# INF5820

## Language technological applications

H2008

Jan Tore Lønning/Stephan Oepen

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

# LOGON-systemet

INF 5820 – H2008
Forelesning 17

3. Nov 2008

# Høstens 2. forelesning

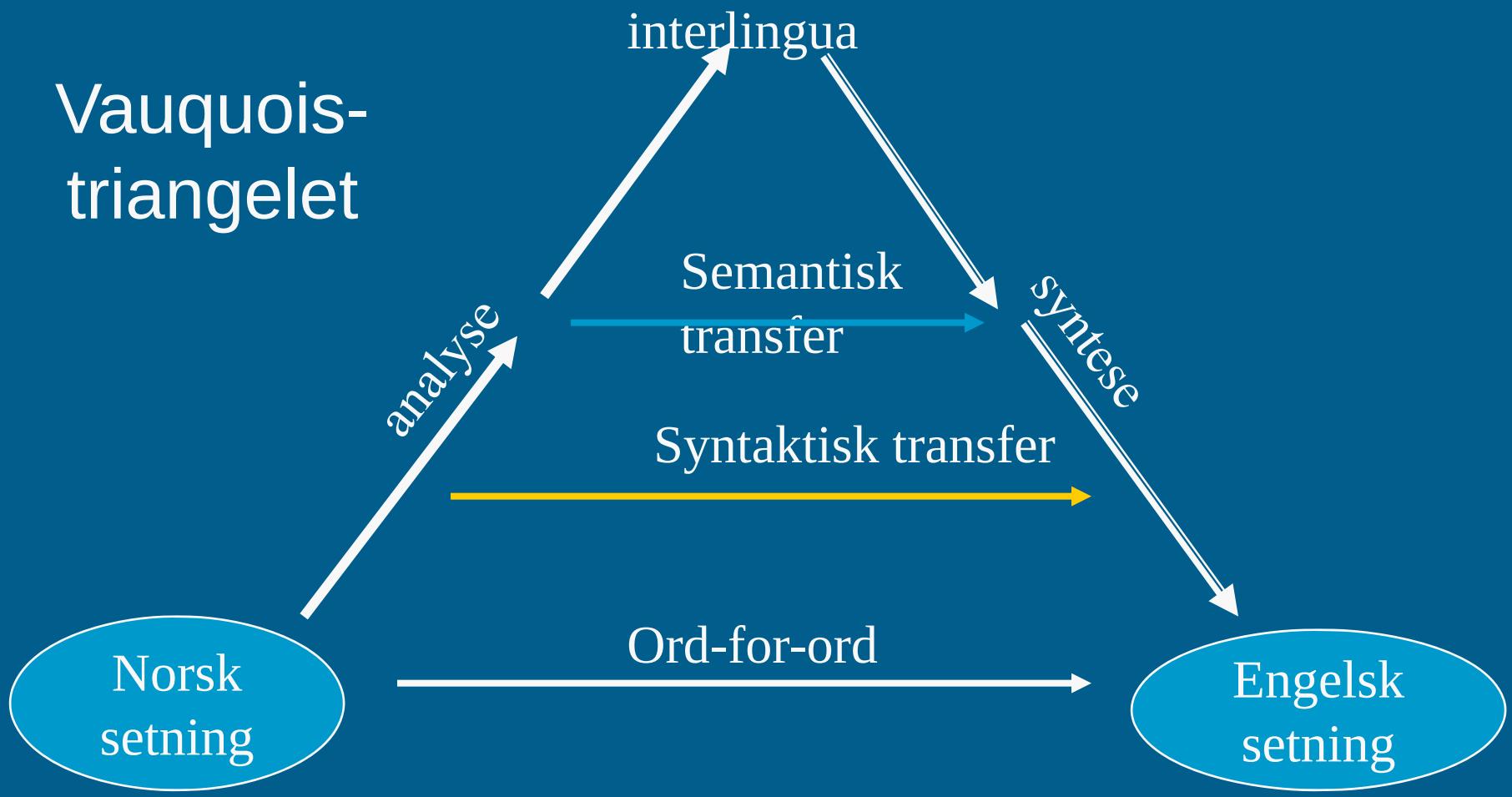
1. Hva er maskinoversettelse
2. Hvorfor er det vanskelig?
3. Tradisjonelle tilnærminger:
    1. Direkte
    2. Interlingua
    3. Transfer
4. Empiriske tilnærminger:
    1. Eksempelbasert MT (EBMT)
    2. Statistisk MT (SMT)
5. LOGON-prosjektet
6. Evaluering
7. Maskinoversettelse i praksis
8. Litt historie

