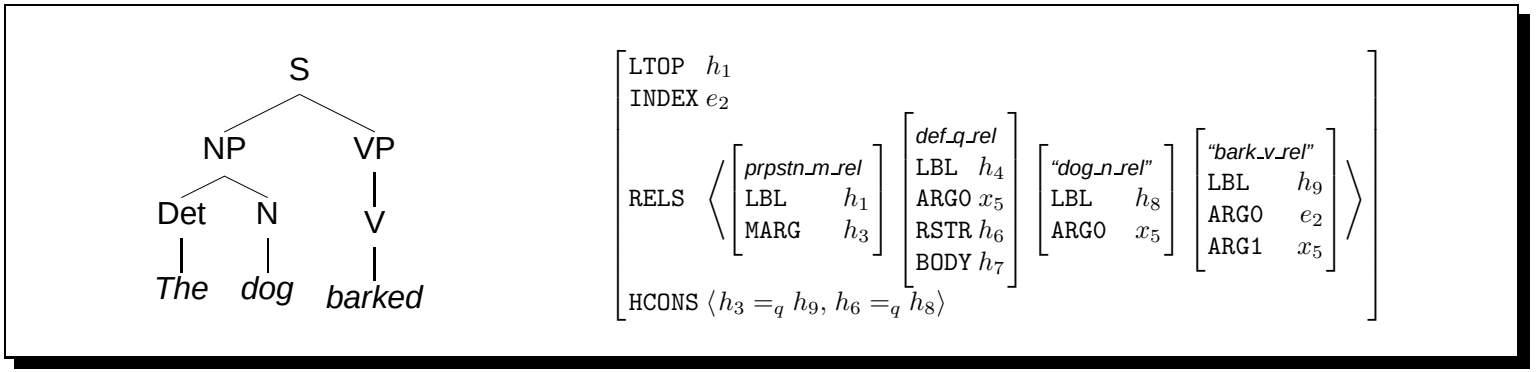


Computational Linguistics (INF2820 — Wrapping Up)

$$\textit{phrase}\begin{bmatrix}\texttt{HEAD} & \boxed{1} \\ \texttt{SPR} & \langle\rangle \\ \texttt{COMPS} & \boxed{3}\end{bmatrix} \longrightarrow \boxed{2}\textit{phrase}\begin{bmatrix}\texttt{SPR} & \langle\rangle \\ \texttt{COMPS} & \langle\rangle\end{bmatrix}, \textit{phrase}\begin{bmatrix}\texttt{HEAD} & \boxed{1} \\ \texttt{SPR} & \langle\boxed{2}\rangle \\ \texttt{COMPS} & \boxed{3}\end{bmatrix}$$

~~Wilhelm~~ Stephan Oepen & Jan Tore Lønning

Universitetet i Oslo

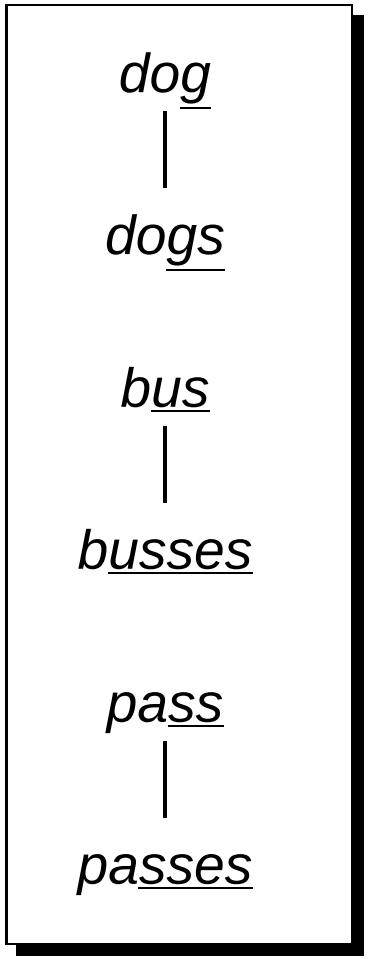
oe@ifi.uio.no

Orthographemic Variation: Inflectional Rules

```
%(letter-set (!s abcdefghijklmnopqrtuvwxyz))

noun-non-3sing_irule :=
%suffix (!s !ss) (!ss !ssses) (ss sses)
non-3sing-word &
[ HEAD [ AGR non-3sing ],
  ARGS < noun-lxm > ].

noun-3sing_irule :=
3sing-word &
[ ORTH #1,
  ARGS < noun-lxm & [ ORTH #1 ] > ].
```



