Grammar-Based Processing and Re-Usable Hybrid MT

— An Experience Report in the Wilderness —

Stephan Oepen

Oslo University and CSLI Stanford

oe@ifi.uio.no

(Machine Translation Marathon – May 16, 2008

Grammar-Based Processing for Machine Translation?

on the west - bank it there lay DecimalErsatz setre almost at the side of each other . [Naïve Norwegian – English SMT]

Do not want to go so far, is Besstrondrundhø an excellent alternative. [Google Translate, Norwegian – English, May 16, 2008]

Grammar-Based Processing for Machine Translation?

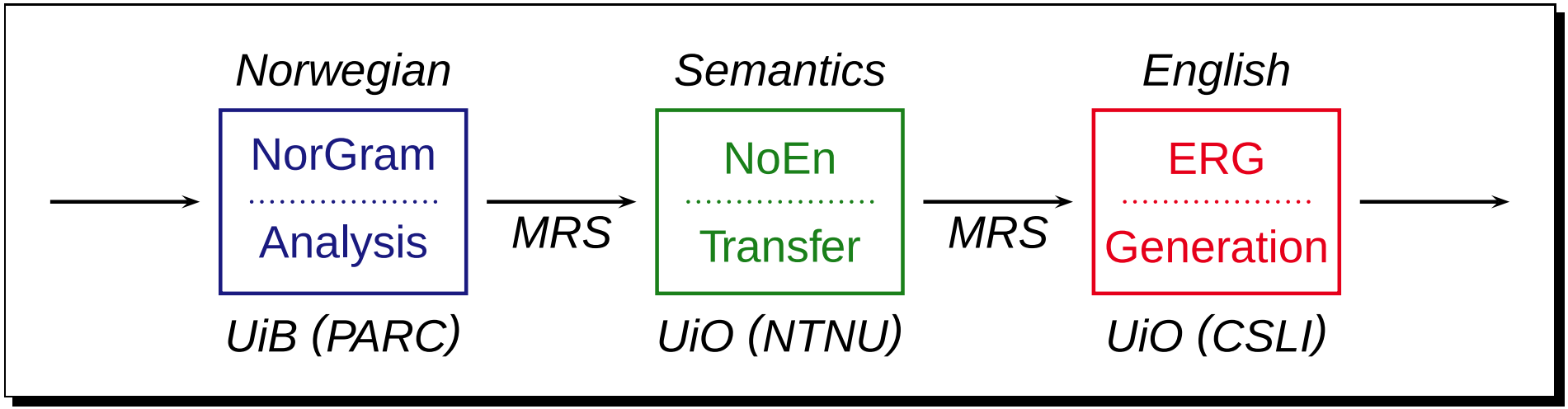
on the west - bank it there lay DecimalErsatz setre almost at the side of each other . [Naïve Norwegian – English SMT]

Do not want to go so far, is Besstrondrundhø an excellent alternative. [Google Translate, Norwegian – English, May 16, 2008]

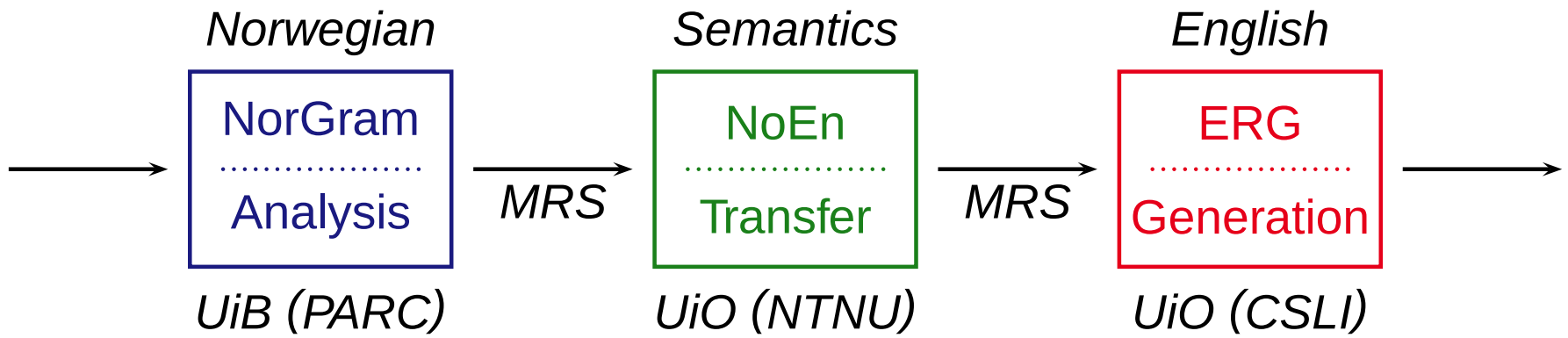
Broad Progress in Computational Linguistics

- Precise, broad-coverage grammars now available (e.g. LFG & HPSG);
- linguistic grammar relates strings to syntactic *and* semantic analyses;
- grammar (pretty much) guarantees wellformedness of system outputs;
- combination with stochastic processes to rank and select hypotheses.