# 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

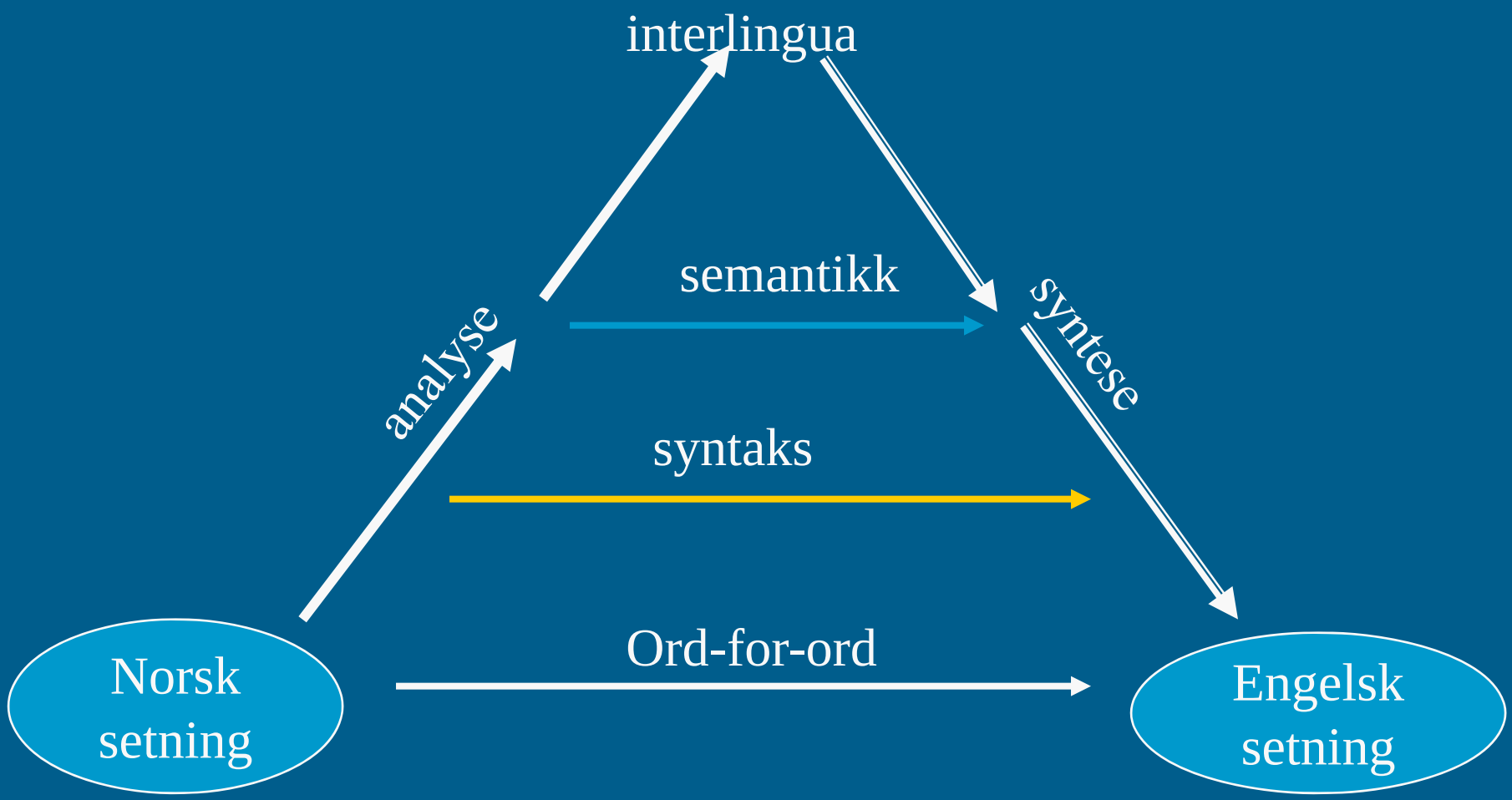
Vauquois-triangelet



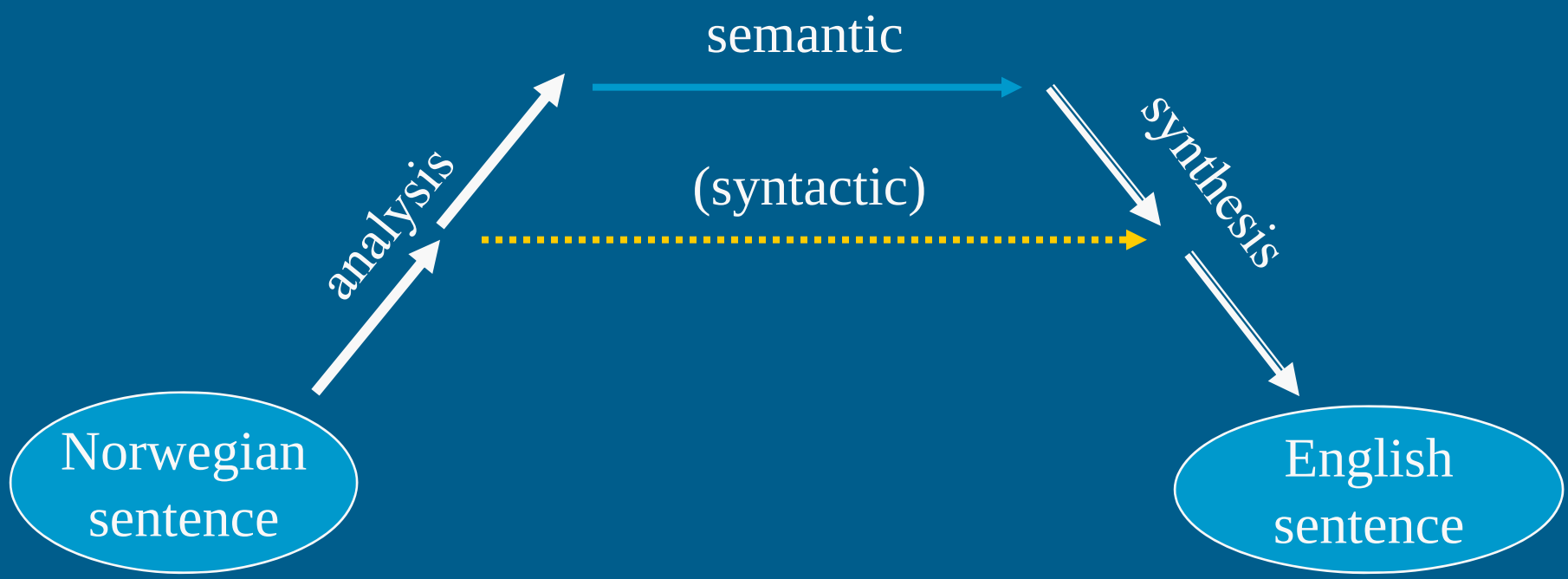