LinGO English Resource Grammar

Linguistic Grammars On-Line (`http://lingo.stanford.edu/erg`)

- LinGO English Resource Grammar (Dan Flickinger et al., since 1993);
- general-purpose HPSG; domain-specific lexica (some 32,000 lexemes);
- development using LKB; high-efficiency $C^{[++]}$ parser for applications;
- domain-specific vocabulary addition and tuning $\rightarrow$ $\sim 85^{+}\%$ coverage;
- average parse times: a few seconds per sentence, for Wikipedia text;

$\longrightarrow$ exact same resource used simultaneously in many (research) 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.

Adding Semantics to Unification Grammars

- **Logical Form**

  For each sentence admitted by the grammar, we want to produce a meaning representation that is suitable for applying rules of inference.

  This fierce dog chased that angry cat.

  $$this(x) \wedge fierce(x) \wedge dog(x) \wedge chase(e,x,y) \wedge past(e) \wedge that(y) \wedge angry(y) \wedge cat(y)$$

- **Compositionality**

  The meaning of each phrase is composed of the meanings of its parts.

- **Existing Machinery**

  Unification is the only means for constructing semantics in the grammar.

(Elementary) Semantics in Typed Feature Structures

- Encode semantic content in the SEM attribute of every word and phrase:

$$\textit{expression}\begin{bmatrix}\texttt{HEAD} & \textit{pos}\\ \texttt{SPR} & \textit{*list*}\\ \texttt{COMPS} & \textit{*list*}\\ \texttt{SEM} & \textit{semantics}\begin{bmatrix}\texttt{RELS}\ \textit{*dlist*}\end{bmatrix}\end{bmatrix}$$

- The value of SEM for a sentence is simply a list of relations in the attribute RELS, with the arguments in those relations 'linked up' appropriately:

$$\begin{bmatrix}\texttt{RELS}\ \left\langle\begin{bmatrix}\texttt{PRED}\ \textit{“the\_rel”}\\ \texttt{ARG0}\ \boxed{1}\textit{entity}\end{bmatrix},\begin{bmatrix}\texttt{PRED}\ \textit{“dog\_rel”}\\ \texttt{ARG0}\ \boxed{1}\end{bmatrix},\begin{bmatrix}\texttt{PRED}\ \textit{“bark\_rel”}\\ \texttt{ARG0}\ \textit{event}\\ \texttt{ARG1}\ \boxed{1}\end{bmatrix}\right\rangle\end{bmatrix}$$

- Semantic relations are introduced by lexical entries, and are appended when grammar rules combine words with other words or phrases.

Linking Semantic Arguments

- Each word or phrase carries an associated variable: its INDEX 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