Old-Fashioned MT: The Norwegian LOGON Project



Old-Fashioned MT: The Norwegian LOGON Project

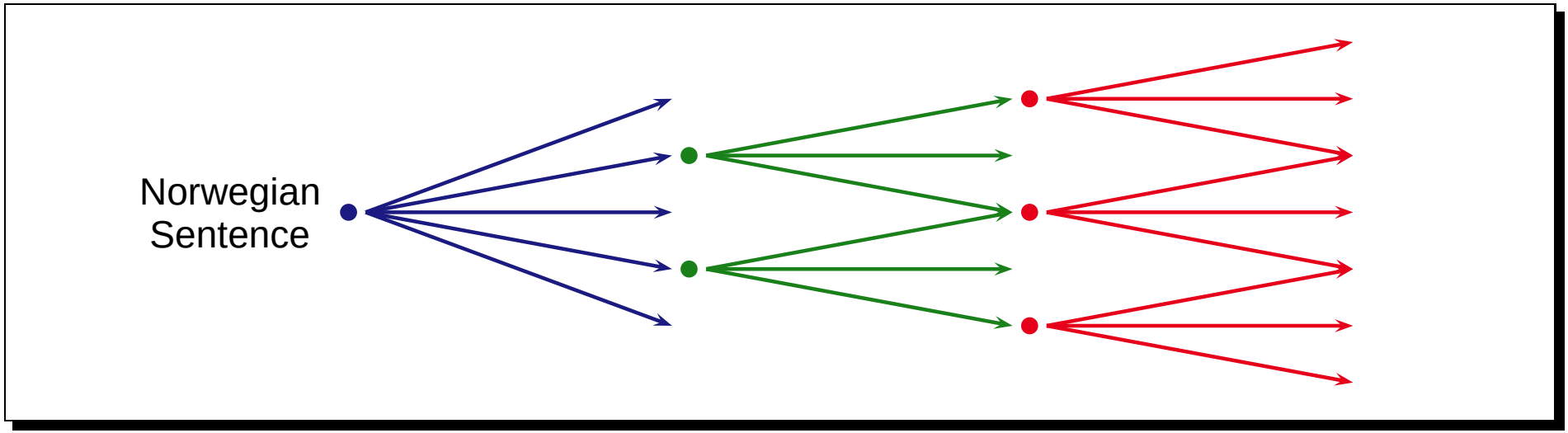


Some LOGON Highlights

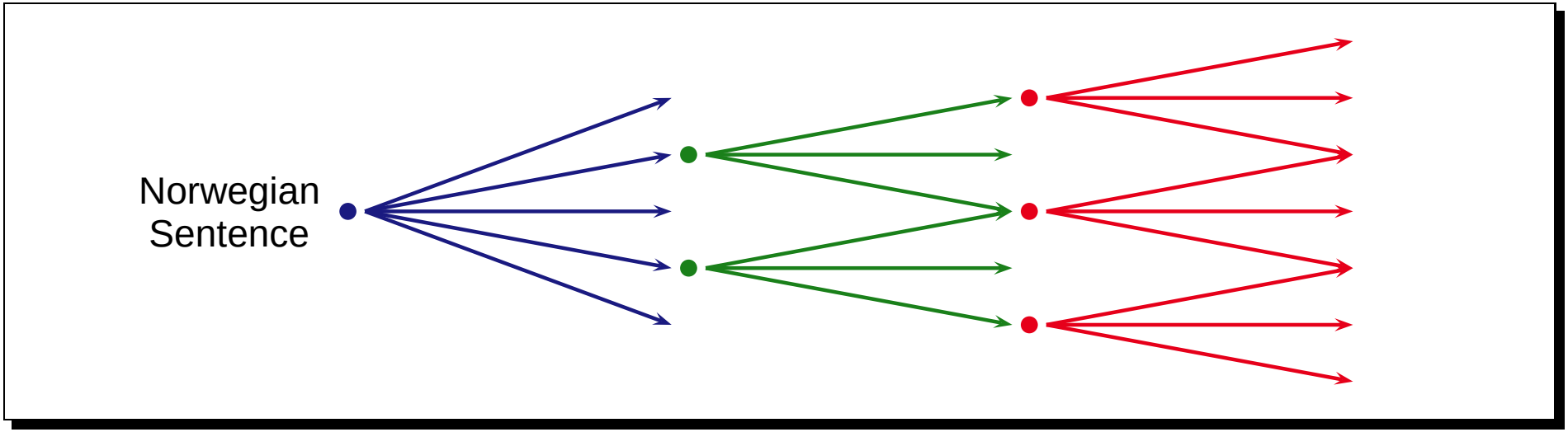
- Re-usable, mono-lingual precision grammars as linguistic back-bone;
- abstract from language-internal idiosyncrasies by semantic transfer;

$\rightarrow$ most likely unique in combination of LFG and HPSG in working system.

Ambiguity Management: Stochastic Processes



Ambiguity Management: Stochastic Processes



Stochastic Elements in LOGON

- At each stage, rank alternate hypotheses according to local probability;
- discriminative re-ranking: normalize and combine scores to re-order;

→ hybrid MT: linguistic back-bone, combined with advanced statistics.

For Example: Various Sources of 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).
  ...
```

For Example: Various Sources of 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).
> Th                                                      .
> Th                                                      )).
  ...
> Sh                                                     0).
> Sh                                                     4:0).
  ...
```

Scraped Off the Internet

the road to the other bergen is short .

