# Kan man lita på språköversättning med maskin?

Aarne Ranta

Institutionen för data- och informationsteknik
Chalmers tekniska högskola och Göteborgs universitet
Café-å-lär, 7 november 2013, Göteborg

## En prognos

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

# En prognos

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

IBM press release 1954

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

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

## Vad jag vill tala om

Lite historia

Hur maskinöversättning fungerar

Vad som är lätt och vad som är svårt

Ansvarsfrågan

Vår forskning: GF, MOLTO, REMU

## Lite historia

**1946 kryptanalys**: ryska = krypterad engelska

- man behöver bara knäcka koden!

# Lite historia

1946 kryptanalys: ryska = krypterad engelska

- man behöver bara knäcka koden!

**1966 ALPAC report**: översättning med dator blir *dubbelt så dyrt* som manuell översättning

- man behöver lingvistisk forskning

# Lite historia

1946 kryptanalys: ryska = krypterad engelska

- man behöver bara knäcka koden!

1966 ALPAC report: översättning med dator blir *dubbelt så dyrt* som manuell översättning

- man behöver lingvistisk forskning

**1989 IBM**: kryptanalys igen - med starkare datorer

- inga språkkunskaper behövs

# Lite historia

1946 kryptanalys: ryska = krypterad engelska

- man behöver bara knäcka koden!

1966 ALPAC report: översättning med dator blir *dubbelt så dyrt* som manuell översättning

- man behöver lingvistisk forskning

1989 IBM: kryptanalys igen - med starkare datorer

- inga språkkunskaper behövs

**2006 Google translate**: numera 60 språk

- IBM-metoden + Googles enorma datamängder

# Två skolor

SMT = Statistical Machine Translation

- kryptanalys
- söker efter ”den sannolikaste översättningen” i ljuset av tidigare data
- exempel: Google translate

RBMT = Rule-Based Machine Translation

- lingvistik och dataspråksöversättning
- söker efter ”den rätta översättningen” som ska återskapa meningen
- exempel: GF (Grammatical Framework)

# Regelbaserad översättning

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!

# Kompilatorer

**Kompilator**: översätter programkod till maskinkod

```
printf("hello world")    --->   01010110110100010101010100101001
```

En kalkylator med språkregler

Här har datorn ersatt människan som översättare

# Regler i naturligt språk

**Morfologi**: böjning

- *nyckel* —> *nyckel, nyckeln, nycklar, nycklarna*
- *känna* —> *känner, kände, känt, känns, kändes, känts*

**Syntax**: kongruens, ordföljd

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

# Hur datorn kan hjälpa med reglerna

Morfologi och syntax har mekaniska regler

En infödd talare kan dem utantill, *i princip*

Men man kan lätt göra fel:

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

# Grammatikregler i automatisk översättning

Grammatiken varierar från språk till språk

- olika kongruensregler
- olika ordföljdsregler

Men språken kan ha samma **struktur**, s.k. **abstrakt syntax**

# Exempel

| | | |
|---|---|---|
| Abstrakt syntax | `Sats(subjekt,verb,objekt)` | `Sats(I,know,you)` |
| Engelska | `subjekt + verb + objekt` | *I know you* |
| Tyska, huvudsats | `subjekt + verb + objekt` | *ich kenne dich* |
| - bisats | `subjekt + objekt + verb` | *ich dich kenne* |
| - efter adverbial | `verb + subjekt + objekt` | *kenne ich dich* |
| Franska, vanligt objekt | `subjekt + verb + objekt` | *je connais cet homme* |
| - pronomen som objekt | `subjekt + objekt + verb` | *je te connais* |

## Problem: rätt analys

*jag åt en pizza med räkor*

## Problem: rätt analys

*jag åt en pizza med räkor*

*jag åt en pizza med vänner*

## Problem: rätt analys

*jag åt* (*en pizza med räkor*)

(*jag åt en pizza*) *med vänner*

## Problem: rätt analys

*jag åt* (*en pizza med räkor*)

