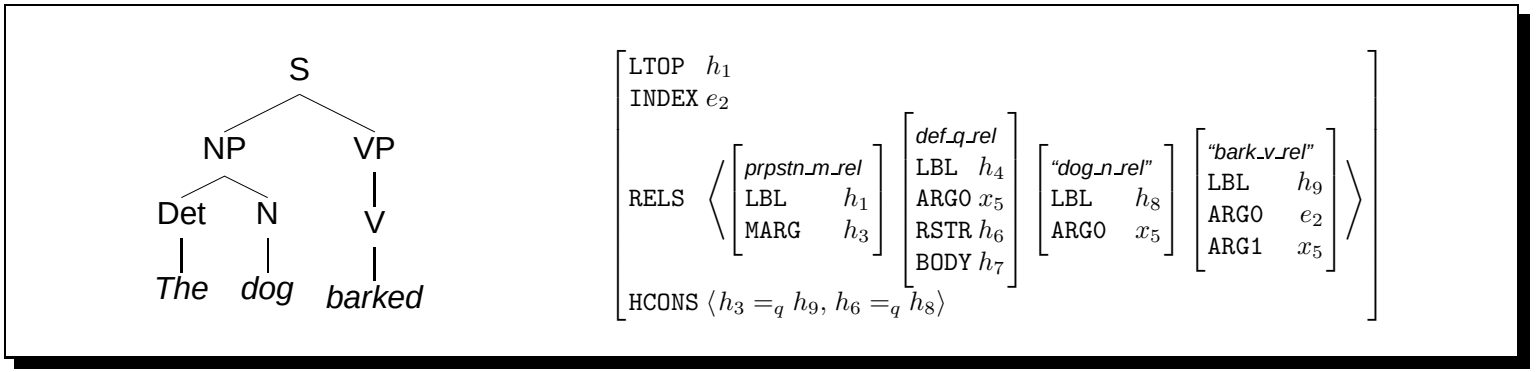



# Computational Linguistics (INF2820 — MT)

$\{$*this(x)* $\wedge$ *fierce(x)* $\wedge$ *dog(x)* $\wedge$ *bark(e,x)* $\}$

**Stephan Oepen**

Universitetet i Oslo & CSLI Stanford

oe@ifi.uio.no

# Linking Semantic Arguments

- Each word or phrase also has an INDEX attribute in SEM
- When heads select a complement or specifier, they constrain its INDEX value – an *entity* variable for nouns, an *event* variable for verbs.
- Each lexeme also specifies a KEY relation (to allow complex semantics)

*transitive-verb-lxm*

$$\begin{bmatrix}
\text{HEAD} & \textit{verb} \\
\text{SPR.FIRST} & \begin{bmatrix}\text{SEM.INDEX}\ \boxed{1}\end{bmatrix} \\
\text{COMPS.FIRST} & \begin{bmatrix}\text{SEM.INDEX}\ \boxed{2}\end{bmatrix} \\
\text{SEM} & \begin{bmatrix}
\text{INDEX}\ \boxed{0}\ \textit{event} \\
\text{KEY}\ \boxed{3} & \begin{bmatrix}\text{PRED}\ \textit{*string*} \\ \text{ARG0}\ \boxed{0} \\ \text{ARG1}\ \boxed{1} \\ \text{ARG2}\ \boxed{2}\end{bmatrix} \\
\text{RELS} & \langle\boxed{3}\rangle
\end{bmatrix}
\end{bmatrix}$$

# Semantics of Phrases

- Every phrase makes the value of its own RELS attribute be the result of appending the RELS lists of its daughter(s) (difference list concatenation);
- Every phrase identifies its semantic INDEX value with the INDEX value of exactly *one* of its daughters (which we will call the *semantic head*);
- As we unify the whole TFS of a complement or specifier with the constraints in the syntactic head, unification takes care of semantic linking.
- Head–modifier structures are analogous: the modifier lexically constrains the INDEX of the head daughter it will modify; the rules unify the whole TFS of the head daughter with the MOD value in the modifier.

# Semantically Void Elements; Construction Semantics

- We have assumed the following all 'mean' the same (truth-conditionally):

*The dog gave that cat those aardvarks.*
*The dog gave those aardvarks to that cat.*
*That cat was given those aardvarks by the dog.*

- Thus, lexical entries for *to*, *by*, *was*, et al. must be semantically empty;

→ differentiation of prepositions; contentful variants needed for modifiers.

- Derivational rules and constructions can introduce additional semantics:

*the chaser of the cat*
*cat food*

→ construction semantics (`CSEM`), treated as another, 'hidden' daughter.

# Language, Linguistics, and Machine Translation?

*With the purchase of Microsoft wanted the supremacy of the search engine giant Google for Internet advertising and Web search break.* [Google Translate: German – English, May 18, 2008]

*Do not want to go so far, is Besstrondrundhø an excellent alternative.* [Google Translate: Norwegian – English, May 18, 2008]

# Language, Linguistics, and Machine Translation?

*With the purchase of Microsoft wanted the supremacy of the search engine giant Google for Internet advertising and Web search break.* [Google Translate: German – English, May 18, 2008]

*Do not want to go so far, is Besstrondrundhø an excellent alternative.* [Google Translate: Norwegian – English, May 18, 2008]

# Language, Linguistics, and Machine Translation?

*With the purchase of Microsoft wanted the supremacy of the search engine giant Google for Internet advertising and Web search break.* [Google Translate: German – English, May 18, 2008]

Mit dem Kauf wollte Microsoft die Vormacht
des Suchmaschinenriesen Google
bei Internet-Werbung und Web-Suche brechen.

# Language, Linguistics, and Machine Translation?

*With the purchase of Microsoft wanted the supremacy of the search engine giant Google for Internet advertising and Web search break.* [Google Translate: German – English, May 18, 2008]

*Do not want to go so far, is Besstrondrundhø an excellent alternative.* [Google Translate: Norwegian – English, May 18, 2008]

**(Formal and) Computational Linguistics**

- Grammar (syntax, semantics, et al.) as tool for language understanding;
- view language as a system of rules, (mostly) shared among speakers;

$\rightarrow$ formal models of grammatical structures for computational processing.