The other road to Bergen is short.

Den other roads against Boron Gene are short.

Other one autobahn against Mountains am abrupt.

Project Organization — A Few Facts

Organizational

- ~~All but one university in Norway participate: Oslo, Bergen, Trondheim~~;
- received about 10 crowns from each Norwegian tax payer; 2003 – 2007.

Scope & Domain

- Functional core demonstrator NO – EN (limited domain and vocabulary);
- support to the regions: facilitate the translation of tourism-related texts.

Existing Resources at Project On-Set

- NorGram (ParGram): Norwegian LFG, since 1999 (Dyvik, Rosén; UiB);
- ERG (DELPH-IN): broad-coverage English HPSG (Flickinger; Stanford).

Main Development Corpora

Original Development Corpus ('tur')

- 104 sentences, picked from DNT brochure by NorGram developers;
- three project-internal translations; average length 10.5 : 12.9 words.

Jotunheimen, Preikestolen, Turglede (JHPSTG)

- Lauritzen, Per Roger (2001): *På tur i Jotunheimen.* (four booklets);
- one original, published translation; two commissioned translations;
- augmented with a few smaller texts; two commissioned translations;
- Norwegian: 4062 sentences (44079 words); English: 10927 (134973).

Some LOGON Sample Translations (Version 0.9)

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.

Minimal Recursion Semantics — By Example

The new serviced cabin opens on Sundays.

$\langle h_1,$
$\{ h_1$:proposition_m(h_2), h_3:_open_v_inchoative(e_1, x_1),
h_4:def_q(x_1, h_5, h_6), h_7:_cabin_n(x_1), h_7:_new_a(x_1), h_7:_service_v(e_2, u_1, x_1),
h_3:temp_loc(e_1, x_2), h_8:proper_q(x_2, h_9, h_{10}), h_{11}:dofw_n(x_2, SUNDAY) $\}$,
$\{ h_2 =_q h_3, h_6 =_q h_7, h_{10} =_q h_{11} \} \rangle$

Minimal Recursion Semantics — By Example

The new serviced cabin opens on Sundays.

$\langle\, h_1,$
$\{\, h_1$:proposition_m(h_2), h_3:_open_v_inchoative(e_1, x_1),
h_4:def_q(x_1, h_5, h_6), h_7:_cabin_n(x_1), h_7:_new_a(x_1), h_7:_service_v(e_2, u_1, x_1),
h_3:temp_loc(e_1, x_2), h_8:proper_q(x_2, h_9, h_{10}), h_{11}:dofw_n(x_2, SUNDAY) $\}$,
$\{\, h_2 =_q h_3,\ h_6 =_q h_7,\ h_{10} =_q h_{11} \,\}\, \rangle$

Background (Copestake et al., 1996, 2003, & 2005)

- Family of flat, underspecified semantics (including UDRS, CLLS, et al.);
- non-recursive composition: predicates, bindings, and scope constraints;
- logical-form representation of 'who did what to whom' $\rightarrow$ normalization;
- large, MRS-enabled grammars available for at least seven languages.

Norwegian Analysis — NorGram

The Grammar

- Actively developed at UiB since 1998; Helge Dyvik and Victoria Rosén;
- part of ParGram consortium → Xerox Linguistic Environment (XLE);
- traditionally heretic: some divergence in feature inventory; s-projection;
- large, semi-manual lexicon, though previously not focused on coverage.

Adaptation for LOGON

- Add MRS projection in co-description spirit; not defining grammaticality;
- MRS elements primarily projected off the f-structure, where possible;
- post-XLE processing: accumulation of bags and some scope relations.

The MRS Transfer Formalism

Background

- *Semantic transfer* as successive rewriting of meaning representation;
- transfer rule: replacement of SL MRS fragment with TL correspondence;

\+ complex, non-monotonic transformations; works well with *flat* structures;

\+ good handle on most translational correspondences *and* divergences;

– resource-sensitive process: rule ordering; reversibility not guaranteed.

Realization

- General-purpose, unification-based rewrite system on MRS structures;
- allow non-determinism (ambiguity), but tight control in transfer grammar;

→ sets of rules; manual ordering of rules and sets relative to each other.

The MRS Transfer Formalism

Background

- *Semantic transfer* as successive rewriting of meaning representation;
- transfer rule: replacement of SL MRS fragment with TL correspondence;

+ complex, non-monotonic transformations; works well with *flat* structures;
+ good handle on most translational correspondences *and* divergences;

– res eed.

$$\langle \_, \{ \text{\_tur\_n}(x_1) \}, \{ \} \rangle : \langle \_, \{ h_0\text{:\_gå\_v}(e_0, x_0, x_1) \}, \{ \} \rangle$$
$$\rightarrow \quad \langle \_, \{ h_0\text{:\_take\_v\_1}(e_0, x_0, x_1) \}, \{ \} \rangle$$

- Ge tures;
- allow non-determinism (ambiguity), but tight control in transfer grammar;

→ sets of rules; manual ordering of rules and sets relative to each other.

English Generation — ERG

The Grammar

- 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 per cent coverage;
- manual inspection and treebanking → up to ten percent 'false' coverage;

→ same resource is used centrally in multiple parallel projects (non-MT).