# Mål

- 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

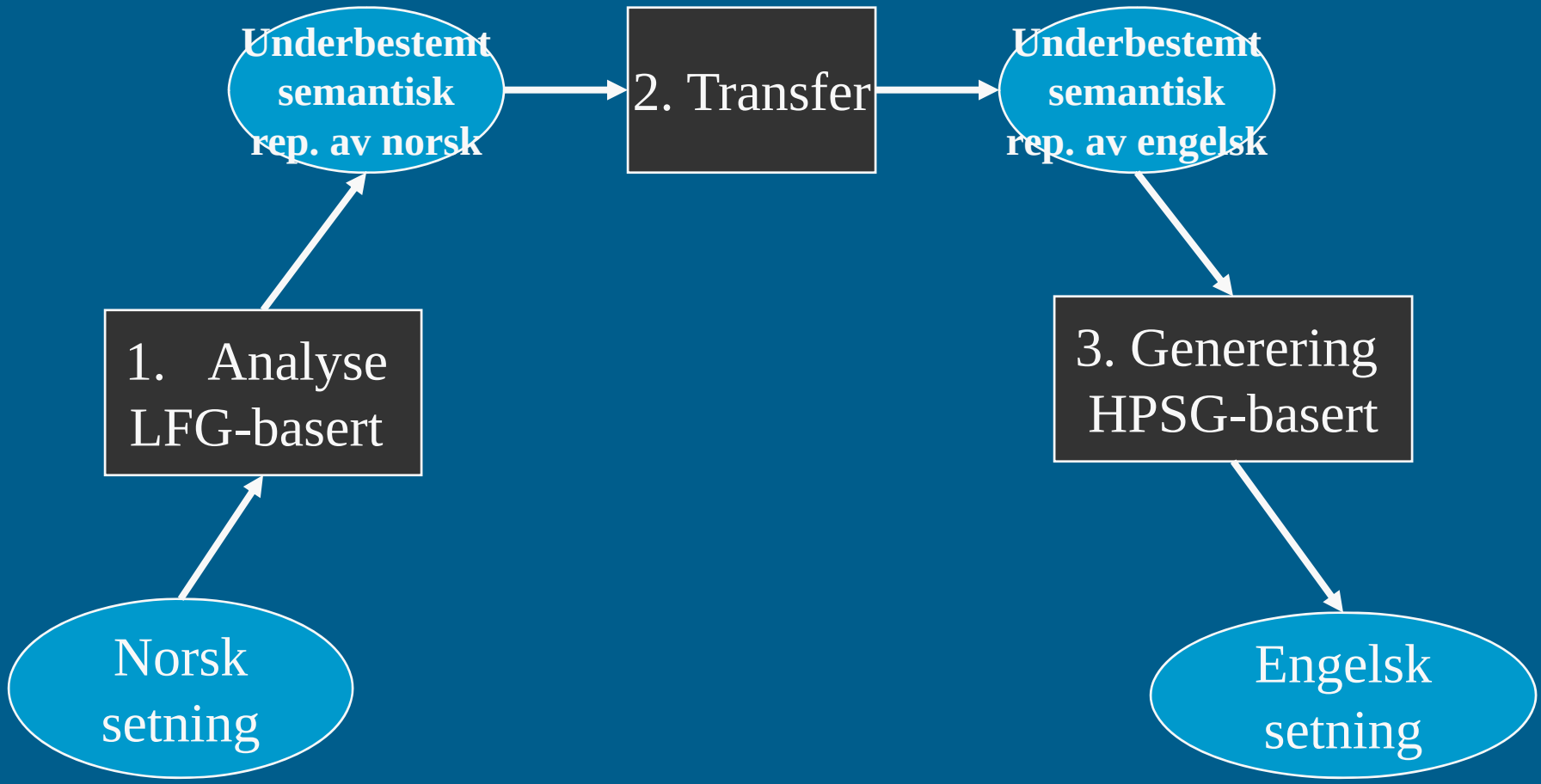
# Alternative strategier



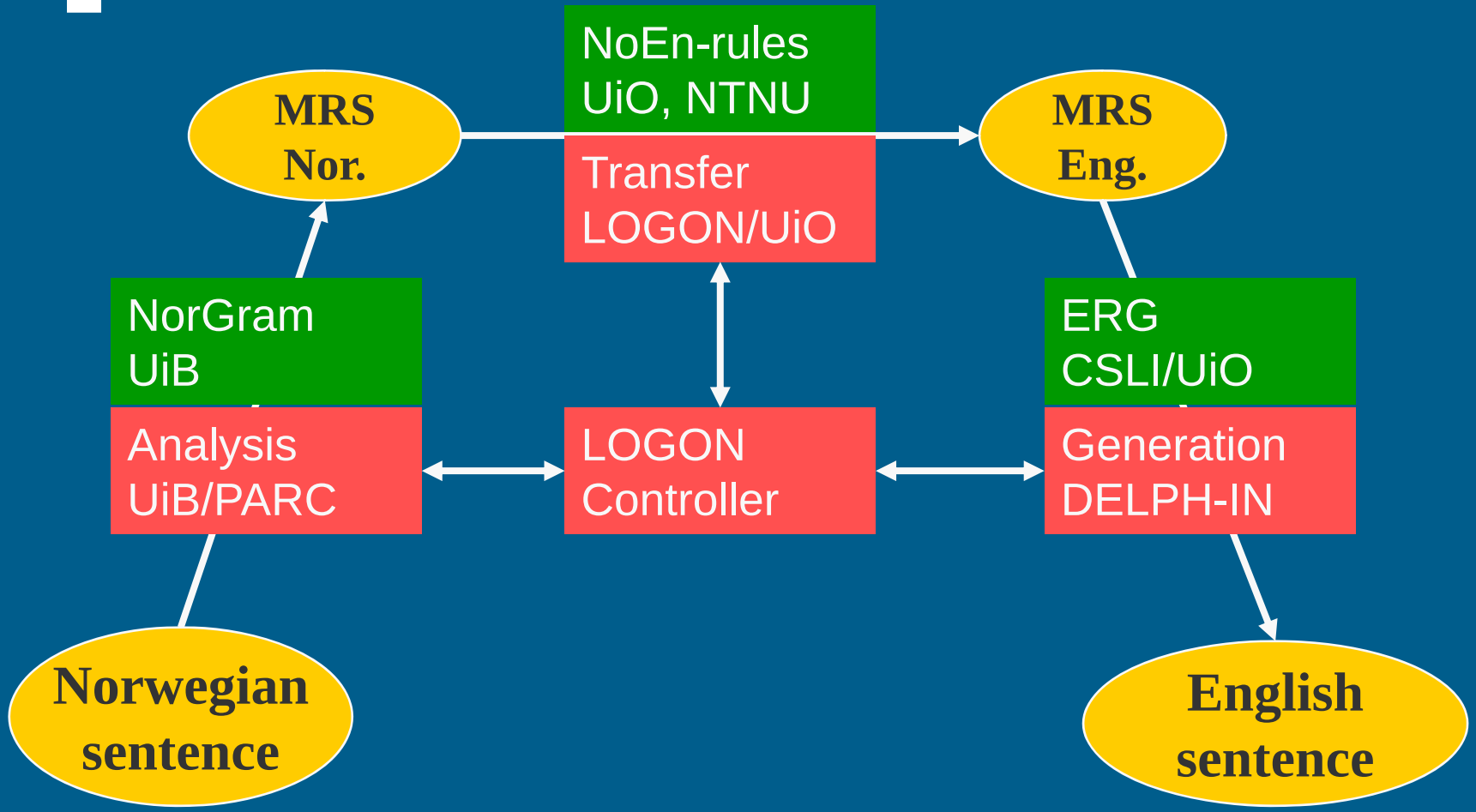
# MT strategies (symbolic)



