INF5820

Language technological applications

H2008

Jan Tore Lønning/Stephan Oepen

jtl@ifi.uio.no/oe@ifi.uio.no

Maskinoversettelse

INF 5820 – H2008
Forelesning 2

Maskinoversettelse

1. Hva er maskinoversettelse
2. Hvorfor er det vanskelig?
3. Tradisjonelle tilnærminger:
    1. Direkte
    2. Interlingua
    3. Transfer
4. Empiriske tilnærminger:
    1. Eksempelbasert MT (EBMT)
    2. Statistisk MT (SMT)
5. LOGON-prosjektet
6. Evaluering
7. Maskinoversettelse i praksis
8. Litt historie

Hva er maskinoversettelse?

- Hva er en oversettelse?
- Vanlig tilnærming:
  1. Hva er problemet?
  2. Hva definerer en korrekt løsning?
  3. Algoritme for å finne en korrekt løsning?
- Problem pkt. 2:
  - Ikke en entydig korrekt oversettelse
  - Mer eller mindre gode

Maskinoversettelse

1. Hva er maskinoversettelse
2. Hvorfor er det vanskelig?
3. Tradisjonelle tilnærminger:
    1. Direkte
    2. Interlingua
    3. Transfer
4. Empiriske tilnærminger:
    1. Eksempelbasert MT (EBMT)
    2. Statistisk MT (SMT)
5. LOGON-prosjektet
6. Evaluering
7. Maskinoversettelse i praksis
8. Litt historie

Eksempel: freetranslation.com

The construction of the park lasted for a number of years. The area east of the two Frogner ponds had already by the turn of the century been opened to the public.

freetranslation.com

Konstruksjonen av parken vart for et antall år. Områdesom øst av den to Frogner damer hatt allerede ved det vender av århundret vært åpnet til offentligheten.

Utvikling (2006→2008)

- The construction of the park lasted for a number of years. The area east of the two Frogner ponds had already by the turn of the century been opened to the public.
- freetranslation.com :
  Konstruksjonen av parken vart for et antall år. Områdesom øst av den to Frogner damer hatt allerede ved det vender av århundret vært åpnet til offentligheten.
- Google:
  Konstruksjonen av parken varte i en årrekke. Området øst for de to Frogner dammer hadde allerede ved århundreskiftet blitt åpnet for allmennheten.

Eksempel 2

- http://www.systransoft.com/index.html
- http://www.systranet.com/systran/net
- http://www.heeg.de/~uta/iX/art.htm
- www.dn.se

Ikke bare ord for ord:

1. Grammatisk struktur- flertydig
2. Gram. struktur bevares ikke
3. Ordvalg - flertydig
4. Mer enn setningsmening

1. Flertydig struktur

De satte pris på dyrene

De uppskattade djuren

gir oss kjærlighet, vennskap og innimellom beskyttelse.

ger oss kärlek, vänskap och ibland skydd.

- Behov for fullstendig grammatisk analyse
- Fra 1950-tallet: Utvikling av grammatikker egnet for dataprosessering

2. Gram. struktur bevares ikke

Han heter Paul.
His name is Paul.
Il s'appelle Paul.

He likes to swim.
Er schwimmt gern.

3. Ordvalg

Jeg skar av makens tomme!
Jag skar av makens tumme!

The box is in the pen.
(Y. Bar-Hillel 1960)

- Problemet er ”AI-hardt” –
  kan ikke vente fullgod løsning

4. Mer enn setningsmening

- Større enheter, avsnitt
- Holde rede på referenter (f.eks. den/det)
- Metaforer, idiomer
- Stil, tone
- Rim, rytme
- Flertydigheter, humor
- …

Typologiske språkforskjeller

- J&M peker på typologiske forskjeller som et mulig problem:
  - Polysentetisk vs. isolerende språk
  - Markering på hodet eller avhengige ledd
- Typologiske forskjeller behøver ikke være et problem for MT:
  - SVO vs. SOV
  - The man – mannen
- Men kan være det:
  - Språk med og uten definitthet
  - aspekt