# The Early Days of Machine Translation (1954)

*Russian is Turned into English by a Fast Electronic Translator*

(New York Times, January 8, 1954)

*The switch is assured in advance by attaching the rule sign 21 to the Russian 'gjeneral' in the bilingual glossary which is stored in the machine, and by attaching the rule-sign 110 to the Russian 'major'. The stored instructions, along with the glossary, say whenever you read a rule sign 110 in the glossary, go back and look for a rule-sign 21. If you find 21, print the two words that follow it in reverse order.*

(Journal of Franklin Institute, March 1954)

# The Early Days of Machine Translation (1954)

*Russian is Turned into English by a Fast Electronic Translator*

(New York Times, January 8, 1954)

*The switch is assured in advance by attaching the rule sign 21 to the Russian 'gjeneral' in the bilingual glossary which is stored in the machine, and by attaching the rule-sign 110 to the Russian 'major'. The stored instructions, along with the glossary, say whenever you read a rule sign 110 in the glossary, go back and look for a rule-sign 21. If you find 21, print the two words that follow it in reverse order.*

(Journal of Franklin Institute, March 1954)

- Georgetown Experiment: first public MT demonstration (with IBM);
- minuscule scale: 250 words, six 'syntactic' rules $\rightarrow$ first MT boom.

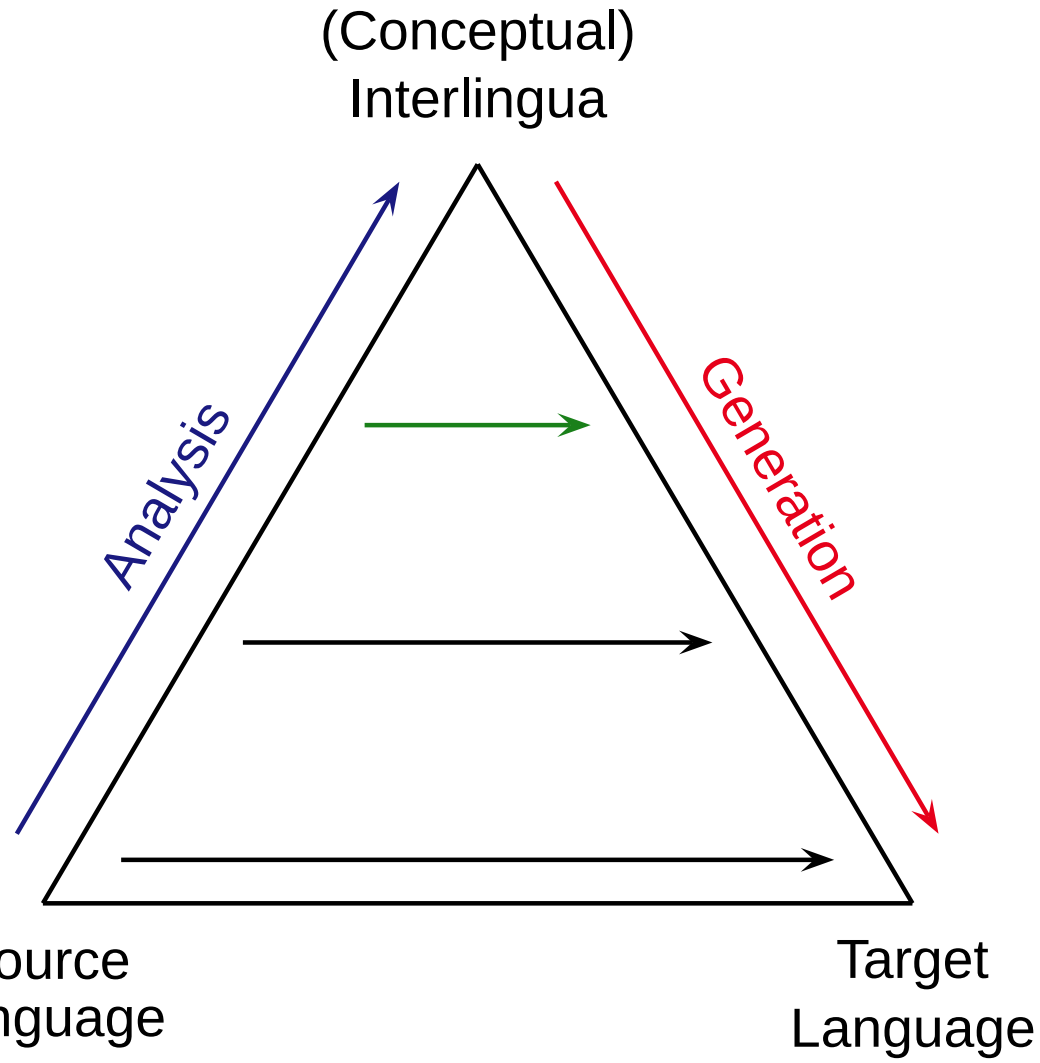
# Dimensions of Machine Translation (Vauquois, 1968)