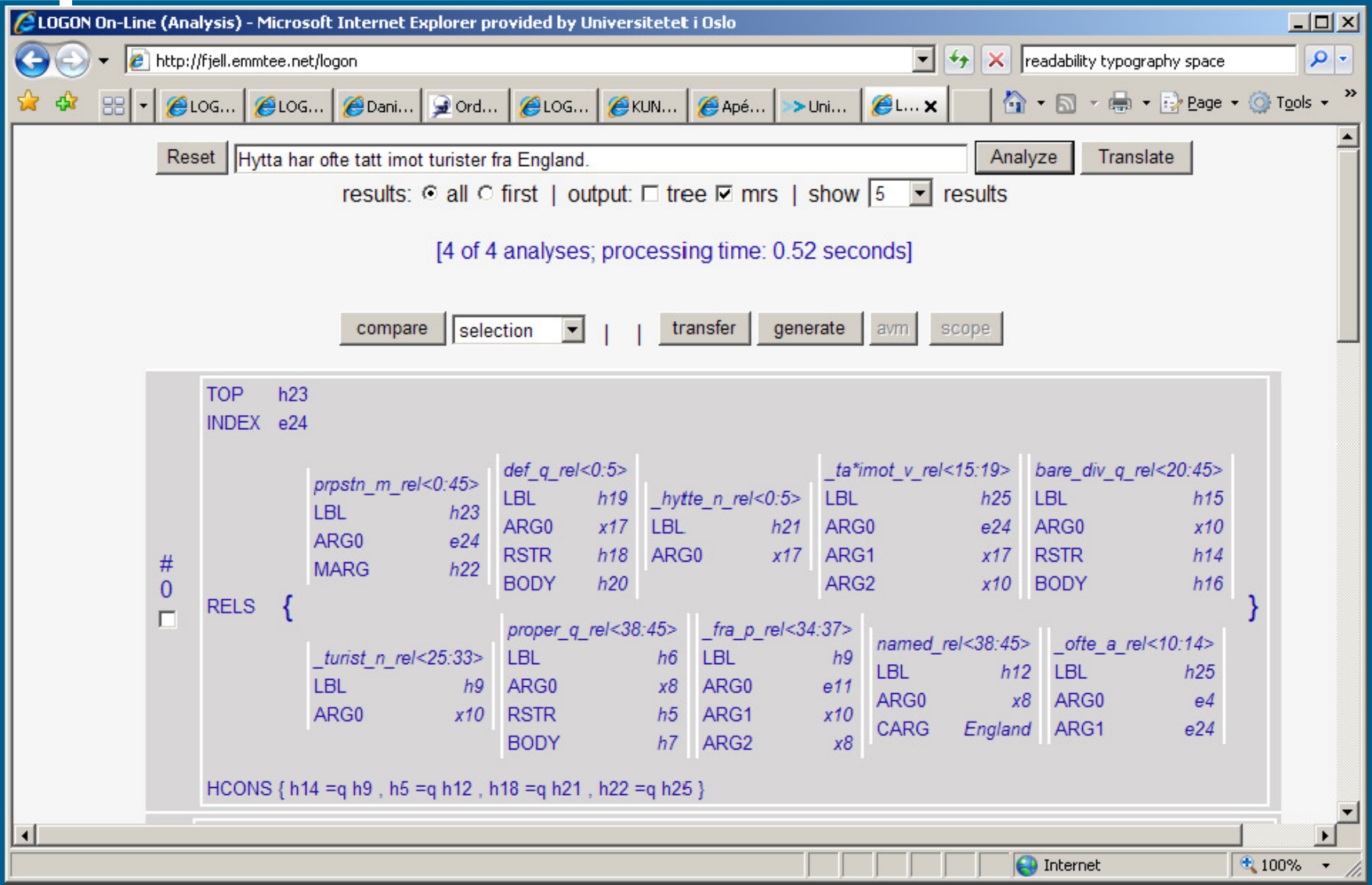
# Basis: Transferbasert oversettelse



# Transfer


NoEn-rules
UiO, NTNU
MRS
Nor.
MRS
Eng.
Transfer
LOGON/UiO
NorGram
UiB
ERG
CSLI/UiO
Analysis
UiB/PARC
LOGON
Controller
Generation
DELPH-IN
Norwegian
sentence
English
sentence

# Minimal Recursion Semantics



# Analysis

- Grammar: NorGram,
  - A multipurpose computational grammar based on LFG
  - Developed at UiB since 1998
  - LOGON has
    - greatly extended grammatical coverage
    - equipped it with an MRS semantics module
    - enhanced efficiency
- Processing
  - The XLE system from PARC
  - Morphological processing developed at UiB on top of earlier projects (tagging, UiB & UiO & NTNU)
  - Compositional analysis of compounds