Mer om ordvalg

- I kildespråket
  - Grammatisk forskjellige:
    - løp, løper, bygg, murer, fisker (Dorr et. al: pkt. 2)
  - Homografer innen samme ordklasse:
    - ball bygg, mangle, himle, fil (pkt. 3.i)
  - Polysemer:
    - hode, …(pkt. 3.ii)
- I overgangen:
  - Valg av realisering:
    - teach vs. learn, farmor vs mormor, (7)
  - Ikke direkte match:
    - morgen-morning, legg-leg, … (jf J&M)

Lingvistiske betraktninger

- Flere dimensjoner:
  - analyse vs transfer vs realisering
  - leksemer vs mer komplekse strukturer
  - det løsbare og det som forblir vanskelig
- pkt 4 kompleks semantisk ambiguitet:
  - leksikalsk homografi som løses av kontekst
- pkt 5 kontekstuell ambiguitet:
  - nødvendigheten av å holde rede på referenten
  - at det kan være mulig noen ganger
- pkt 6 kompleks kontekstuell flertydighet:
  - syntaktisk flertydighet (som i pkt 1)
  - og hvordan den kan løses

Lingvistiske betraktninger

- Typologiske forskjeller: pkt. 8 & 9
- Klasse 3:
  - tematisk mismatch, eller argument switching
  - head switching
  - strukturell divergens
  - kategori divergens
  - konflatusjonell divergens

Maskinoversettelse

1. Hva er maskinoversettelse
2. Hvorfor er det vanskelig?
3. Tradisjonelle tilnærminger:
    1. Direkte
    2. Interlingua
    3. Transfer
4. Empiriske tilnærminger:
    1. Eksempelbasert MT (EBMT)
    2. Statistisk MT (SMT)
5. LOGON-prosjektet
6. Evaluering
7. Maskinoversettelse i praksis
8. Litt historie

1. Realskolealgoritmen

S N B E			V Pr	V PP	H D 3p E		O A U F	
Jenta	fra	byen	har	gitt	ham		noen	røde epler
Mädchen	von	Stadt	haben	geben	er		einige rot	Apfel
Das Mädchen	von der	Stadt	hat		ihm		einige rote	Äpfel
								gegeben

1. Identifiser verb, syntaktisk funksjon og kasus,
2. og morfosyntaktiske trekk: bestemthet, tall, pers., form, tid, …
3. Finn oversettelser av leksemene.
4. Egenskaper ved leksemene: kjønn, styring, samsvar, ..
5. Bøyning: Kasus, tall, person, kjønn, bestemth., tid, samsvar, …
6. Ordstilling.

1. Direkte oversettelse

- Hovedidé: Det er ord som skal oversettes
- To-språklig ordbok
- En viss morfologisk analyse
- To trinn:
  - Finn ordene
  - Finn rekkefølgen mellom dem
  - (Husk dette når vi kommer til SMT)
- J&M: Decision list algorithm

2. Interlingua

- Et universelt meningsrepresentasjonsspråk (lingua franca)
- Setningen i kildespråket analyseres
- Analysen resulterer i en interlingua representasjon
- Fra denne genereres setning i andre språk