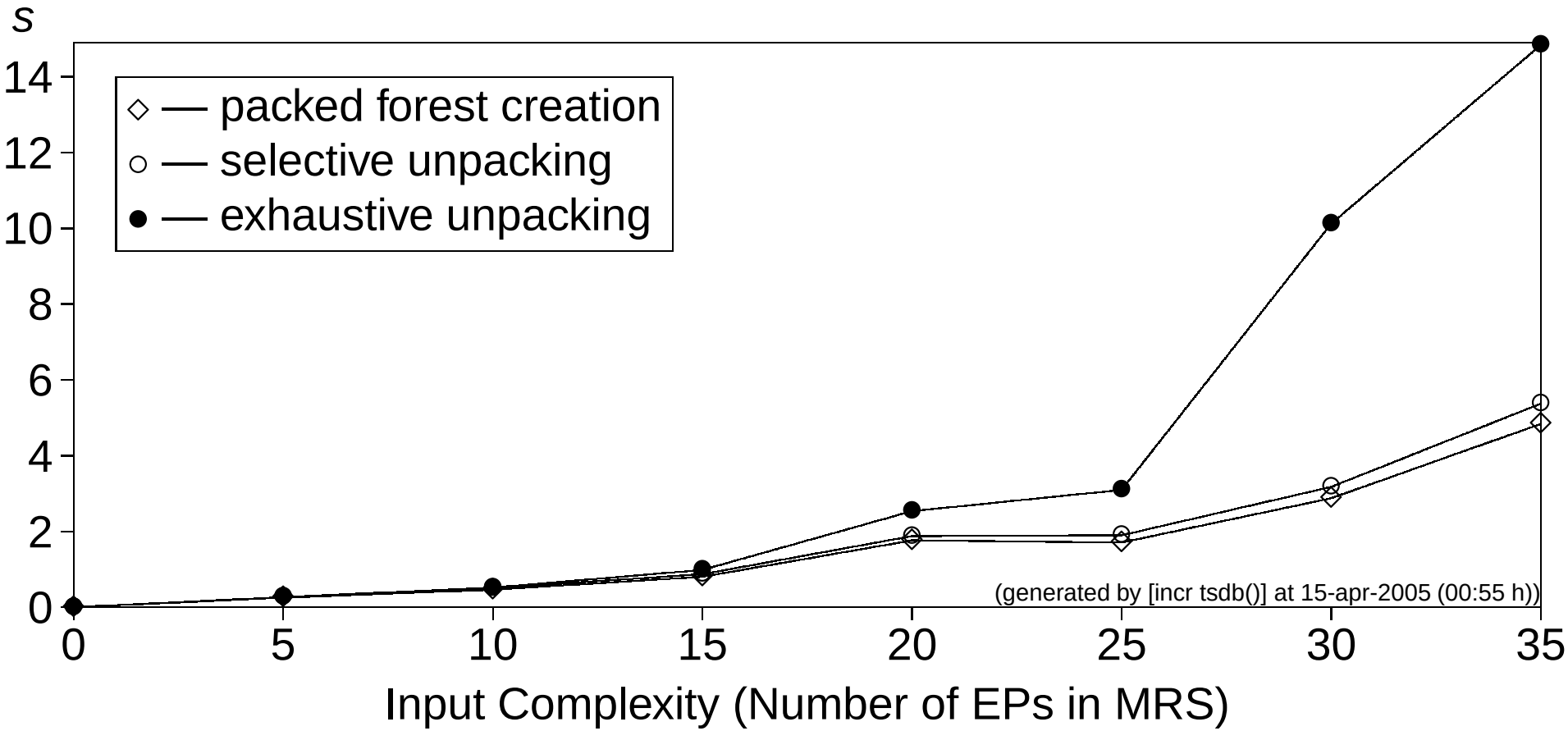
The Generator

- Standard LKB chart generator (Kay 1996) and (Carroll et al. 1999);

\+ subsumption-based ambiguity packing at edge level; adjoin into forest;

\+ MaxEnt realization ranking; selective unpacking of *correct* n-best list.

Polynomial Time (Practical) Generation



→ **Average Time for 15-Word Sentences around One Second**

LOGON 'Current' State of Play — Facts and Figures

- August 2003 – January 2007, six active developers, ~170 person months;

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

Linguistic Resources & Stochastic Models

- NorGram (Dyvik et al.): 8^{+} years; new MRS projection → $67.5 + 15.1\,\%$;

+ adaption of discriminant treebanking to LFG; native XLE parse selection.

