering: interlingua

Derive n^2 translation functions from n functions

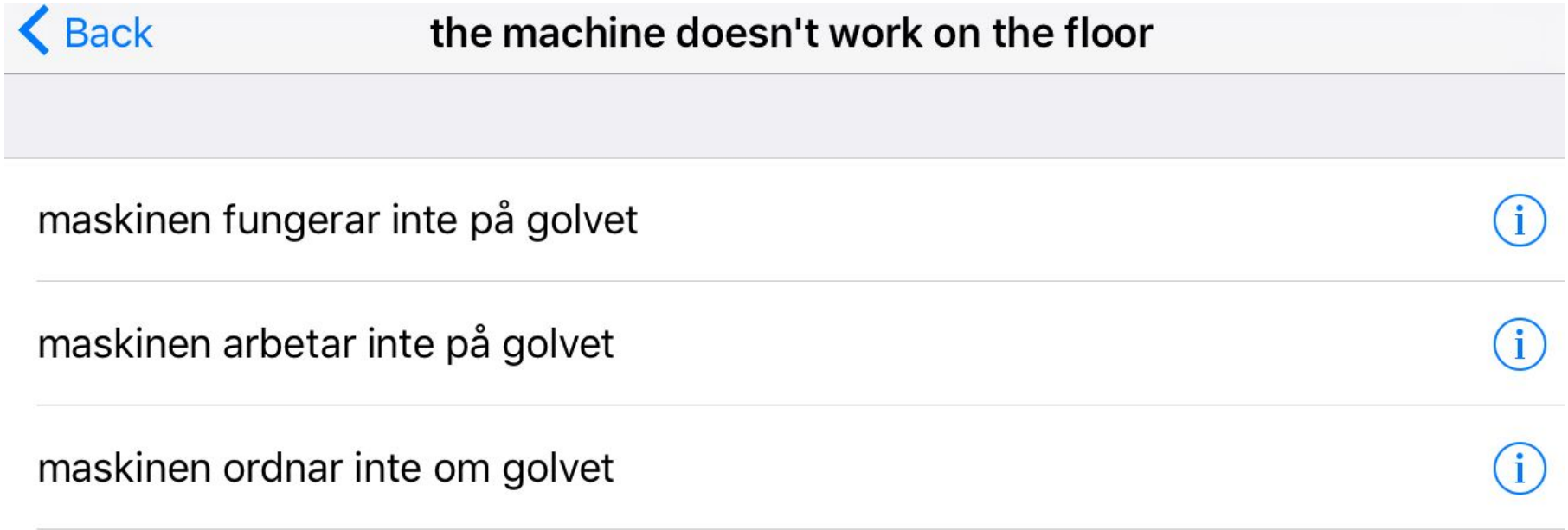

English
Dutch
Finnish
Chinese
French
Catalan
German
Interlingua
Bulgarian
Hindi
Thai
Italian
Swedish
Japanese
Spanish


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

Quality evaluation



Inspectability



Inspectability

Cost estimates

First grammar on cloud IDE	*minutes*
Commercial application grammar	*days*
Resource grammar	*months*
Wide coverage consumer grammar	*days*

Future trends of growth

Share with WordNet, Universal Dependencies, etc

Bootstrap with statistics and machine learning

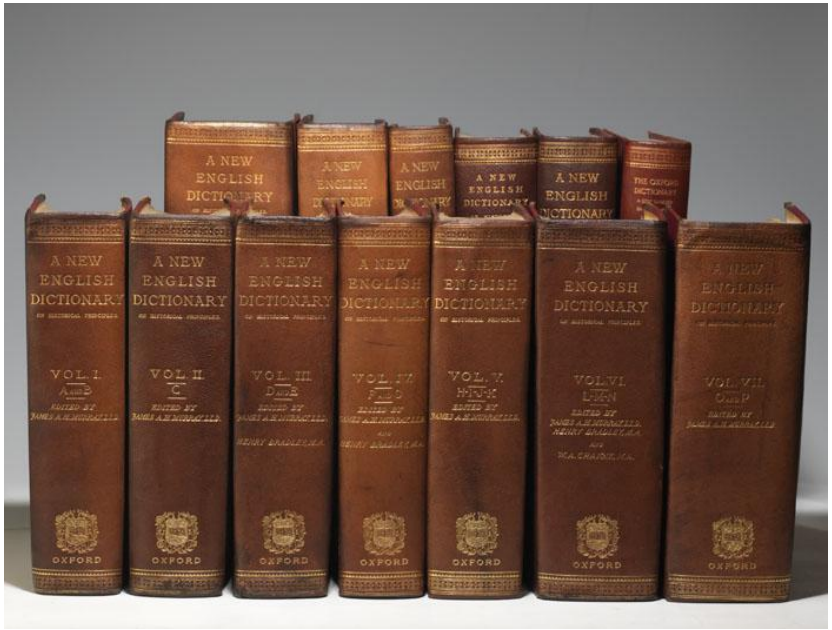
- but keep control of quality

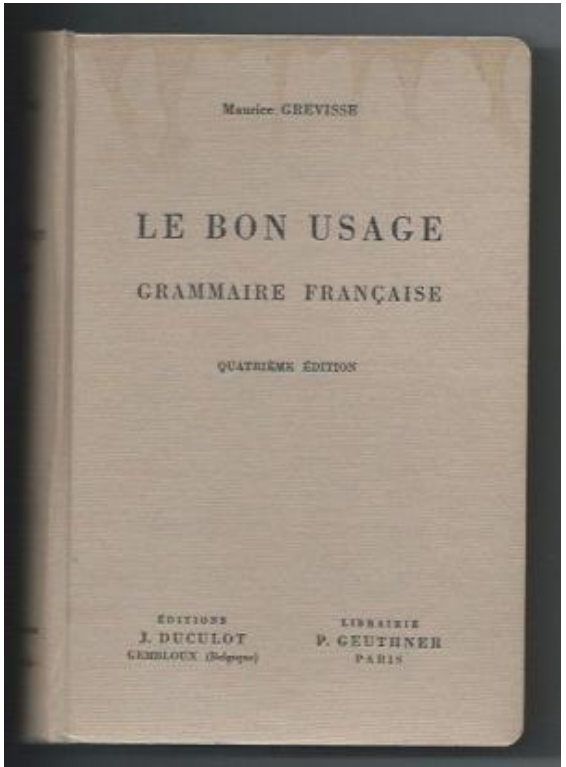
Optimize the use of human work - instead of eliminating it!

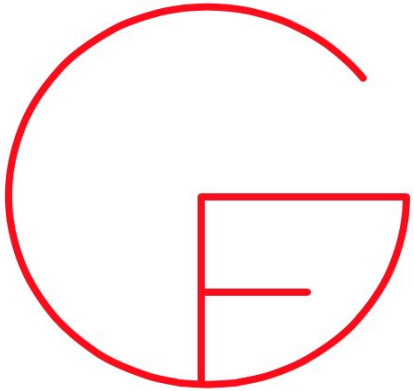
A hundred years perspective

We can build resources that will be useful 100 years from now

Just like dictionaries and grammars written 100 years ago







Open source, freely usable grammar software

32 languages

http://www.grammaticalframework.org/

digital Grammars

Language technology to rely on.

Customized GF solutions for producers of information

http://www.digitalgrammars.com/