2. Interlinguas styrke

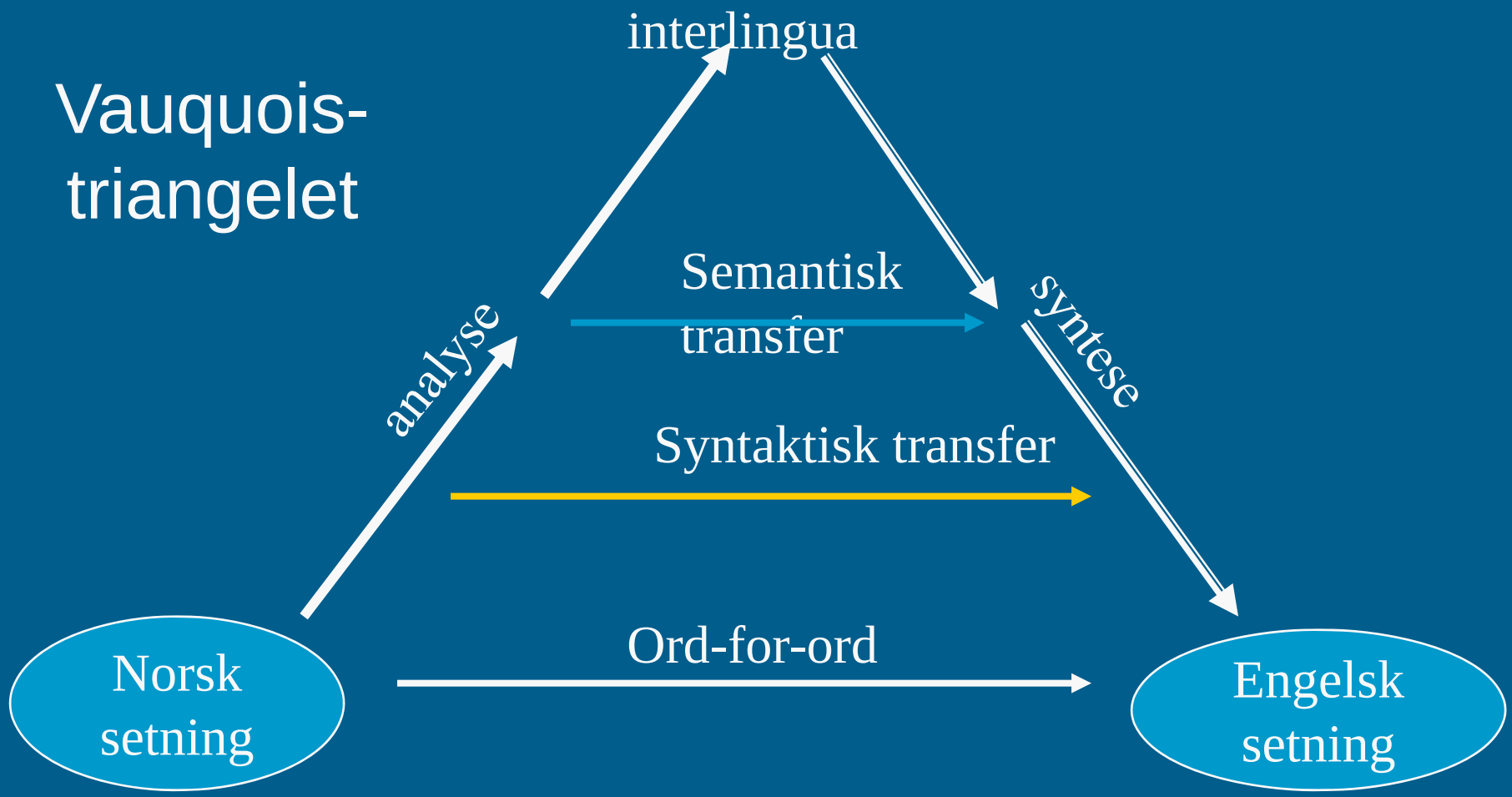
- Oversettelse mellom mange språk.
- Det trengs én analysemodul og én genereringsmodul per språk.
- Eksempel 17 språk:
  - Direkte 17*16 moduler (=272)
  - Interlingua 2*17 (=34)
- Språk 18:
  - Direkte +(2*17)
  - Interlingua +2

3. Transfer

- Problem for interlingua:
  - en språkuavhengig, universell representasjon
- Transfer tilnærming:
  - språkavhengige representasjoner
  - Kontrasten mellom to språk som transferregler
- Syntaktisk transfer tilnærming:
  - Utvider den direkte tilnærmingene med syntaktisk analyse
- Semantisk tilnærming
  - Semantiske representasjoner, men språkavhengige

Alternative strategier

Vauquois-triangelet



Maskinoversettelse

1. Hva er maskinoversettelse
2. Hvorfor er det vanskelig?
3. Tradisjonelle tilnærminger:
    1. Direkte
    2. Interlingua
    3. Transfer
4. Empiriske tilnærminger:
    1. Eksempelbasert MT (EBMT)
    2. Statistisk MT (SMT)
5. LOGON-prosjektet
6. Evaluering
7. Maskinoversettelse i praksis
8. Litt historie

Eksempelbasert MT

- No: Jenta har lest lekser i en time.
- Eng: ?
- Eksempler:
  - Jenta har spist et eple hver dag
  - The girl has eaten an apple a day
  - Per hadde lest lekser
  - Per had studied
  - Kari sang i en time.
  - Kari sang for an hour.
- Ikke bare konstituenter

Statistikkbasert maskinoversettelse (SMT)

- http://people.csail.mit.edu/koehn/publications/tutorial2003.pdf 3-5
- http://www.iccs.informatics.ed.ac.uk/~pkoehn/publications/esslli-slides-day1.pdf 7ff.