(*jag åt en pizza*) *med vänner*

(*jag åt en pizza*) *med kniv och gaffel*

## Problem: idiomatiska kombinationer

| | |
|---|---|
| *var så god* | **be so good* |
| *jag heter X* | *my name is X* |
| *hur långt är det till X* | *how far is X* |
| *finlandsbåt* | *ruotsinlaiva* |

# Problem: det krävs kunskap och arbete

Känt sedan länge:

- grammatik
- lexikon

Nya idéer behövs:

- semantisk analys
- idiomatiska kombinationer

# Statistisk översättning

Automatisk analys

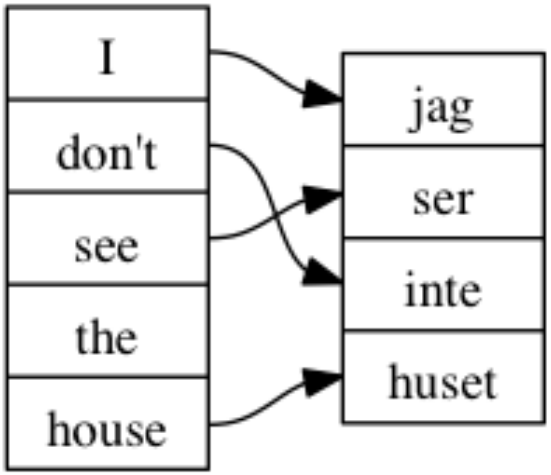
Inga språkkunskaper

Lexikon: **ordlinjering**

Syntax: **n-gram**

# Ordlinjering

Hitta ord som motsvarar varandra



# Val av linjering

| | |
|---|---|
| *house* | *hus*, *huset* |
| *houses* | *hus*, *husen* |
| *is* | *är* |
| *are* | *är* |
| *am* | *är* |
| *red* | *röd*, *rött*, *röda* |
| *this* | *det här*, *det där*, *denna*, *detta* |

*this house is red*:

*den här hus är röda*? *det här huset är rött*?

# Syntax med n-gram

*n*-gram = sekvens av *n* ord (*n* = 1,2,3,4,...)

3-gram i svensk text:

- **den här hus*
- *det här huset*
- **huset är röd*
- *huset är rött*

*det här huset* + *huset är rött* −> *det här huset är rött*

## Hjälper också med idiom

Frekventa n-gram fångas väl i modellen

*vice president*

-> *vice ordförande* (Google translate)

-> *skruvstädspresident* (GF baseline translator)

## Problem: glesa data

Ordlinjering: 1 miljon ord av parallell text

n-grammodell: 10 miljon ord text

Det behövs mer om språket har många böjningar eller varierande ordföljd

## Glesa data och böjningar

Alla finska ordformer har inte setts (Google translate 6 November 2013)

*yö, yön, yötä, yöksi, yönä, yössä, yöstä, yöhön, yöllä, yöltä, yölle, yöttä, yöt, öiden, öitä, öiksi, öinä, öissä, öistä, öihin, öillä, öiltä, öille, öittä, öine, öin*

*night, night, night, night, night, night, night, night, night, night, nights, Yotta, night, nights, nights nights, nights, nies, nights, nights, nights, with, company against loss, nities, öittä, öine, night*

## Problem: ”long distance dependencies”

Svensk kongruens försvinner

## Problem: ”long distance dependencies”

Svensk kongruens försvinner

(Google translate 6 November 2013)

| | |
|---|---|
| *This house is big.* | *Detta hus är stort.* |
| *This house is very big.* | *Detta hus är mycket stor.* |

## Problem: ”long distance dependencies”, del 2

Lite allvarligare: tyska *um* och *bringen* hamnar isär

## Problem: ”long distance dependencies”, del 2

Lite allvarligare: tyska *um* och *bringen* hamnar isär

(Google translate 6 November 2013)

| | |
|---|---|
| *Er bringt dich um.* | *He will kill you.* |
| *Er bringt deinen Freund um.* | *He brings to your friend.* |

## Problem: alla ord är lika viktiga

Modellen optimeras för ordsekvenser

Ett missat ord betyder bara lite

## Problem: alla ord är lika viktiga

Modellen optimeras för ordsekvenser

Ett missat ord betyder bara lite

(Google translate 6 November 2013)

| | |
|---|---|
| *Min fru är svensk.* | *Meine Frau ist Schwedisch.* |
| *Min fru är inte svensk.* | *Meine Frau ist Schwedisch.* |