*Interlingua Translation*

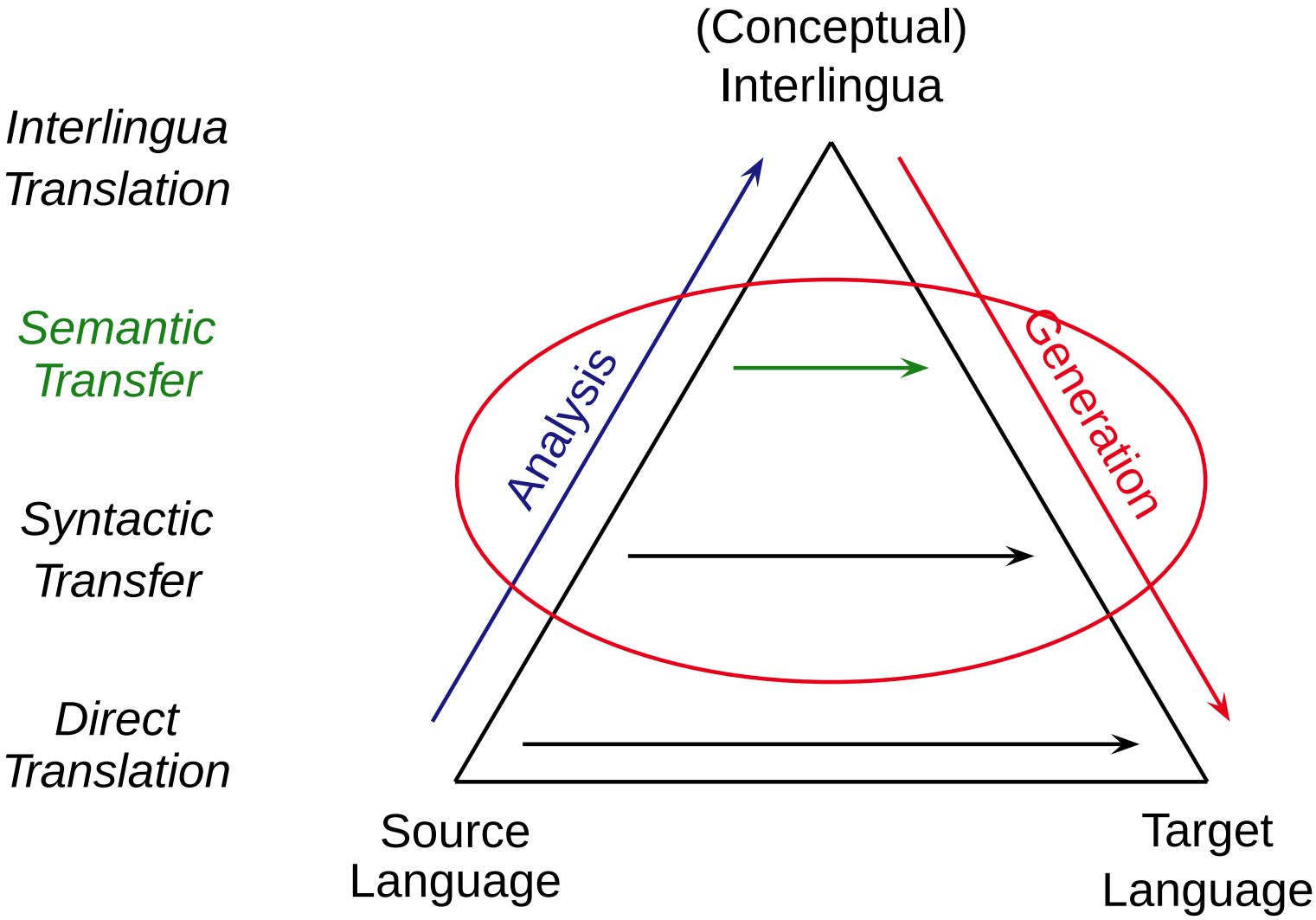
*Semantic Transfer*

*Syntactic Transfer*

*Direct Translation*




# Dimensions of Machine Translation (Vauquois, 1968)



# An Example — What Might Be in a Rule?

**Some Linguistic Observations**

- Noun – noun compounding is a *productive process* in many languages;

  *Such·maschinen·riese* — *search engine giant*

- nearly unrestricted, applies to new words; thus, cannot be enumerated.

# An Example — What Might Be in a Rule?

**Some Linguistic Observations**

- Noun – noun compounding is a *productive process* in many languages;

  *Such·maschinen·riese* — *search engine giant*

- nearly unrestricted, applies to new words; thus, cannot be enumerated.

*river bank* — *kawa no soba*

*river water* — *kawa no mizu*

# An Example — What Might Be in a Rule?

**Some Linguistic Observations**

- Noun – noun compounding is a *productive process* in many languages;

  *Such·maschinen·riese* — *search engine giant*

- nearly unrestricted; applies to new words; thus, cannot be enumerated.

*river bank* — *kawa no soba* — *bord de rivière*

*river water* — *kawa no mizu* — *eau de rivière*

# An Example — What Might Be in a Rule?

**Some Linguistic Observations**

- Noun – noun compounding is a *productive process* in many languages;

  *Such·maschinen·riese* — *search engine giant*

- nearly unrestricted, applies to new words; thus, cannot be enumerated.

*river bank* — *kawa no soba* — *bord de rivière*

*river water* — *kawa no mizu* — *eau de rivière*

$N_1\ N_2 \longrightarrow N_1$ *no* $N_2$

$N_1\ N_2 \longrightarrow N_2$ *de* $N_1$

# Interlingua Translation — Appealing But Impractical

**A Few Cross-Linguistic Examples**

*cousin* — *fetter* | *kusine*

*rice* — *padi* (grain) | *beras* (uncooked) | *nasi* (cooked) | ...

*Jeg fisker gjerne.* — *I like to fish.*

# Interlingua Translation — Appealing But Impractical

**A Few Cross-Linguistic Examples**

*cousin* — *fetter* | *kusine*

*rice* — *padi* (grain) | *beras* (uncooked) | *nasi* (cooked) | ...

*Jeg fisker gjerne.* — *I like to fish.*

**Interlingua vs. Transfer**

- Languages 'carve up' the world differently, lexically and structurally;

→ fully abstract 'conceptual' representation is (put mildly) impractical;

- mono-lingual grammatical knowledge independent of language pair;

→ syntactic or semantic *transfer* accounts for translational divergences.

# A Detour: Advances in Computational Linguistics

**The Grand Challenges**

→ MT research raised foundational questions for language processing:

? **representation** formalizing and encoding of linguistic knowledge;

? **declarativity** separation of linguistic and processing information;

? **reversability** using the same grammar for parsing and generation;

? **computation** (at least) real-time processing of large-scale data;

? **re-usability and standardization** application-independent tools;

? **sustainability** long-term multi-developer and -site collaboration.