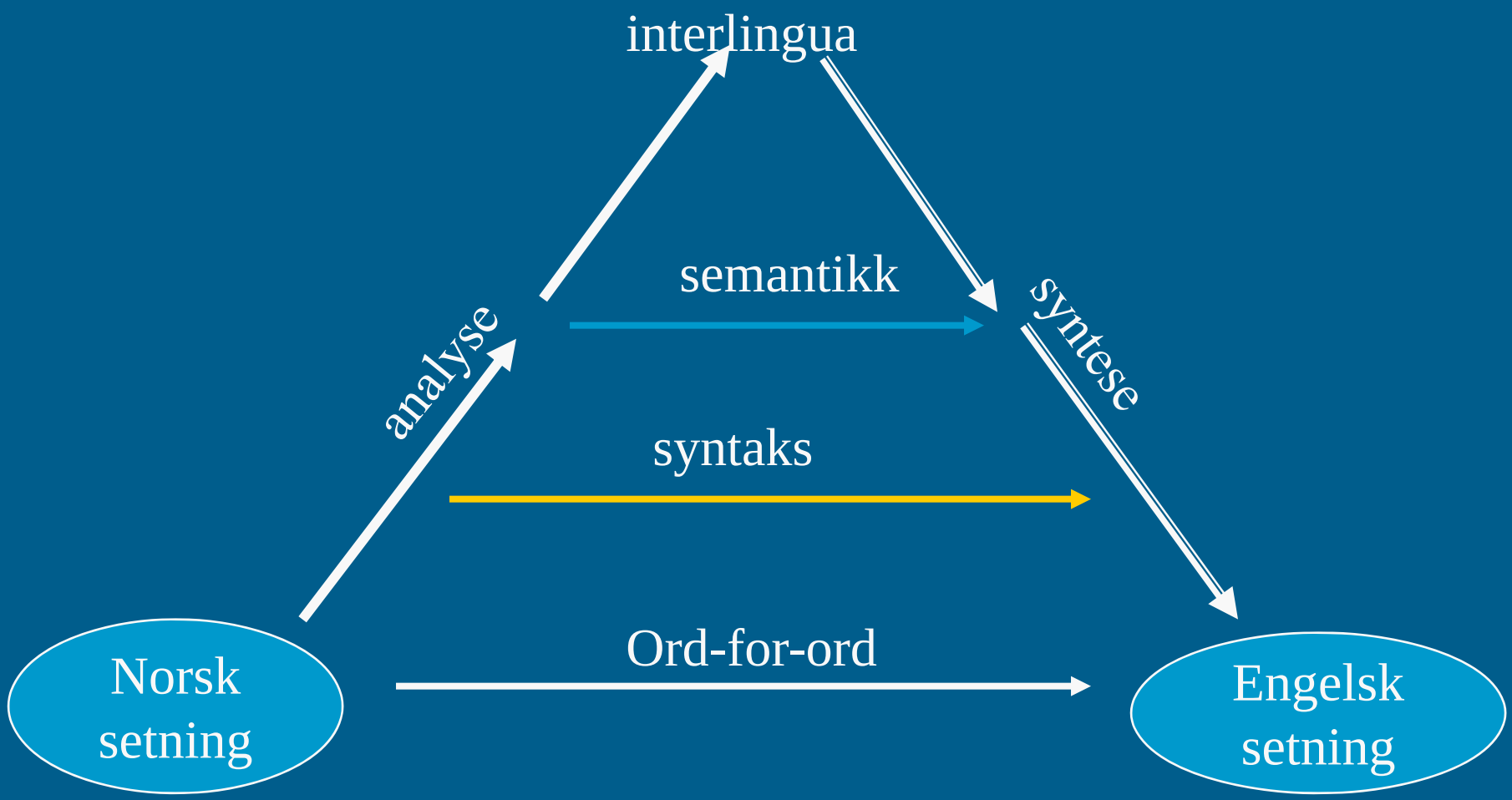
Maskinoversettelse

1. Hva er maskinoversettelse
2. Hvorfor er det vanskelig?
3. Tradisjonelle tilnærminger:
    1. Direkte
    2. Interlingua
    3. Transfer
4. Empiriske tilnærminger:
    1. Eksempelbasert MT (EBMT)
    2. Statistisk MT (SMT)
5. LOGON-prosjektet
6. Evaluering
7. Maskinoversettelse i praksis
8. Litt historie