# Generation

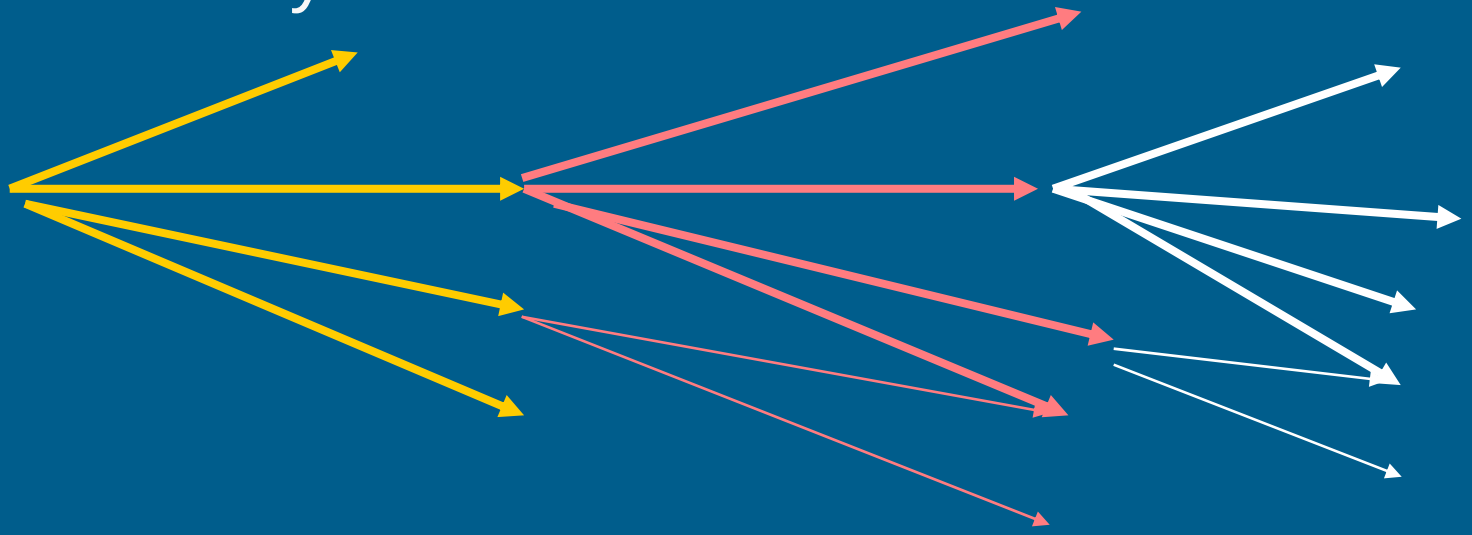
- Grammar
  - The English Resource Grammar (ERG)
  - A multipurpose computational grammar based on HPSG
  - Continuously developed since 1994 (CSLI Stanford)
  - Refined, domain-adapted, and extended by LOGON
  - Open source, used in other ongoing projects
- Processing
  - Adapted technology from DELPH-IN consortium
  - LOGON: forty times faster generation algorithms

# Transfer

- Grammar
  - Hand-coded transfer rules (7000 rules)
  - Semi-automatic acquisition of transfer correspondences
    - for open class words
    - from a dictionary (Kunnskapsforlagets store No-En)
    - (ca 10 000)
- Processing
  - Typed unification-based formalism for rewriting of MRSs
  - Design and implementation from scratch
  - Non-deterministic rewriting of MRS-fragments

# Flertydighet

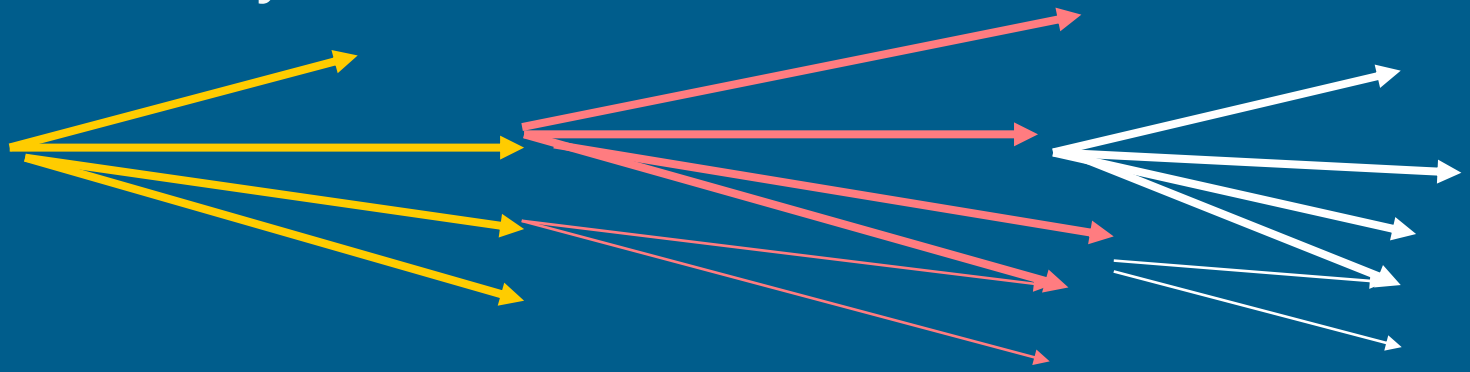
1. Analysis 2. Transfer 3. Generation



- Hvordan velge den rette eller beste på hvert trinn?

# Ambiguity

1. Analysis
2. Transfer
3. Generation



- Each step generates many different hypotheses
- Stochastic models score the alternative outcomes of each component: Parsing, Transfer, Generation
- The per-component scores are calculated together and the final outcomes are ranked.
- Component models are trained on corpora and treebanks.

# Stochastic ranking

|< |Hytta har ofte tatt imot fotturister fra England.| --- 4 x 24 x 180 = 85 [180]
|> |Often, the cabin has received hikers from England.| {0.65} <1.00> (0:0:0).
|> |Often, the cabin has received walkers from England.| {0.61} <1.00> (0:2:0).
|> |The cabin often has received hikers from England.| {0.58} <1.00> (0:0:1).
|> |The cabin has often received hikers from England.| {0.56} <1.00> (0:0:2).
|> |The cabin has received hikers from England often.| {0.55} <1.00> (0:0:4).
|> |The cabin often has received walkers from England.| {0.55} <1.00> (0:2:1).
|> |Often, the cabin has received hikers since England.| {0.54} <1.00> (0:1:0).
|> |Often, the cabin has received ramblers from England.| {0.53} <1.00> (0:3:0).
|> |The cabin has often received walkers from England.| {0.52} <1.00> (0:2:2).
|> |The cabin has received walkers from England often.| {0.52} <1.00> (0:2:4).
|> |The cabin has received hikers often from England.| {0.50} <1.00> (2:0:1).
|> |Often, the cabin has moved from England against hikers.| {0.50} <1.00> (3:0:1).
|> |The cabin often has received hikers since England.| {0.47} <1.00> (0:1:1)

# Demo

- handon.emmtee.net

# Lyspunkt 1

- Vær forsiktig med gaupene!
- Be careful about the lynxes!
- Valg av preposisjon gjøres av de enspråklige grammatikkene som vet hvilke preposisjoner som velges av adjektivene. Preposisjonene er semantisk tomme.