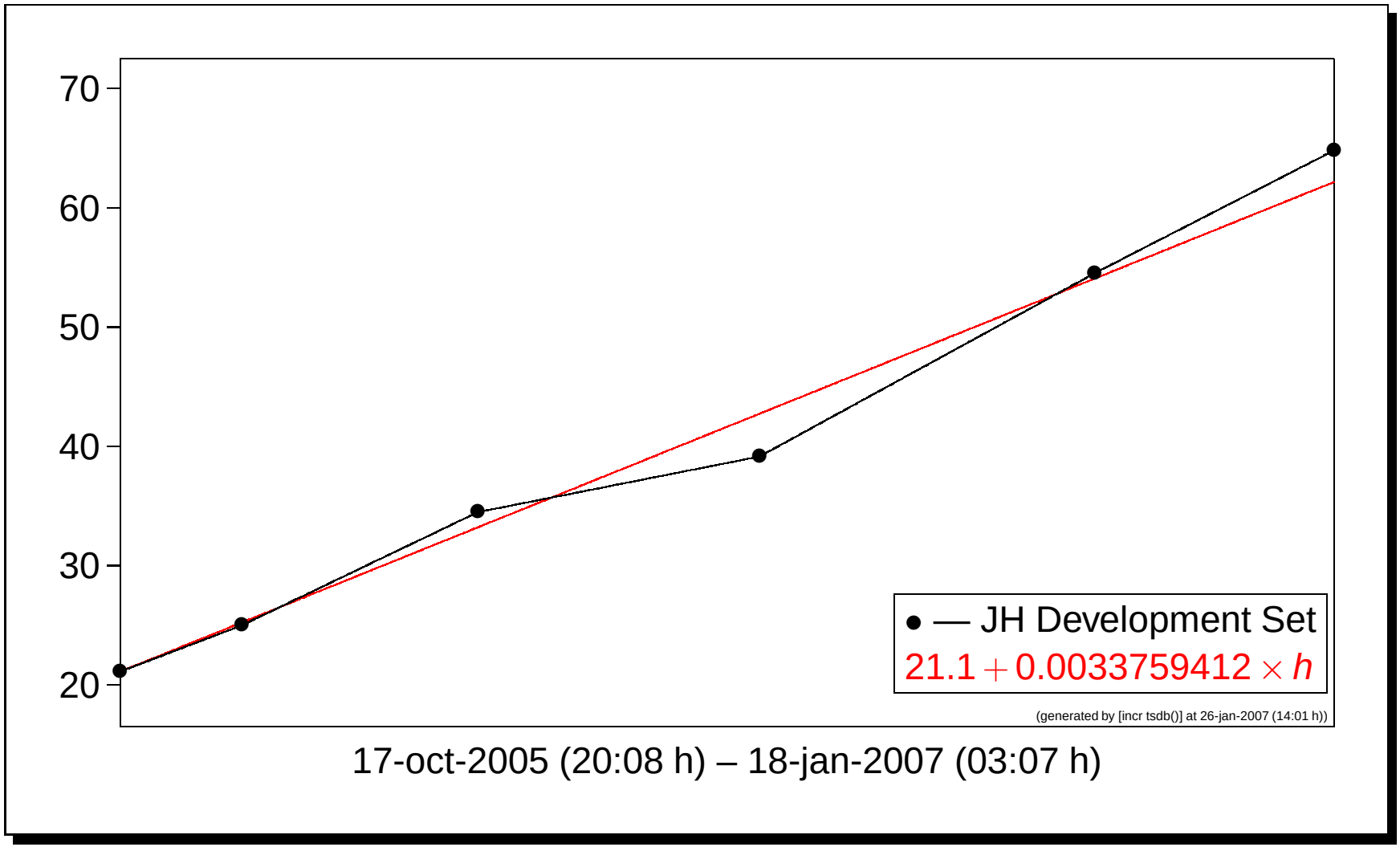
- LinGO ERG (Flickinger et al.): 13^{+} years; domain vocabulary → $94.3\,\%$;

+ structural MaxEnt plus BNC language model (Velldal & Oepen, 2006).

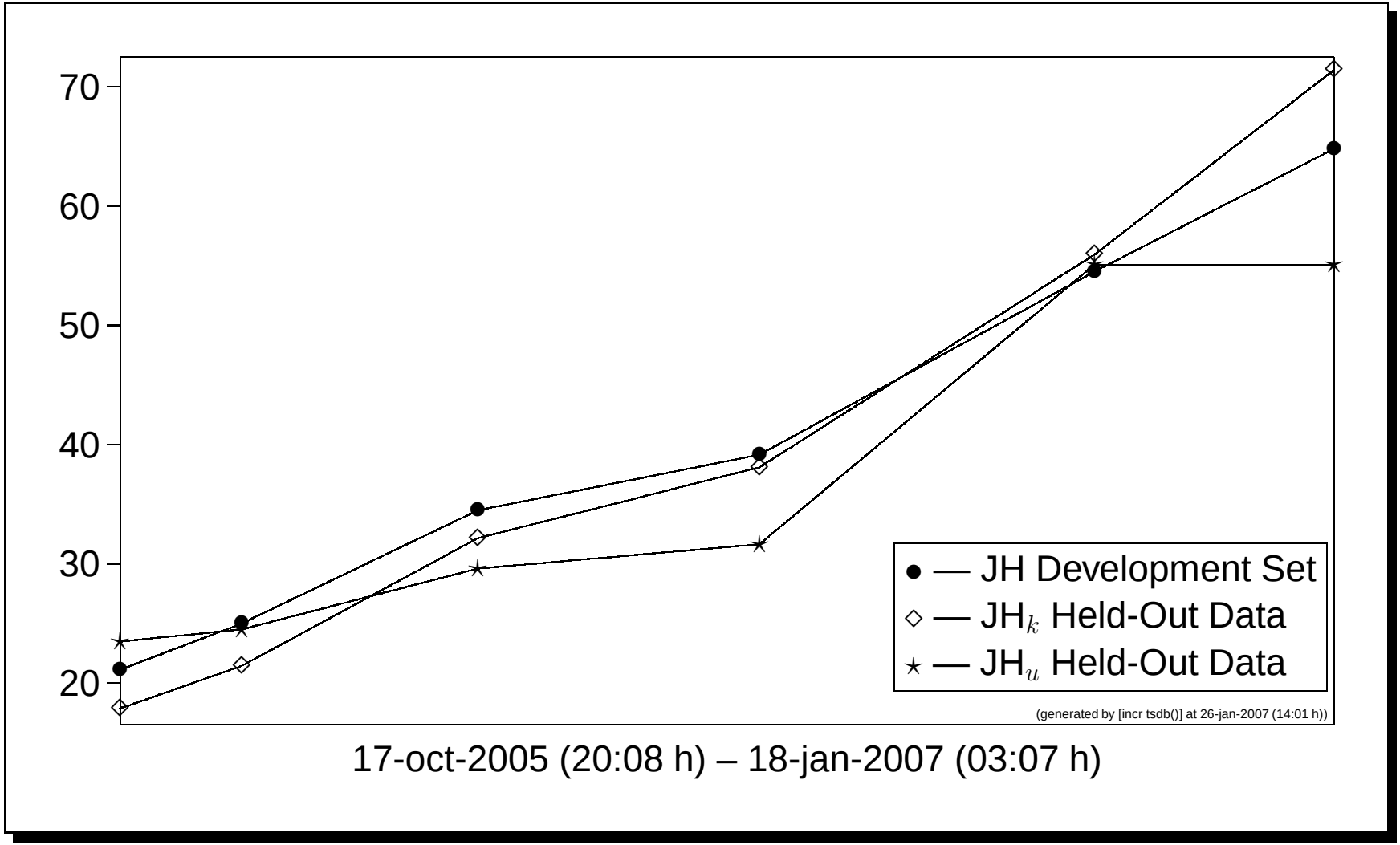
- 7627 hand-built transfer rules, 9222 from bi-lingual dictionary → $92.4\,\%$;