## Mellanspråk

Problem: för lite parallell data svenska-bulgariska

Lösning: använd engelska som mellanspråk

- svenska-bulgariska = svenska-engelska + engelska-bulgariska

## Problem: fel ordlinjering

Google translate, Summer 2010

*Jag är svensk. -> Ich bin ein Amerikaner.*

## Förklaring

1. Parallella texter är **lokaliserade**, inte bara översatta.

2. Användning av engelska som mellanspråk.

*Jag är svensk -> I am American -> Ich bin ein Amerikaner*

## Ansvarsfrågan

Fransk e-handelbutik:

- *prix 99 euros*

Svensk översättning:

- *pris 99 kronor*

Kan den svenske konsumenten kräva att få varan för 99 kronor?

# Producent vs. konsument

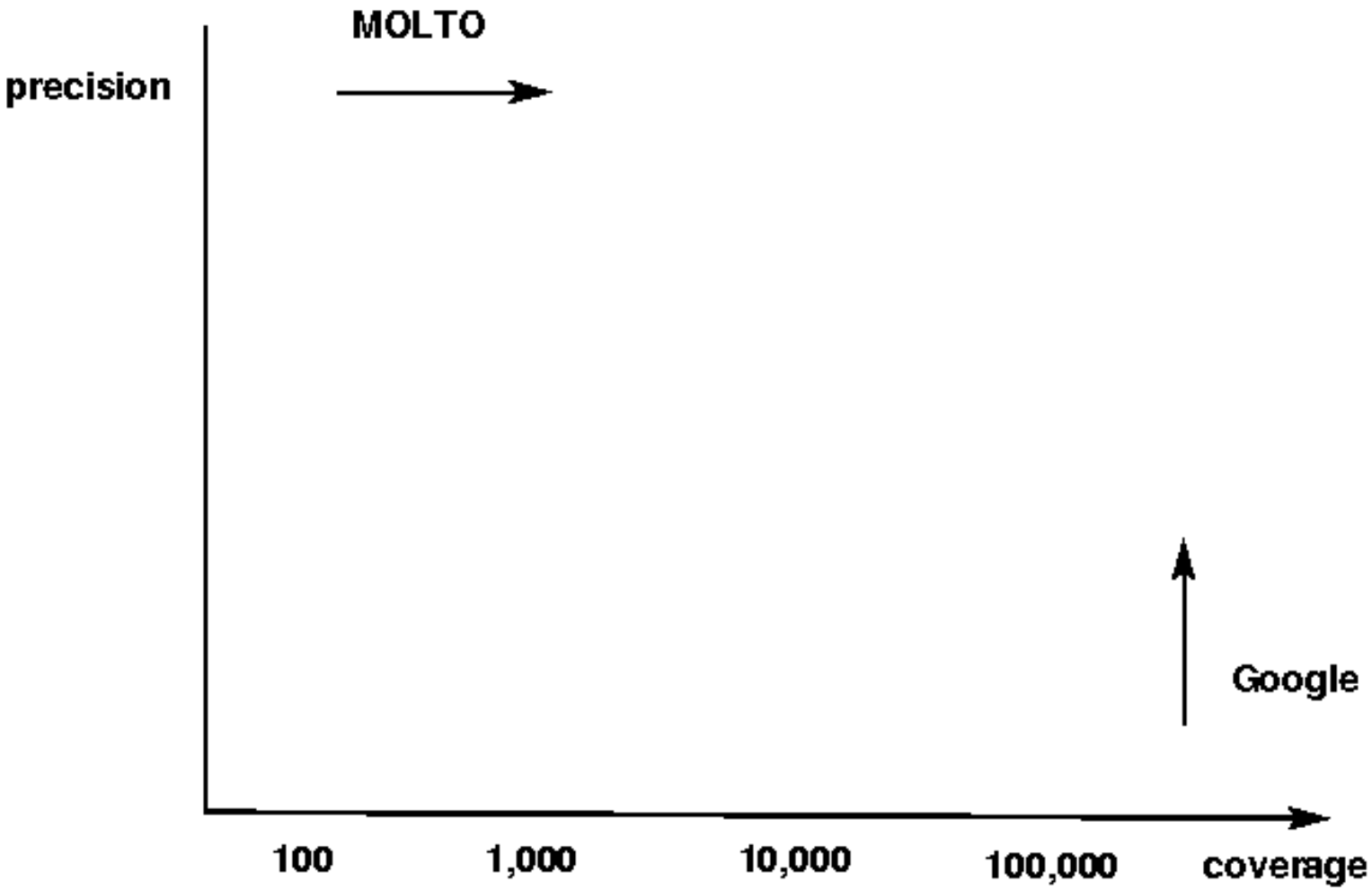
*prix 99 euros - pris 99 kronor*

Kan den svenske konsumenten kräva att få varan för 99 kronor?

Beror på vem som har gjort översättningen

- **producenten** (den franska butiken): Ja!
- **konsimenten** (den svenske användaren): Nej!

# Två forskningslinjer




MOLTO = Multilingual Online Translation, EU-projekt 2010-2013

# Ett exempel från MOLTO

Texter om konstverk (t.ex. Göteborgs stadsmuseum, Wikipedia)

Källa: formell beskrivning i databas (Göteborgs stadsmuseum, Wikipedia)

```
MkGenText GSM980019Obj AnnaLindskog OilPainting
  (MkColour Black) (MkSize (SIntInt 435 365))
  (MkMaterial Canvas) (MkYear (YInt 1885))
  (MkMuseum GoteborgsCityMuseum)
```

Mål: 15 språk

- PaintingEng: The girl was painted on canvas by Anna Lindskog in 1885. It is of size 435 by 365 and it is painted in black. This oil painting is displayed at the City Museum of Gothenburg.

- PaintingFin: Maalauksen Flickan on maalannut Anna Lindskog kankaalle vuonna 1885. Se on kokoa 435 kertaa 365 ja se on maalattu mustalla. Tämä öljymaalaus on esillä Göteborgin kaupunginmuseossa.