# A Detour: Advances in Computational Linguistics

Broad Progress

**The Grand Challenges**

→ MT research raised foundational questions for language processing:

+ **representation** formalizing and encoding of linguistic knowledge;
+ **declarativity** separation of linguistic and processing information;
+ **reversability** using the same grammar for parsing and generation;
+ **computation** (at least) real-time processing of large-scale data;
+ **re-usability and standardization** application-independent tools;
+ **sustainability** long-term multi-developer and -site collaboration.

# A Detour: Advances in Computational Linguistics

**The Grand Challenges**

→ MT research rais

\+ **representation**

\+ **declarativity** se

\+ **reversability** us

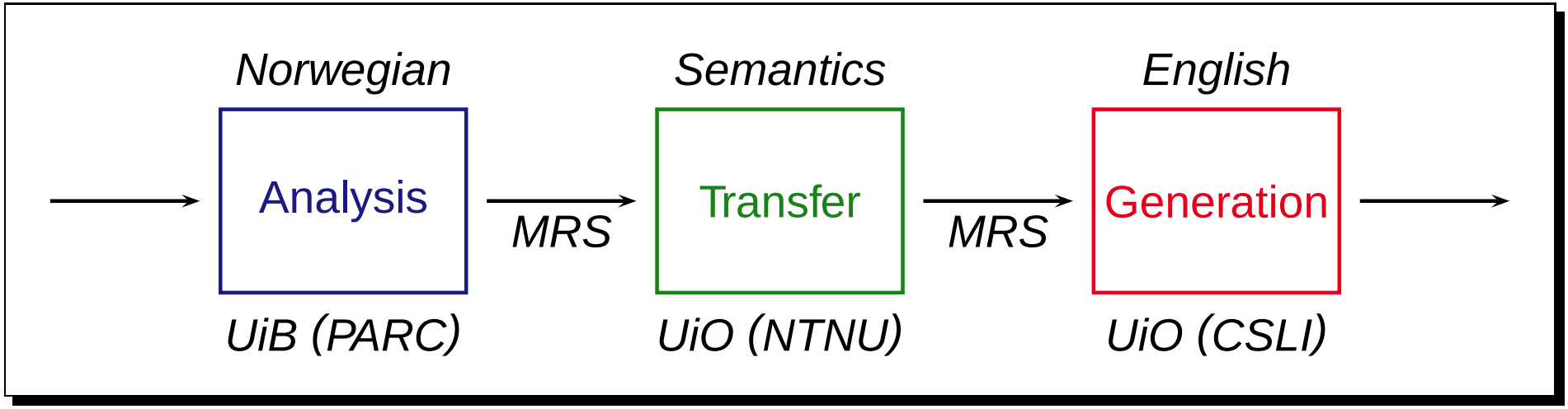
\+ **computation** (a

\+ **re-usability and**

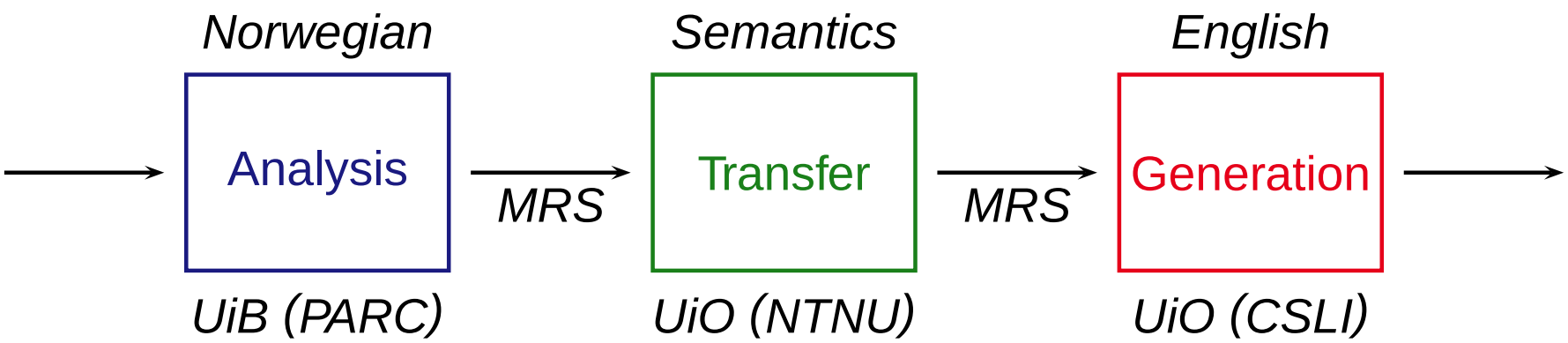
\+ **sustainability** l

Since (around) the early 1990s, federations of computational linguists deploy advanced grammatical formalisms, high-effiency tools, shared (interface) representations, and rigid development and evaluation methodologies— to the analysis of a growing set of languages, applied to many diverse tasks and applications.

# An MT Example — The Norwegian LOGON Project



# An MT Example — The Norwegian LOGON Project




**Some LOGON Highlights**

- Re-usable, mono-lingual precision grammars as linguistic back-bone;
- abstract from language-internal idiosyncrasies by semantic transfer;

→ 'plug & play' of general-purpose resources for flexible MT framework.