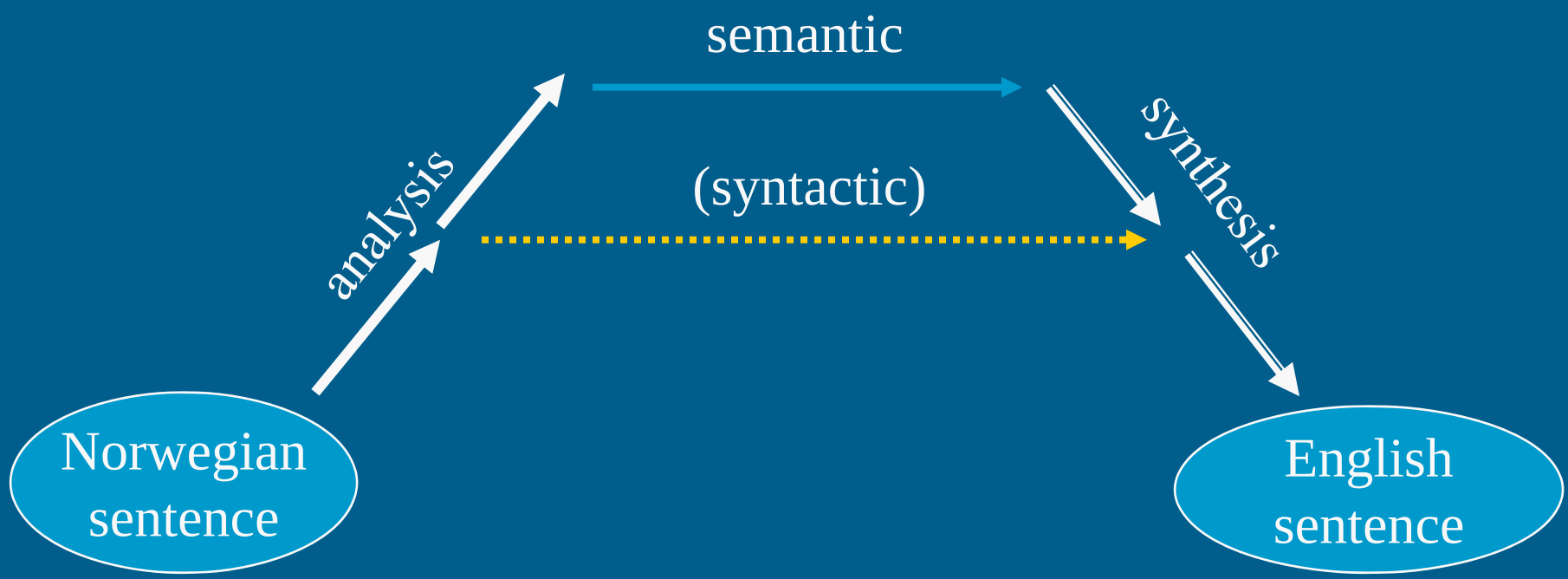
Mål

- Maskinoversettelse norsk → engelsk
  - Mange områder av språkteknologi trengs:
  - Samvirke i en demonstrator
  - Likheter og forskjeller mellom norsk og andre språk
- Turisttekster/turbeskrivelser
- Høykvalitet, (begrenset dekning)
- 2003-2007