+ LM of dependency tuples: MRS 'fluency' by similarity to domain MRSs.

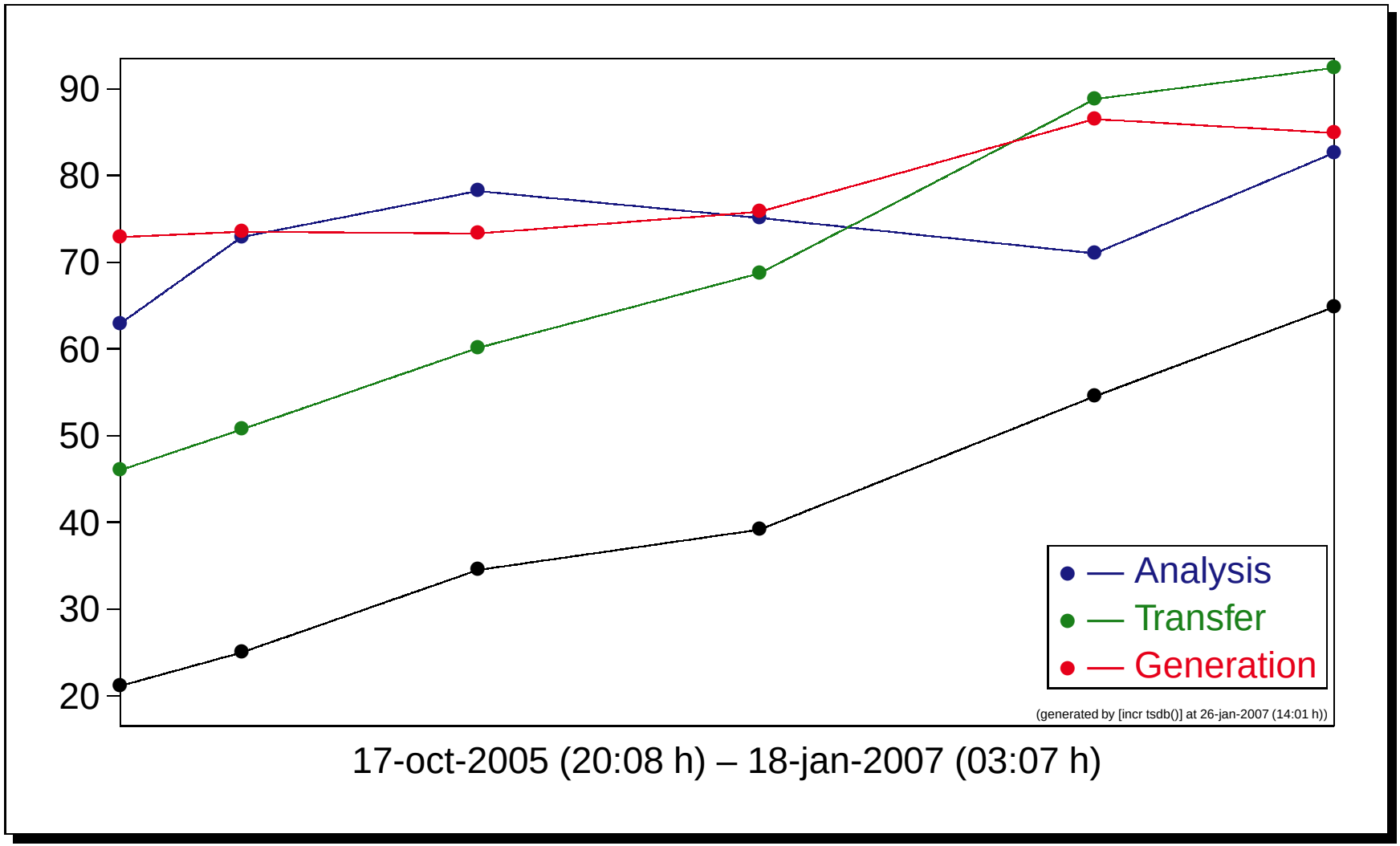
The Quest for Coverage: End-to-End Throughput