# The Real Challenge — Language Ambiguity

```
< Den andre veien mot Bergen er kort. --- 12 x 30 x 25 = 25
> The other path towards Bergen is short. {0.58} (1:1:0).
> The other road towards Bergen is short. {0.56} (1:0:0).
> The second road towards Bergen is short. {0.55} (2:0:0).
> That other path towards Bergen is a card. {0.54} (0:1:0).
> That other road towards Bergen is a card. {0.54} (0:0:0).
> The second path towards Bergen is short. {0.51} (2:1:0).
> The other road against Bergen is short. {0.48} (1:2:0).
> The second road against Bergen is short. {0.48} (2:2:0).
  ...
> Short is the other street towards Bergen. {0.33} (1:4:0).
> Short is the second street towards Bergen. {0.33} (2:4:0).
  ...
```

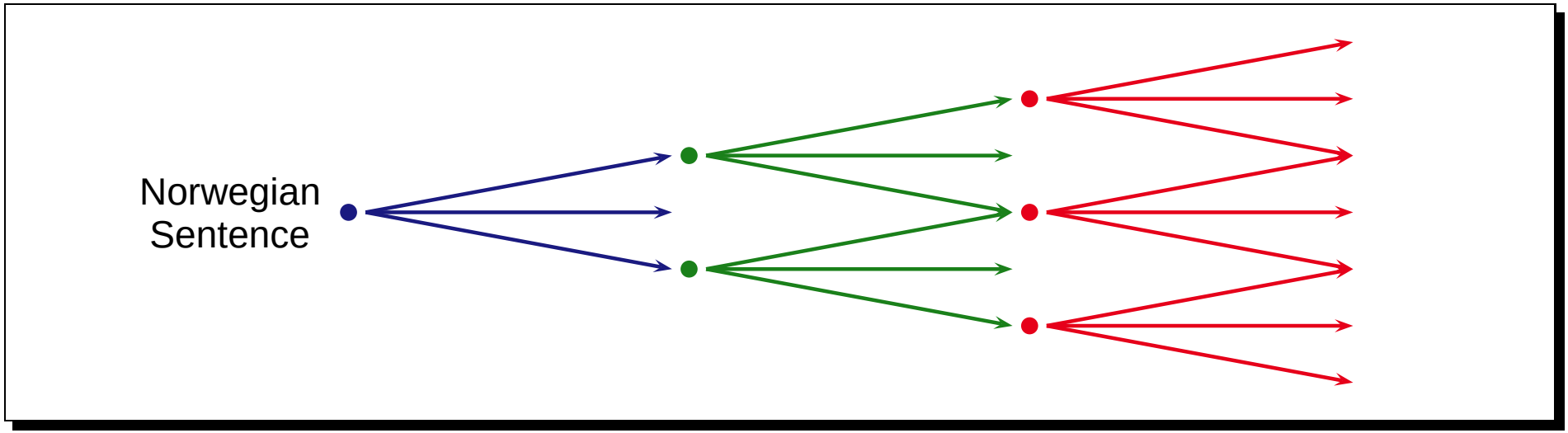
# The Real Challenge — Language Ambiguity

```
< Den andre veien mot Bergen er kort. --- 12 x 30 x 25 = 25
> The other path towards Bergen is short. {0.58} (1:1:0).
> The other road towards Bergen is short. {0.56} (1:0:0).
> The second road towards Bergen is short. {0.55} (2:0:0).
> That other path towards Bergen is a card. {0.54} (0:1:0).
> That other road towards Bergen is a card. {0.54} (0:0:0).
> The second path towards Bergen is short. {0.51} (2:1:0).
> The other road against Bergen is short. {0.48} (1:2:0).
> Th                                                      )).
  ...
> Sh                                                     0).
> Sh                                                    4:0).
  ...
```

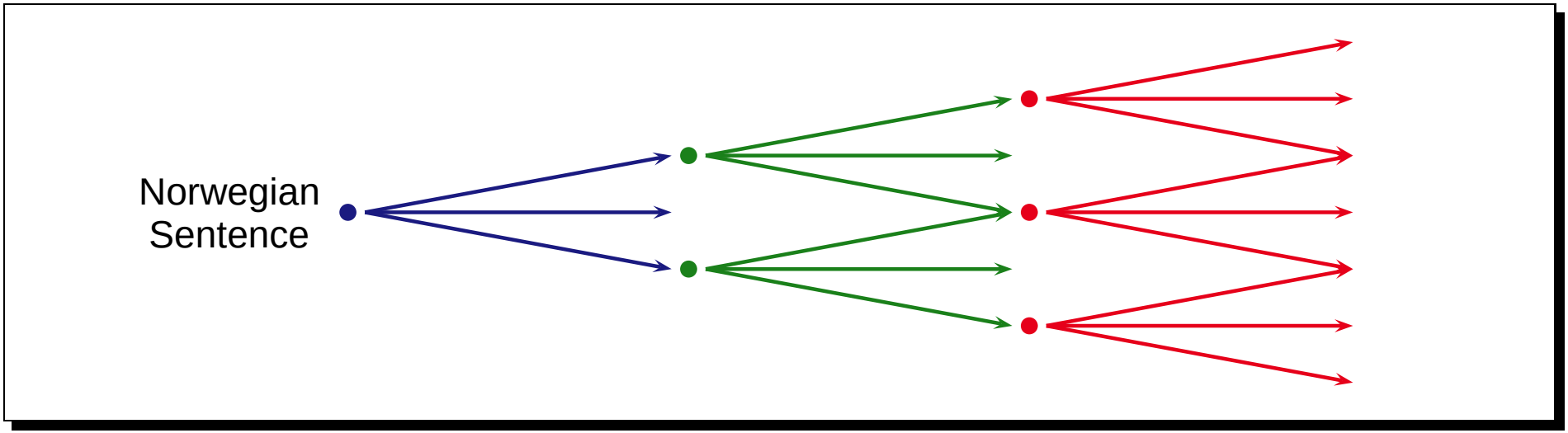
## Scraped Off the Internet

```
the road to the other bergen is short .
Den other roads against Boron Gene are short.
Other one autobahn against Mountains am abrupt.
```

# Ambiguity Management: Stochastic Processes



# Ambiguity Management: Stochastic Processes




**Combining Rule-Based and Statistical Elements**

- Linguistic back-bone grammatically 'circumscribes' the search space;
- advanced statistical models help navigate: rank candidate translations;

$\rightarrow$ hybrid Machine Translation: aim to combine strengths from both worlds.

# Some Sample Translations (And Errors)

1 *Velkommen til Jotunheimen!*

Welcome to Jotunheimen.