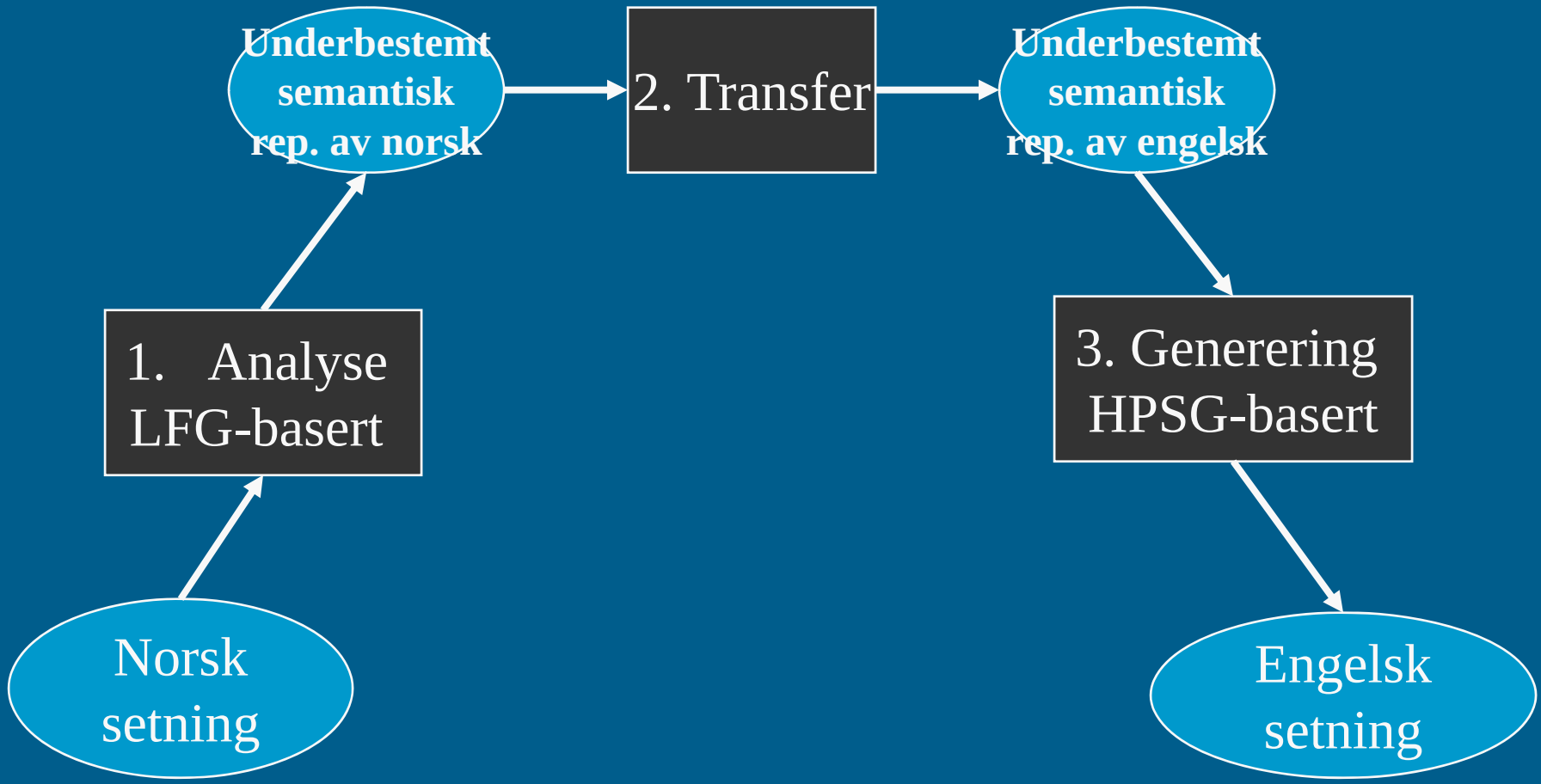
Alternative strategier



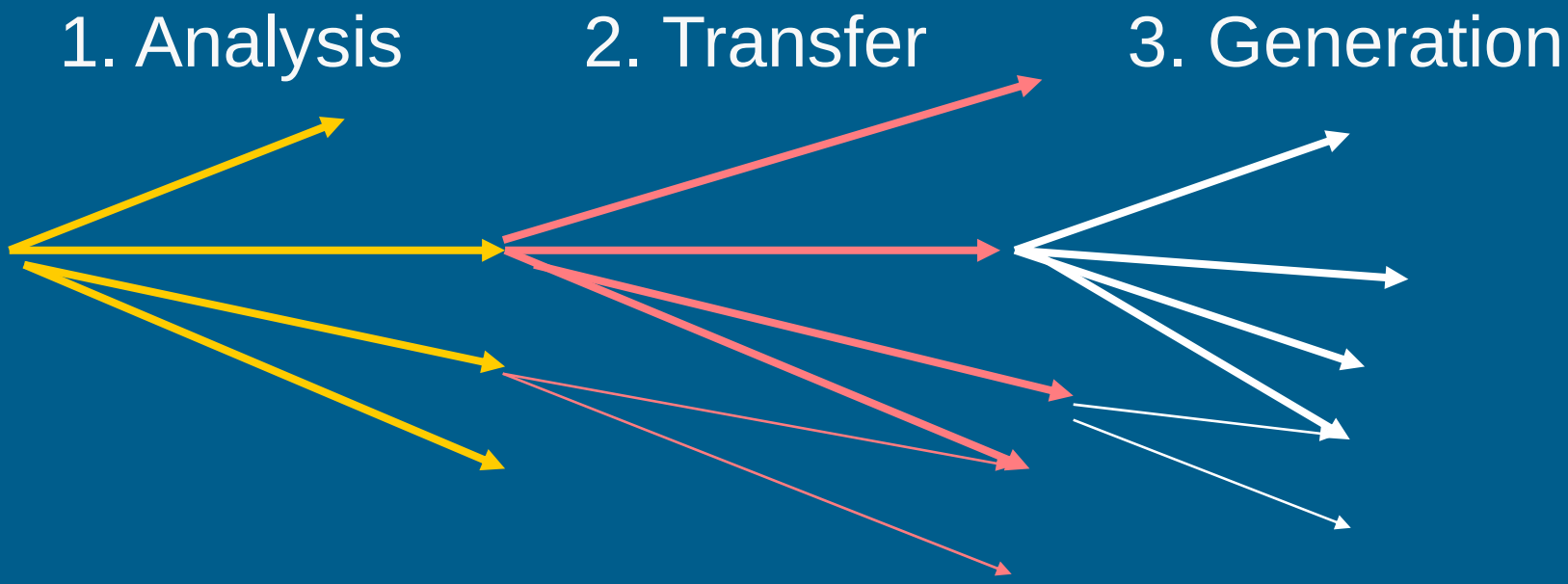
MT strategies (symbolic)



Basis: Transferbasert oversettelse



2.2 Flertydighet



- Hvordan velge den rette eller beste på hvert trinn?

Demo

- handon.emmtee.net

- |< |Toppen er luftig, og har en utrolig utsikt!| (83) --- 2 x 24 x 12 = 12
- |> |the top is airy and has an incredible view| [85.9] <0.70> (1:0:0).
- |> |the summit is airy and has an incredible view| [87.4] <1.00> (1:4:0).
- |> |the top is breezy and has an incredible view| [87.7] <0.46> (1:6:0).
- |> |the top is airy and has an unbelievable view| [88.9] <0.70> (1:1:0).
- |> |the peak is airy and has an incredible view| [89.1] <0.96> (1:2:0).
- |> |the summit is breezy and has an incredible view| [89.1] <0.66> (1:10:0).
- |> |the summit is airy and has an unbelievable view| [90.3] <1.00> (1:5:0).
- |> |the top is breezy and has an unbelievable view| [90.7] <0.46> (1:7:0).