Comparing Development and Held-Out Data



Evolution of Individual Components (on JH)

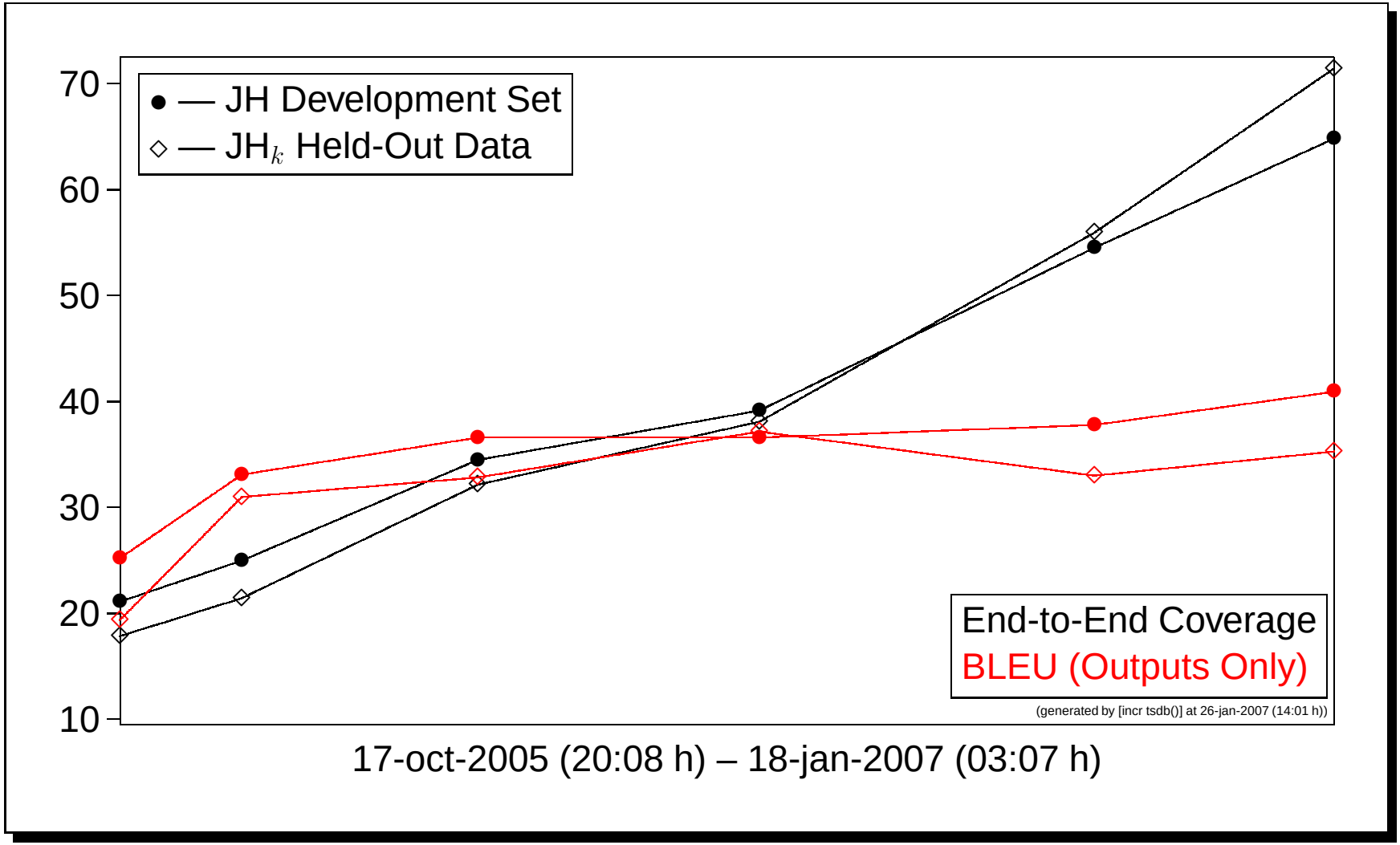


End-to-End Coverage vs. Input Length

'gold/logon/jh' Coverage Profile						
Aggregate	**total items** ♯	**positive items** ♯	**word string** ϕ	**distinct analyses** ϕ	**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))

One (Allegedly) Objective Measure (on JH)



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	first	top	judge
fidelity	1.28	1.59	1.83	1.94	2.05
fluency	1.27	1.31	1.62	1.69	1.79

End-to-End Re-ranking: Features (Very Briefly)

The Core: Per-Component Scores

- **Parse Selection** Unnormalized MaxEnt score (i.e. sum of weights);
- **Transfer Outputs** n-gram perplexity against 'semantic' model (LM);
- **Realization Ranking** unnormalized MaxEnt score (sum of weights).