# Lyspunkt 2

- Ask ankom på mandag.
- Ask ankom på ettermiddagen.
- Bruken av underspesifikasjon i generatorinput. Transfer vet at både mandag og ettermiddag er temporale. Derfor blir på overført som et abstrakt predikat temp_loc_sp. Valget av temporal preposisjon i engelsk (on monday, in the afternoon, at five), blir gjort av genereringsgrammatikken.

# Lyspunkt 3-sammensetninger

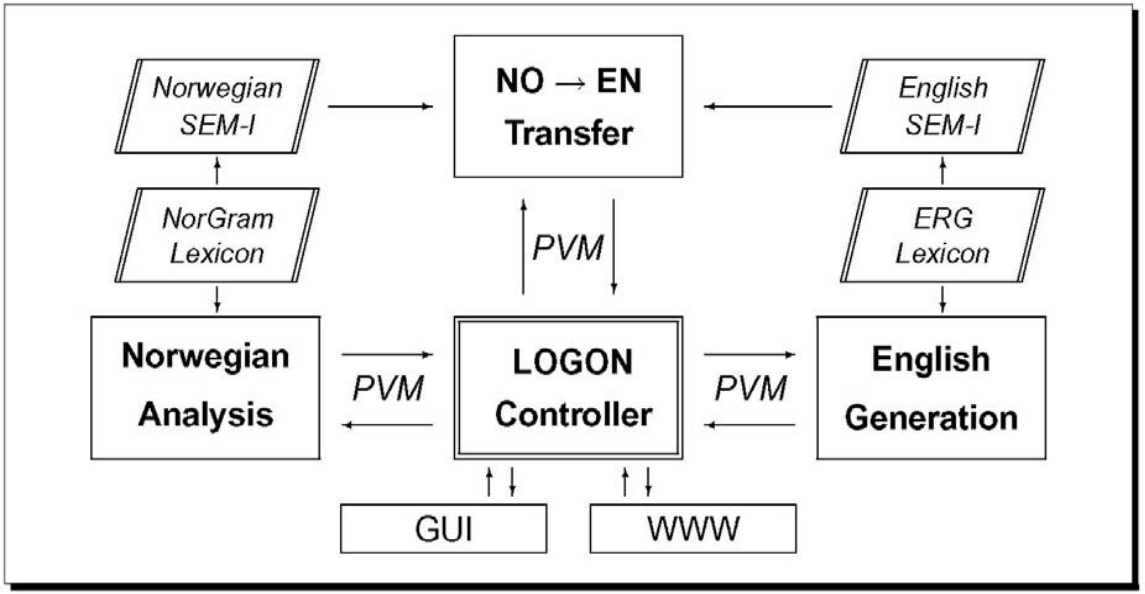
- båndtvang → leash law (N+N as N+N with idiosyncratic translation)
- sandstrand → sandy beach (N+N as Adj+N)
- vareutvalg → selection of merchandise (N+N as N+of+N)
- utgangspunkt → starting point (TL nominalization)
- ved → fire wood (SL lexicalization)
- turgåer → hiker (TL lexicalization)

# Morfologi og sammensetninger

- Analyse i LFG/XLE.
  - Opprinnelig: integrert morfologi etter modell av engelsk.
- I år integrert morfologi for norsk utviklet i andre prosjekt:
  - Taggerprosjekt, Nomen nesco (Enhet for digital dok UiO, Tekstlab UiO, Aksis UiB)
- Muliggjør:
  - Oversetting av produktive sammensetninger
  - Gjenbruk av andre språkteknologiske produkt basert på samme leksikon/morfologi, f.eks. tagger

# 4. SEMI

## Component Organization — Control Flow




UIO — 6-NOV-03 (oe@hf.uio.no)

LOGON Core Architecture (4)

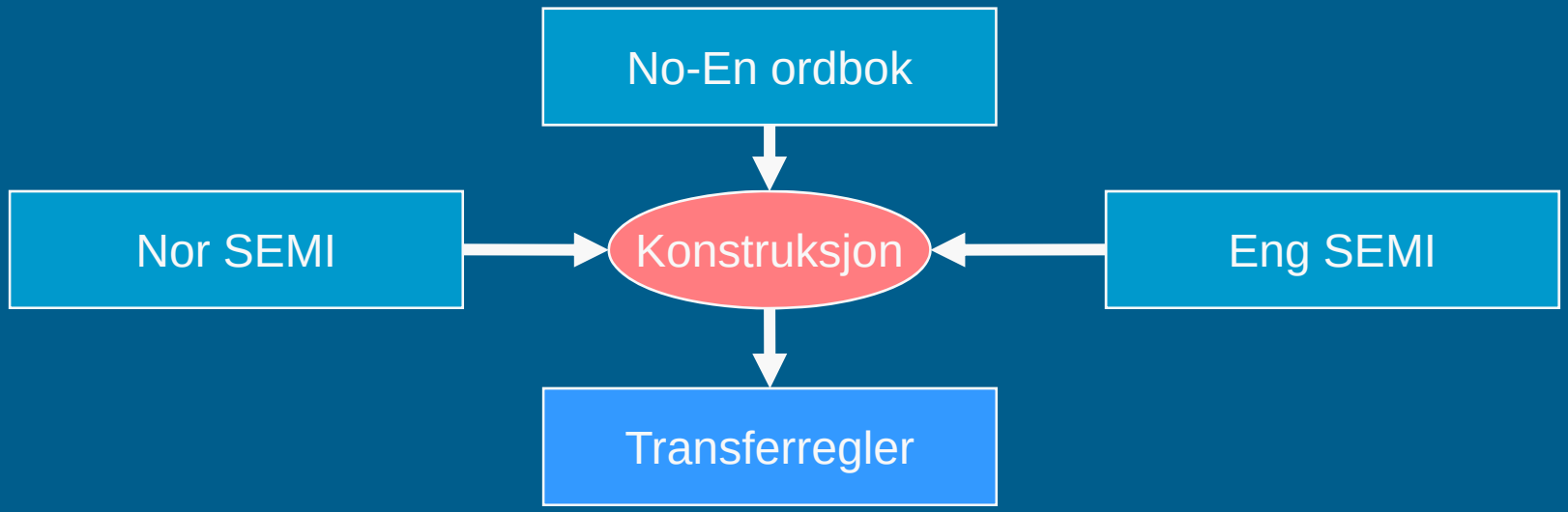
# 4. SEMI

- En abstrakt representasjon av leksikonets semantikk
- Basis for skriving av transferregler
- Basis for automatisk innlæring av transferregler
- Filtrerer ved oversetting