1037 *På vestbredden lå det der tre setre nesten ved siden av hverandre.*

On the west bank, 3 mountain pastures lay there almost beside each other.

1048 *Vil du ikke gå så langt, er Besstrondrundhø et utmerket alternativ.*

If you don't want to go so far, Besstrondrundhø is an excellent alternative.

1376 *Den toppen er et fint turmål om du bor på Bessheim eller Gjendesheim.*

That summit, a nice trip tongue is if you stay at Bessheim or Gjendesheim.

# Some Sample Translations (And Errors)

1 *Velkommen til Jotunheimen!*

Welcome to Jotunheimen.

1037 *På vestbredden lå det der tre setre nesten ved siden av hverandre.*

On the west bank, 3 mountain pastures lay there almost beside each other.

1048 *Vil du ikke gå så langt, er Besstrondrundhø et utmerket alternativ.*

If you don't want to go so far, Besstrondrundhø is an excellent alternative.

1376 *Den* *r Gjen-*
*desh*

That r Gjen-
desh

**Google Translate**

Do not want to go so far,
is Besstrondrundhø an excellent alternative.

# (Minimal Recursion) Semantics — By Example

*The new mountain cabin opens on Sunday.*

$\langle\, h_1,$
$\{\, h_1$:proposition_m($h_2$), $h_3$:_open_v_inchoative($e_1$, $x_1$),
$h_4$:_cabin_n($x_1$), $h_4$:_new_a($x_1$), $h_4$:unspec($x_1$, $x_2$), $h_5$:_mountain_n($x_2$),
$h_3$:temp_loc($e_1$, $x_3$), $h_6$:dofw_n($x_3$, SUNDAY), ... $\}$,
$\{\, h_2 =_q h_3 \,\}\,\rangle$

# (Minimal Recursion) Semantics — By Example

*The new mountain cabin opens on Sunday.*

$\langle\, h_1,$
$\{\, h_1$:proposition_m($h_2$), $h_3$:_open_v_inchoative($e_1$, $x_1$),
$h_4$:_cabin_n($x_1$), $h_4$:_new_a($x_1$), $h_4$:unspec($x_1$, $x_2$), $h_5$:_mountain_n($x_2$),
$h_3$:temp_loc($e_1$, $x_3$), $h_6$:dofw_n($x_3$, SUNDAY), ...$\}$,
$\{\, h_2 =_q h_3 \,\}\,\rangle$

## Some Reflections

- logical-form representation of 'who did what to whom' $\rightarrow$ normalization;
- *unspec(...)* relation introduced by compounding (e.g. 'no' in Japanese);
- some interlingua elements, but most predicates require lexical transfer;
- large-scale, MRS-enabled grammars available for several languages.

# (Minimal Recursion) Semantics — By Example

*The new mountain cabin opens on Sunday.*

$\langle\, h_1,$
$\{\, h_1$:proposition_m($h_2$), $h_3$:_open_v_inchoative($e_1$, $x_1$),
$h_4$:_cabin_n($x_1$), $h_4$:_new_a($x_1$), $h_4$:unspec($x_1$, $x_2$), $h_5$:_mountain_n($x_2$),
$h_3$:temp_loc($e_1$, $x_3$), $h_6$:dofw_n($x_3$, SUNDAY), ... $\},$
$\{\, h_2 =_q h_3 \,\} \,\rangle$

## Some Reflections

- logical-form representation of 'who did what to whom' → normalization;
- *unspec(...)* relation introduced by compounding (e.g. 'no' in Japanese);
- some inte cal transfer;
- large-sca nguages.

## Transfer Rule Example (Simplified)

{ _mountain_n } ⟷ { _fjell_n }

# One Grammar for Analysis and Generation

**The Linguistic Knowledge**

- LinGO English Resource Grammar (Dan Flickinger et al., since 1993);
- general-purpose HPSG; domain-specific lexica (some 32,000 lexemes);
- LOGON vocabulary addition and fine-tuning → ∼95 percent coverage;
- manual inspection and treebanking → up to ten percent 'false' coverage;

→ exact same resource used simultaneously in other (non-MT) projects.

**An Open-Source Repository (`http://www.delph-in.net/`)**

- Harmonize theory, formalism, and tools: exchange ling- and software;
- world-wide initiative, now twelve languages under active development.

# LOGON 'Current' State of Play — Facts and Figures

- August 2003 – January 2007, six active developers, ~170 person months;
- limited domain and vocabulary: ~5k sentences edited tourism booklets;

$\rightarrow$ end-to-end: $0.83 \times 0.92 \times 0.85 = 65\,\%$ (71 % vs. 56 % on held-out sets).

# LOGON 'Current' State of Play — Facts and Figures

Partial coverage system potentially useful tool for translators.

- A nths;
- limited domain and vocabulary: ~5k sentences edited tourism booklets;

$\rightarrow$ end-to-end: $0.83 \times 0.92 \times 0.85 = 65\,\%$ (71 % vs. 56 % on held-out sets).

# LOGON 'Current' State of Play — Facts and Figures

- August 2003 – January 2007, six active developers, ~170 person months;
- limited domain and vocabulary: ~5k sentences edited tourism booklets;

→ end-to-end: $0.83 \times 0.92 \times 0.85 = 65\,\%$ (71 % vs. 56 % on held-out sets).

**Some Reflections on Efforts Expended**

- initially: create architecture and interfaces, initiate grammar adaptation;

+ cross-linguistic harmonization: semantic theory for various phenomena;

+ transfer grammar: manual rule writing and semi-automated acquisition;

→ 7627 hand-built transfer rules, 9222 from bi-lingual dictionary → 92.4 %;

+ annotate training data for domain-adapted statistical rankers at all levels.

# LOGON 'Current' State of Play — Facts and Figures

- August 2003 – January 2007, six active developers, ~170 person months;
- limited domain and vocabulary: ~5k sentences edited tourism booklets;

→ end-to-end: $0.83 \times 0.92 \times 0.85 = 65\,\%$ (71 % vs. 56 % on held-out sets).