- |> |the peak is breezy and has an incredible view| [90.8] <0.66> (1:8:0).
- |> |the peak is airy and has an unbelievable view| [92.0] <0.96> (1:3:0).
- |> |the summit is breezy and has an unbelievable view| [92.1] <0.66> (1:11:0).
- |> |the peak is breezy and has an unbelievable view| [93.8] <0.66> (1:9:0).
- |= 64:19 of 83 {77.1+22.9}; 58:9 of 64:19 {90.6 47.4}; 55:9 of 58:9 {94.8 100.0} @ 64 of 83 {77.1} <0.51 0.67>.

- |< |De slipper å bære.| (70) --- 3 x 4 x 9 = 6 [9]
- |> |they do not have to carry something| [40.6] <0.84> (0:0:1).
- |> |you do not have to carry something| [41.8] <0.53> (1:0:1).
- |> |those do not have to carry something| [51.6] <0.53> (2:1:1).
- |> |they don't have to carry something| [55.2] <0.80> (0:0:0).
- |> |you don't have to carry something| [65.8] <0.43> (1:0:0).
- |> |those don't have to carry something| [66.3] <0.43> (2:1:0).
- |= 57:13 of 70 {81.4+18.6}; 51:6 of 57:13 {89.5 46.2}; 48:6 of 51:6 {94.1 100.0} @ 54 of 70 {77.1} <0.53 0.69>.