- "_finish_v_off_rel" : ARG0 e, ARG1 u, [ ARG2 i ].
- "_finish_v_up_rel" : ARG0 e, ARG1 u, [ ARG2 i ].
- "_finite_a_1_rel" : ARG0 e, ARG1 u.
- "_finnish_a_1_rel" : ARG0 e, ARG1 u.
- "_finnish_n_1_rel" : ARG0 x.
- "_finn_n_1_rel" : ARG0 x.
- "_fin_n_1_rel" : ARG0 x.
- "_fireplace_n_1_rel" : ARG0 x.
- "_fireside_n_1_rel" : ARG0 x.

”SEM-I Rational MT:Enriching Deep Grammars with a Semantic Interface for Scalable Machine Translation”
Dan Flickinger, Jan Tore Lønning, Helge Dyvik, Stephan Oepen, Francis Bond,
*Proc. MT Summit 2005, Phuket*

# 6. Automatisk konstruksjon av transferrelasjoner




- Så langt for substantiv og adjektiv
- Lite selektivt
- Torbjørn Nordgård

# 7. Genereringsalgoritmer

- Unifikasjonsbaserte grammatikkformalismer:
  - Sentrale i datalingvistikk siste 25 år,
  - Formelt veldefinerte, deklarative formalismer,
  - Velegnet for lingvistiske generaliseringer
  - Velegnet for prosedyrer: parsing og generering
- Men
  - Komputasjonelt kostbare
  - Problem for grammatikker med stor dekningsgrad
- Samtidig
  - En stadig utvikling mot mer effektive algoritmer
  - Kraftigere maskiner

# 7. Genereringsalgoritmer

- John Carroll and Stephan Oepen
- *High-Efficiency Realization for a Wide-Coverage Unification Grammar*, Second International Joint Conference on Natural Language Processing,
- En gjennomgang av nye algoritmer for chart-generering med unifikasjonsbaserte og typebaserte grammatikker
- Definerer en prosedyre for selektiv nummerering av de n-beste resultatene (unngår eksponensiell vekst)
- Empirisk evaluering
- IJCNLP best paper award!

# 8. Rangering av resultater

- |< |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>.

# 8. Statistisk rangering

- Omtrent 30 engelske realiseringer per MRS
- Første forsøk med rangering: BNC og n-gram
- Nye forsøk med tre-bank og max-ent-modeller

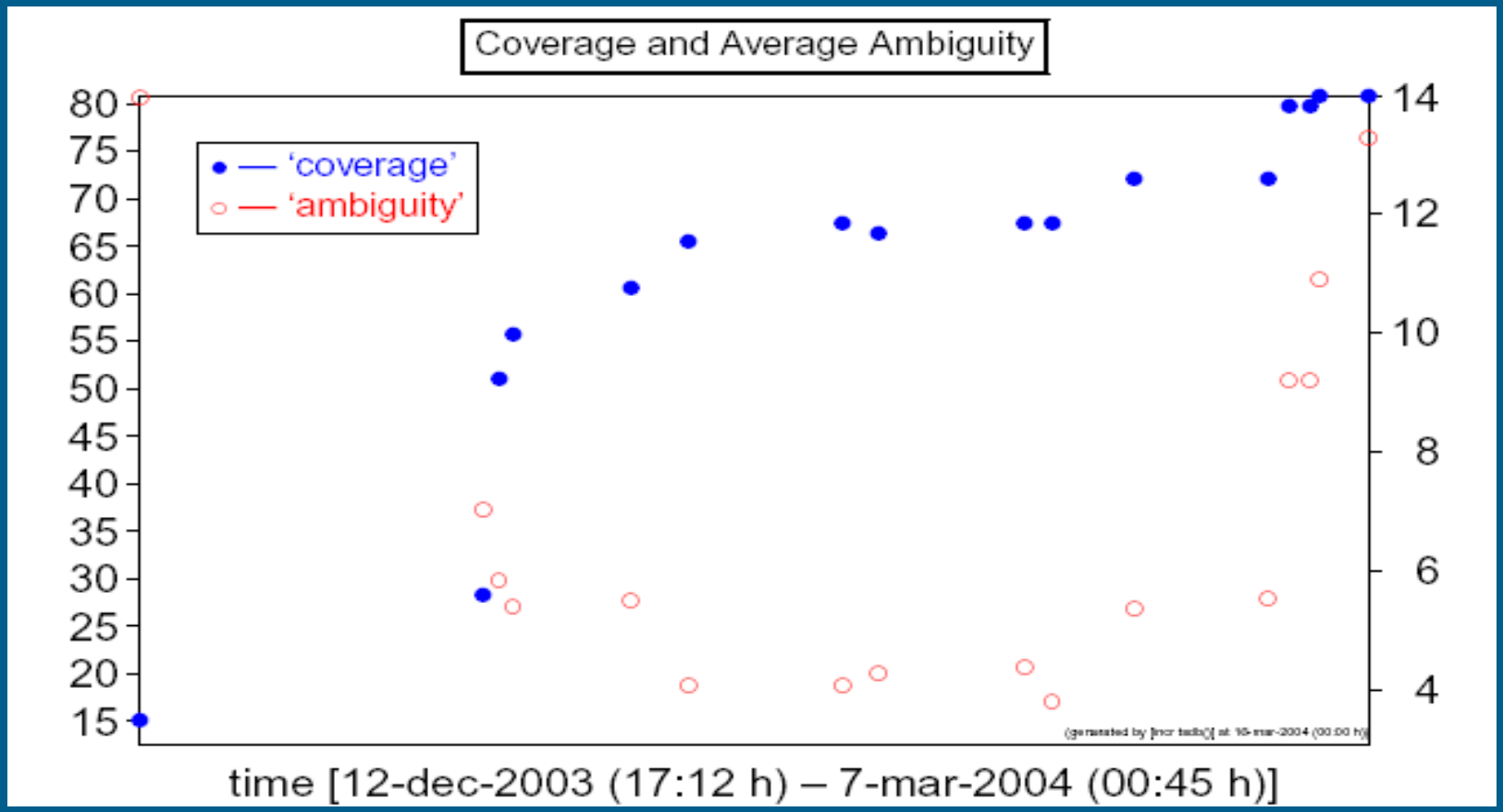
| model configuration | exact | BLEU |
|---|---|---|
| language model of (Velldal et al., 2004) | 48.46 | 0.878 |
| basic model of (Velldal et al., 2004) | 51.36 | 0.897 |
| basic plus lexical type bi-grams | 58.05 | 0.898 |
| basic blus grandparenting | 59.83 | 0.906 |
| basic plus both of the above | 61.58 | 0.903 |
| basic plus language model | 60.71 | 0.915 |
| basic plus all of the above | 65.63 | 0.920 |