**Some Reflections**

- initially: create architecture and interfaces, initiate grammar adaptation;

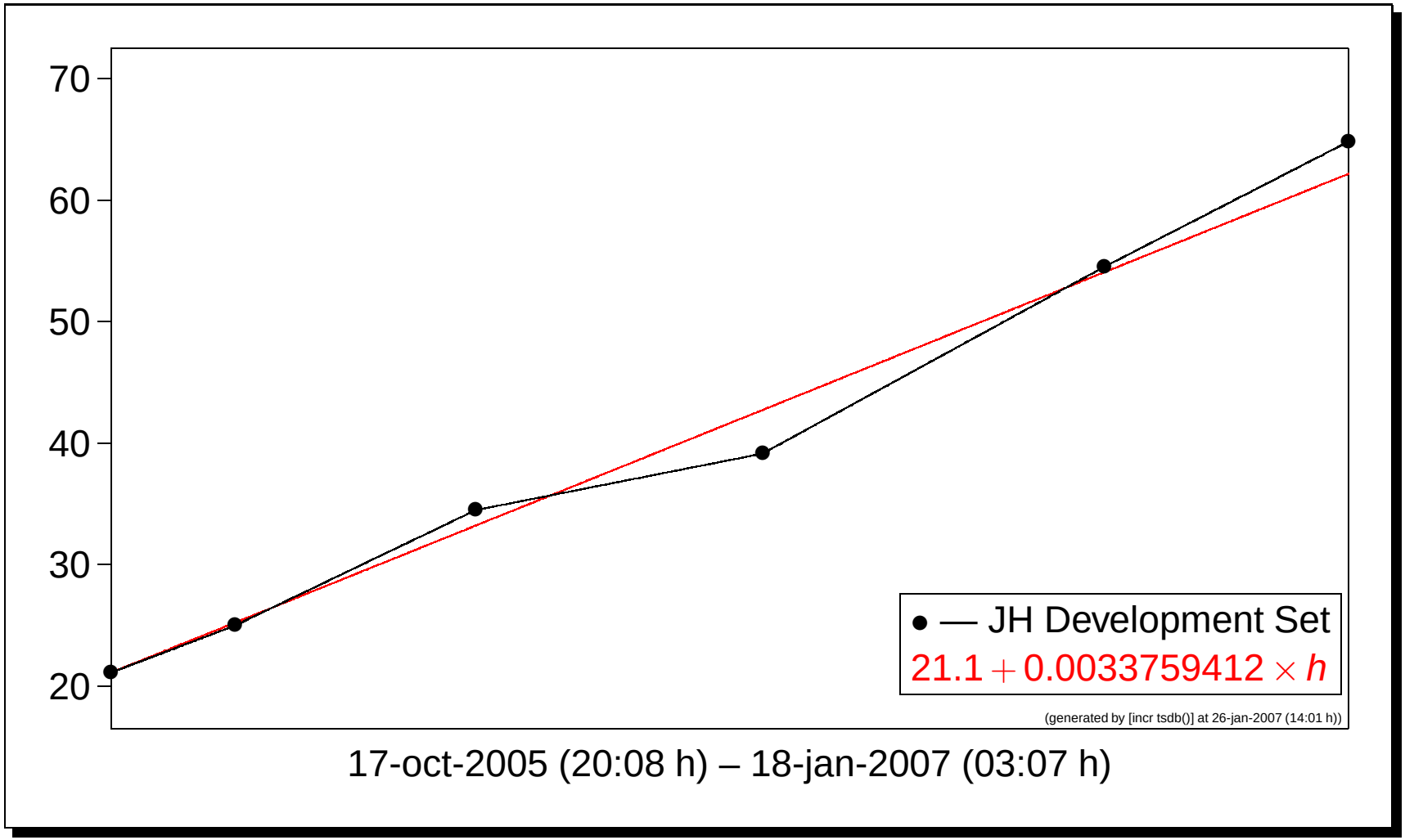
+ cross-linguistic harmonization: semantic theory for various phenomena;

+ transfer grammar: manual rule writing and semi-automated acquisition;