Maskinoversettelse

1. Hva er maskinoversettelse
2. Hvorfor er det vanskelig?
3. Tradisjonelle tilnærminger:
    1. Direkte
    2. Interlingua
    3. Transfer
4. Empiriske tilnærminger:
    1. Eksempelbasert MT (EBMT)
    2. Statistisk MT (SMT)
5. LOGON-prosjektet
6. Evaluering
7. Maskinoversettelse i praksis
8. Litt historie

Evaluering

- Hvor godt er dette MT-systemet?
- Er dette systemet brukbart
  - For hva
- Hvilket er best av disse to systemene
  - For hva?

Menneskelig evaluering

- ”Black box”:
  - Spørre mennesker om å rangere resultatet med hensyn til ulike parametre.
- ”Glass box”:
  - Se på deler og komponenter
  - Følge opp utviklingen av systemet
- Vurdere ulike egenskaper:
  - Tid
  - Pris
  - Domene

Mot automatisk evaluering

- Problem: Hva er en god oversettelse?
- Mangler fasit.
- Ulike automatiske metoder for å måle korrespondansen mot en mengde referanseoversettelser

Maskinoversettelse

1. Hva er maskinoversettelse
2. Hvorfor er det vanskelig?
3. Tradisjonelle tilnærminger:
    1. Direkte
    2. Interlingua
    3. Transfer
4. Empiriske tilnærminger:
    1. Eksempelbasert MT (EBMT)
    2. Statistisk MT (SMT)
5. LOGON-prosjektet
6. Evaluering
7. Maskinoversettelse i praksis
8. Litt historie

Maskinoversettelse i praksis

Mål: Fully automatic, high-quality, (unrestricted) translation (=FAHQT)

Maskinoversettelse i praksis

Mål: Fully automatic, high-quality, (unrestricted) translation

1. Begrenset
   - (eks, METEO)

Maskinoversettelse i praksis

Mål: Fully automatic, high-quality, (unrestricted) translation

1. Begrenset
2. Lav kvalitet:
   - Resymé
   - Etterretning
   - Web

Maskinoversettelse i praksis

Mål: Fully automatic, high-quality, (unrestricted) translation

1. Begrenset
2. Lav kvalitet:
   - Resymé
   - Etterretning
   - Web
3. Semi-automatisk

Maskinoversettelse i praksis

Forredigering

Etterredigering

- Regningssvarende,
  - men kjedelig arbeid
- Kvalitet?
  - Hvordan måle?

- Faktorer:
  - Språkpar
  - Teksttype
  - kvalifikasjoner, …

Interaktiv, datastøttet oversettelse

- ”Translator’s workbench”:
  - Datastøttet menneskelig oversettelse,
  - Menneskehjulpet maskinoversettelse
- ”Translation memory”

- Visjon: ideell arbeidsdeling menneske-maskin

Maskinoversettelse

1. Hva er maskinoversettelse
2. Hvorfor er det vanskelig?
3. Tradisjonelle tilnærminger:
    1. Direkte
    2. Interlingua
    3. Transfer
4. Empiriske tilnærminger:
    1. Eksempelbasert MT (EBMT)
    2. Statistisk MT (SMT)
5. LOGON-prosjektet
6. Evaluering
7. Maskinoversettelse i praksis
8. Litt historie

Historien

- 1950-årene: stor optimisme (FAHQT)
- 1960-årene: for vanskelig
  - Bar-Hillel
  - ALPAC-rapporten
- 1980-årene-fornyet interesse:
  - Japan
  - EU, Eurotra

Vår tid (1992 →)

- Anvendelser
  - Hyllevare for PC-er
  - WWW
  - Interaktive oversettelsesverktøy
  - Nye markeder: Kina
- Teori
  - Taleoversettelse, eks. VerbMobil, tysk prosjekt
  - SMT, EMT

SMTs tidsalder

- Fra 1990
- Med som et alternativ på slutten av VerbMobil
- På markedet fra ca. 2003
- Google:
  - SMT fra ca 2005
  - Overbevisende kvalitet
  - Mange språkpar
  - Men forutsigbare feil