- PaintingFre: Le tableau Flickan a été peint sur toile par Anna Lindskog en 1885. Il est de taille 435 sur 365 et il est peint en noir. Cette peinture à l' huile est exposée dans le musée municipal de Göteborg.

- PaintingIta: Il quadro Flickan è stato dipinto su tela da Anna Lindskog nel 1885. Misura 435 per 365 ed è dipinto in nero. Questo dipinto ad olio è esposto nel museo municipale di Goteburgo.

- PaintingSwe: Flickan målades på duk av Anna Lindskog år 1885. Den är av storlek 435 gånger 365 och den är målad i svart. Den här oljemålningen är utställd på Göteborgs stadsmuseum.

# Det traditionella metodvalet

Producent: regelbaserat

- förutsägbart
- kontrollerbart
- begränsat

Konsument: statistiskt

- obegränsat
- ”tillräckligt bra”

## Hybridmetoder

Kombinera det bästa av RBMT och SMT

RBMT: grammatik

SMT: idiomatiska fraser, automatiskt skapade regler

**Den viktigaste forskningslinjen både här och annanstans**

*2013-10-29 Ramona Enaches disputation med Keith Hall från Google som opponent*

# Demo 1

GF cloud

- prediktiv parsning
- ordlinjering (word alignment)
- böjningar
- 29 språk

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

Android app: Phrasedroid

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

# Demo 2

Hybrid översättning med GF

Android app (prototyp): tal-till-tal översättning på Android med engelska, finska, kinesiska och svenska

# REMU = Reliable Multilingual Digital Communication

VR Framework Grant 2013-2017

# Global Week: Coding the Grammars of the World

Tuesday 12 November 2013 at 4:00 PM - 5:00 PM

University main building Vasaparken
Torgny Segerstedtsalen, 2nd floor

Lecturers:
Aarne Ranta, Ramona Enache, Inari Listenmaa and John J. Camilleri

# Thanks: World-Wide GF Community