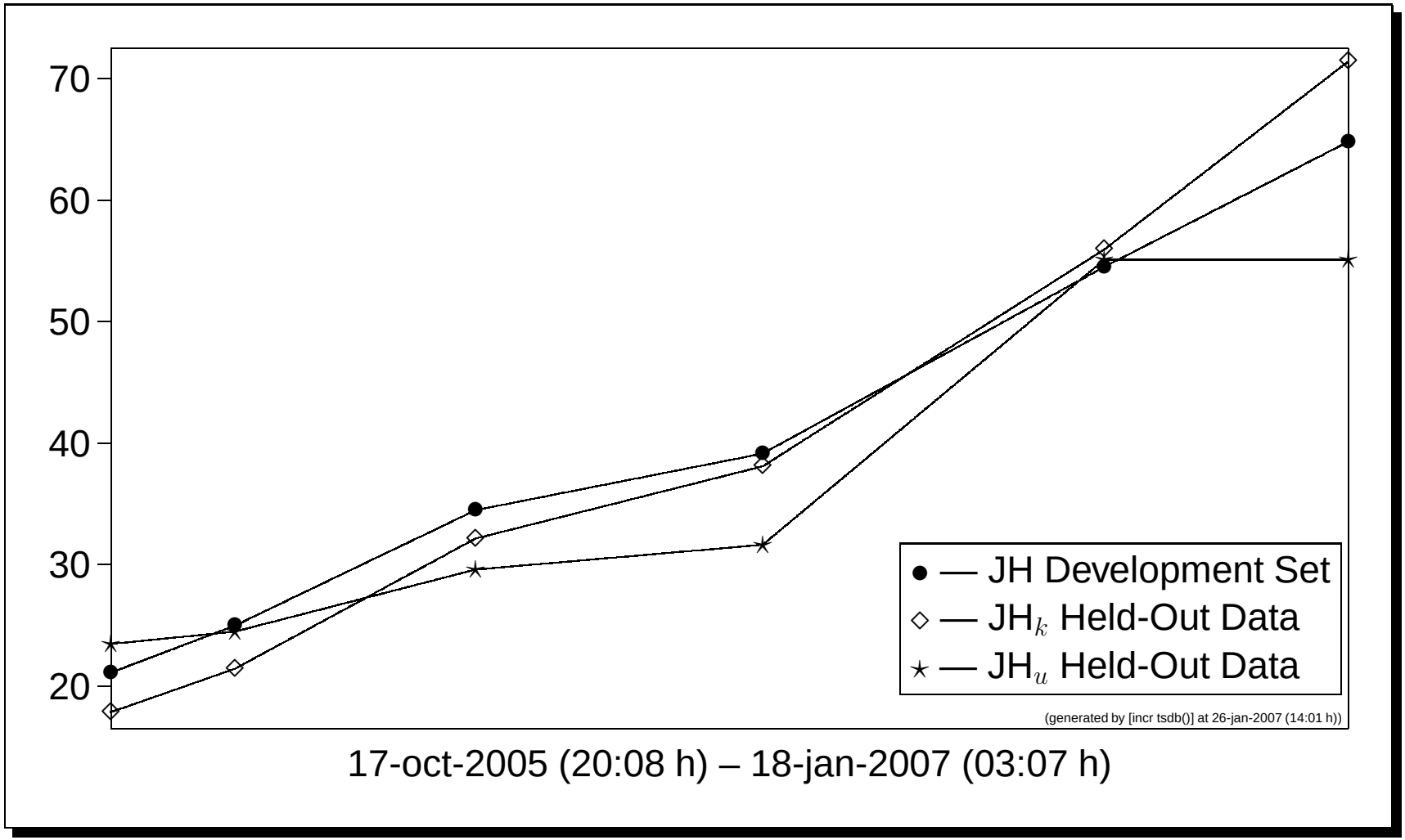
→ 762 (Estimated) Up to Two Thirds of Effort Directly Re-usable: 2.4 %;

+ anr Software, Grammar Extensions, Transfer 'Ontology', et al. levels.

# The Quest for Coverage: End-to-End Throughput




# Comparing Development and Held-Out Data




# LOGON End-to-End Coverage vs. Input Complexity

| ‘gold/logon/jh’ Coverage Profile | | | | | | |
|---|---|---|---|---|---|---|
| **Aggregate** | **total items** ♯ | **positive items** ♯ | **word string** $\phi$ | **distinct analyses** $\phi$ | **total results** ♯ | **overall coverage** % |
| $35 \leq$ *i-length* $< 60$ | 32 | 32 | 39.94 | 0.00 | 0 | 0.0 |
| $30 \leq$ *i-length* $< 35$ | 56 | 56 | 31.48 | 198.00 | 1 | 1.8 |
| $25 \leq$ *i-length* $< 30$ | 137 | 137 | 26.82 | 1180.73 | 11 | 8.0 |
| $20 \leq$ *i-length* $< 25$ | 235 | 235 | 21.87 | 2543.04 | 50 | 21.3 |
| $15 \leq$ *i-length* $< 20$ | 369 | 369 | 16.93 | 667.42 | 173 | 46.9 |
| $10 \leq$ *i-length* $< 15$ | 416 | 416 | 12.11 | 321.63 | 302 | 72.6 |
| $5 \leq$ *i-length* $< 10$ | 454 | 454 | 6.68 | 36.87 | 418 | 92.1 |
| $0 \leq$ *i-length* $< 5$ | 447 | 447 | 2.13 | 3.81 | 436 | 97.5 |
| **Total** | **2146** | **2146** | **12.64** | **266.00** | **1391** | **64.8** |

(generated by [incr tsdb()] at 26-jan-2007 (14:04 h))

# A Human Judgment Study (on Unseen Test Data)

**Experimental Setup**

- 250 held-out sentences, eight judges (English native, Norwegian fluent);
- **fidelity** 'to what degree is the original meaning preserved' (0 to 3);
- **fluency** 'to what degree is the translation natural sounding' (0 to 3);
- custom web interface; judges required to comment on values below 3.

| | OA | SMT | LOGON | | Judge |
|---|---|---|---|---|---|
| **fidelity** | 1.28 | 1.59 | 1.83 | 1.94 | 2.05 |
| **fluency** | 1.27 | 1.31 | 1.62 | 1.69 | 1.79 |

# Preliminary Conclusions — Outlook

**LOGON Results To Date**

- General-purpose NLP resources feasible as rule-based MT back-bone;
- when successful end-to-end, high-quality output(s) typically available;

→ improved stochastic models needed for disambiguation and re-ranking;

→ need to determine scalability, cost of adaptation, re-usability in transfer.

# Preliminary Conclusions — Outlook

**LOGON Results To Date**

- General-purpose NLP resources feasible as rule-based MT back-bone;
- when successful end-to-end, high-quality output(s) typically available;

→ improved stochastic models needed for disambiguation and re-ranking;

→ need to determine scalability, cost of adaptation, re-usability in transfer.

**Confluence of Approaches (MT and CL)**

- Fashion of the year: *hybridization*, balance of linguistics and statistics;
- currently rather low activity level of R&D on 'linguistic' MT, world-wide;

→ rule-based paradigm depends on *sustained*, long-term development.

## Based on Research and Contributions of

Dorothee Beermann, Francis Bond, John Carroll,
Ann Copestake, Helge Dyvik, Liv Ellingsen,
Dan Flickinger, Kristin Hagen, Petter Haugereid,
Lars Hellan, Janne Bondi Johannessen,
Gunn Inger Lyse, Jan Tore Lønning, Paul Meurer,
Torbjørn Nordgård, Lars Nygaard,
Christian Ore, Woodley Packard, Daniel Ridings,
Victoria Rosén, Erik Velldal, and others.