Maximum entropy models for realization ranking.
Erik Velldal and Stephan Oepen.
In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, 2005

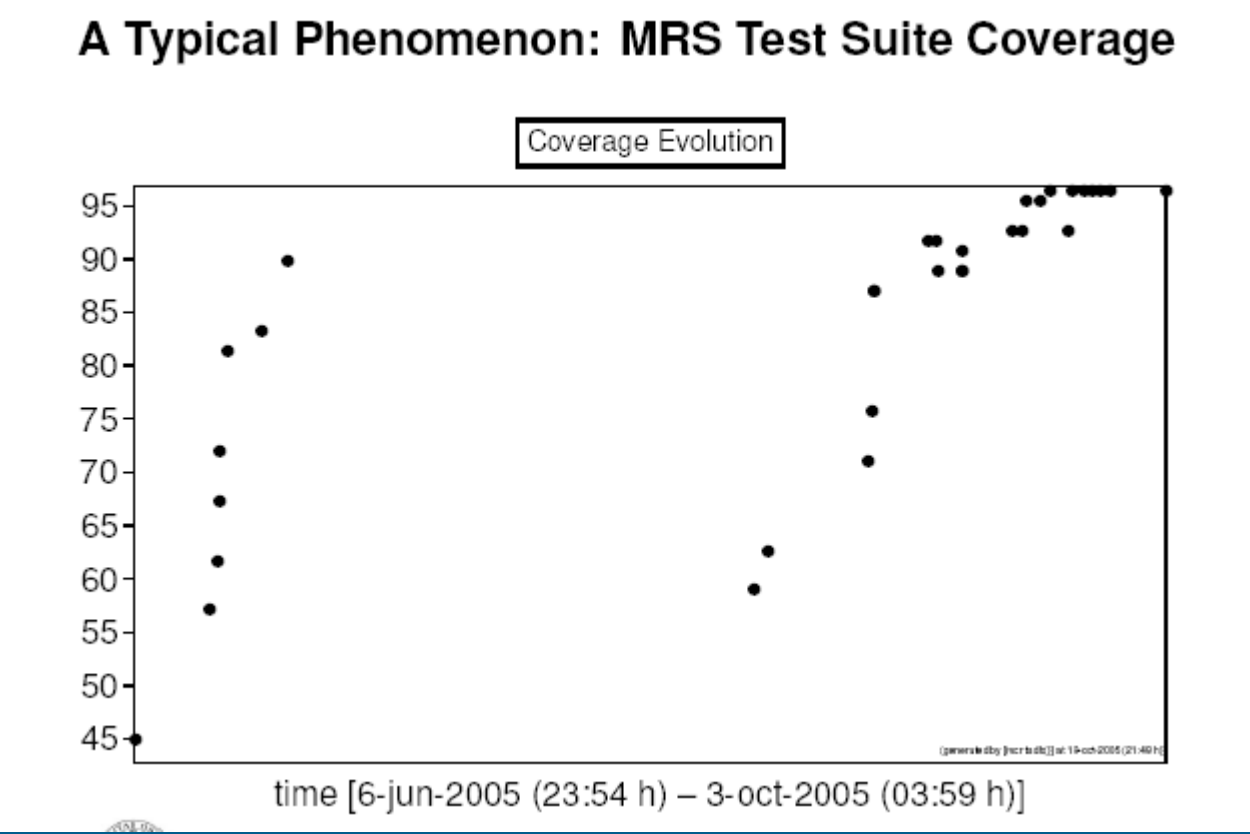
# 9. Profilering og regresjonstesting

- Behov:
  - 3 ulike grammatikker og diverse moduler i stadig utvikling
  - hyppig, rask, systematisk regresjonstesting
- Fra parsing (tidligere):
  - Håndkonstruert testsett av strenger med et vist antall analyser
    - (også 0 for ikke-velformet)
  - For et fast sett setninger (med eller uten fasit) mål grammatikken mot tidligere versjoner (”competence” & ”performance”)
- Utvidet dette til oversettelse:
  - Måler resultater mot tidligere versjoner
  - Bitekster, men ikke direkte mål å gjenskape oversettelsene

# NorGram evolution over time



# Sisyphus






Jan Tore Lønning & Stephan Oepen

# 9. Profilering og regresjonstesting

- Parallell prosessering (grid) tillater mange testkjøringer over en test suite daglig.
- Alle mulige egenskaper ved grammatikker og system kan studeres i et uniformt format.
- DEMO

”Holistic regression testing for high-quality MT. Some methodological and technological reflections”, Stephan Oepen, Helge Dyvik, Dan Flickinger, Jan Tore Lønning, Paul Meurer, and Victoria Rosén.
*Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, Budapest, Hungary, May 2005.

- |< |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>.

# Transfermaskinen

- Reglene er ordnet og gjennomløpes en gang:

1. Start M := start-MRS-et, n:=0
2. n:= n+1, hvis ingen regel(n), avslutt
3. Hvis regel(n) kan anvendes på M, la M:= regel(n) anvendt på M, gjenta
4. Hvis regel(n) er opsjonell lage en forgrening:
    a. En gren går til 3
    b. Den andre til 2
5. Hvis regel(n) ikke kan anvendes på M, gå til 2