$\rightarrow$ Using unnormalized scores to prevent 'vote dilution' across branches.

Additional Properties (Tried So Far)

- **Language Model** tri-gram model, trained on BNC plus domain corpus;
- **Lexical Translation Probabilities** GIZA^{++} word-to-word alignment;
- **Distortion** re-ordering, based on MRS EPs and sub-string pointers;
- **Harmony** string length and MRS size proportions: $|e|/|f|$ and $|E|/|F|$.

MaxEnt Re-Ranking: Preliminary Results

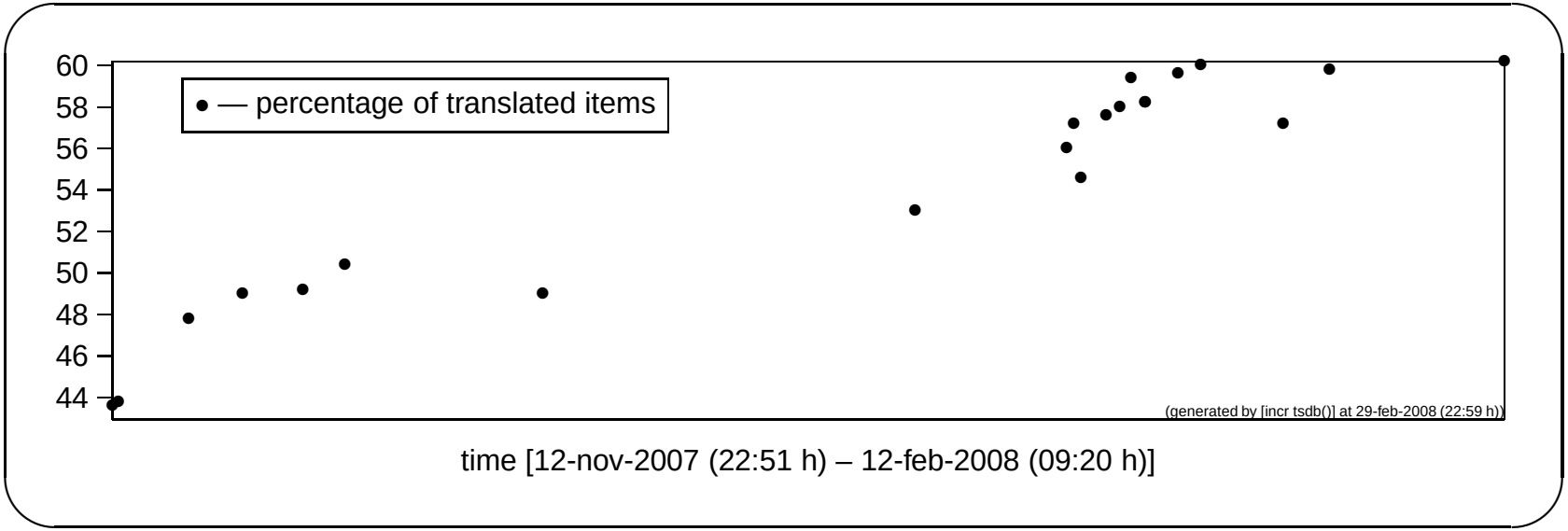
- Log-linear re-ranker, trained on development corpus (Och & Ney, 2002);
- use NEVA (sentence-level BLEU) to rank and 'label' candidate outputs.

data	#	words	coverage	strings
JH_d	2146	12.6	64.8%	266
JH_t	182	11.7	63.2%	114.6

data	#	chance	first	LL	top	judge
JH_d	1391	34.18	40.95	44.10	49.89	–
JH_t	115	30.84	35.67	38.92	45.74	46.32

On Scalability — A Limited Experiment

- Post-LOGON, try to double vocabulary size (5,000 → 10,000 lexemes);
- short pilot: about eight person months (kEUR 100); one new developer;
- web harvest ~270k tokens NO (~430k EN); general tourism information;
- mechanic, partiallyautomized lexicon extensions to all three grammars.



Preliminary Conclusions — Outlook

LOGON Results To Date

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

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

→ (way) too much unpacking; need ability to explore larger search space.

Towards Re-Usable (MT) Technology

- All LOGON modules (but the XLE) available publicly as open-source;
- baby JA – EN system built in one afternoon; now developed at NICT;
- automatic DE – EN transfer acquisition (DFKI); Matrix-Based MT (UW).

On-Line: `http://www.emmtee.net/` **and** `http://www.delph-in.net/`

A (Not Quite) Half-Baked Vision

End-to-End Ambiguity Factoring (Packing)

- Chart parsers and generators internally manipulate a *packed forest*;

→ weighted *and–or graph*: exhaustive search computationally tractable;

\+ selective unpacking: exact inference, avoid exponential combinatorics;

– currently no (interesting) packing at transfer level; breadth-first rewriting;

- MRS structurally similar to many predicate–argument representations.

Re-Usable Technology for (Transfer-Based) MT

? Disjunctive MRS(-like) graph structures as weighted and–or graph;

? weighted rewriting process (aka semantic transfer) as graph search;

? cascaded weighted and–or search: generalized (open-source) toolkit.

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.