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

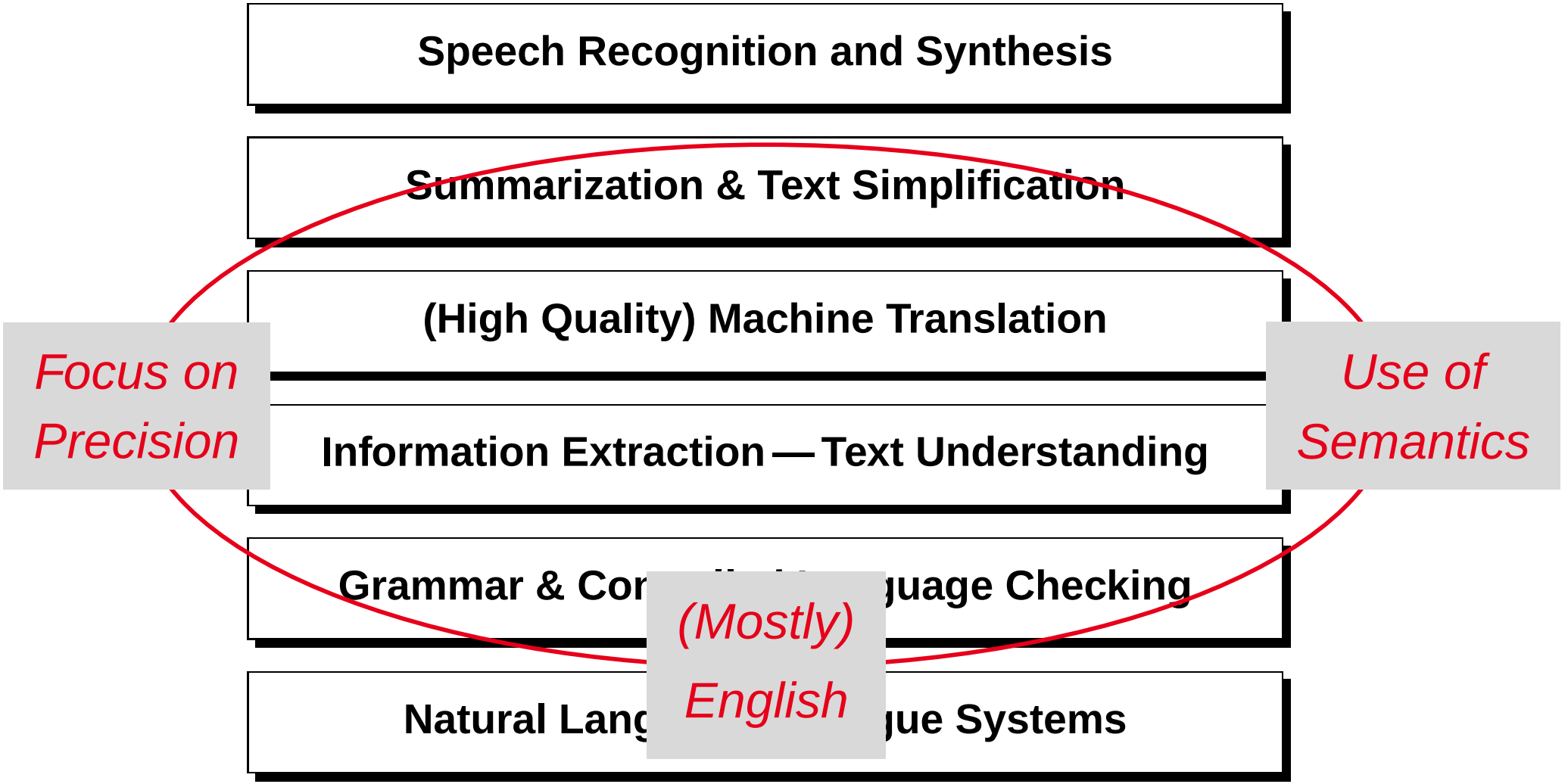
Semantics of Phrases

- 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.

Families of Language Processing Tasks

- **Speech Recognition and Synthesis**
- **Summarization & Text Simplification**
- **(High Quality) Machine Translation**
- **Information Extraction — Text Understanding**
- **Grammar & Controlled Language Checking**
- **Natural Language Dialogue Systems**

Families of Language Processing Tasks



What Makes Natural Language a Hard Problem?

```
< 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).
  ...
```

What Makes Natural Language a Hard Problem?

```
< 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 other way towards Bergen is short.

the road to the other bergen is short .

Den other roads against Boron Gene are short.

Other one autobahn against Mountains am abrupt.

The Rationalist vs. Empiricist Stand-Off

Every time I fire a linguist,
system performance goes up.

[Fred Jelinek, 1980s]

The Rationalist vs. Empiricist Stand-Off

Every time I fire a linguist,
system performance goes up.

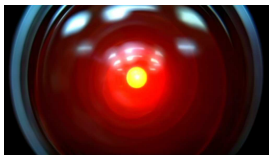
[Fred Jelinek, 1980s]

Competition of Paradigms

- Rationalist: formally encode linguistic and extra-linguistic knowledge;
- empiricist: statistical models trained on distributional data (corpora);
- older and wiser Jelinek today: *Some of my best friends are linguists.*

→ hybrids: combination of approaches required for long-term success.

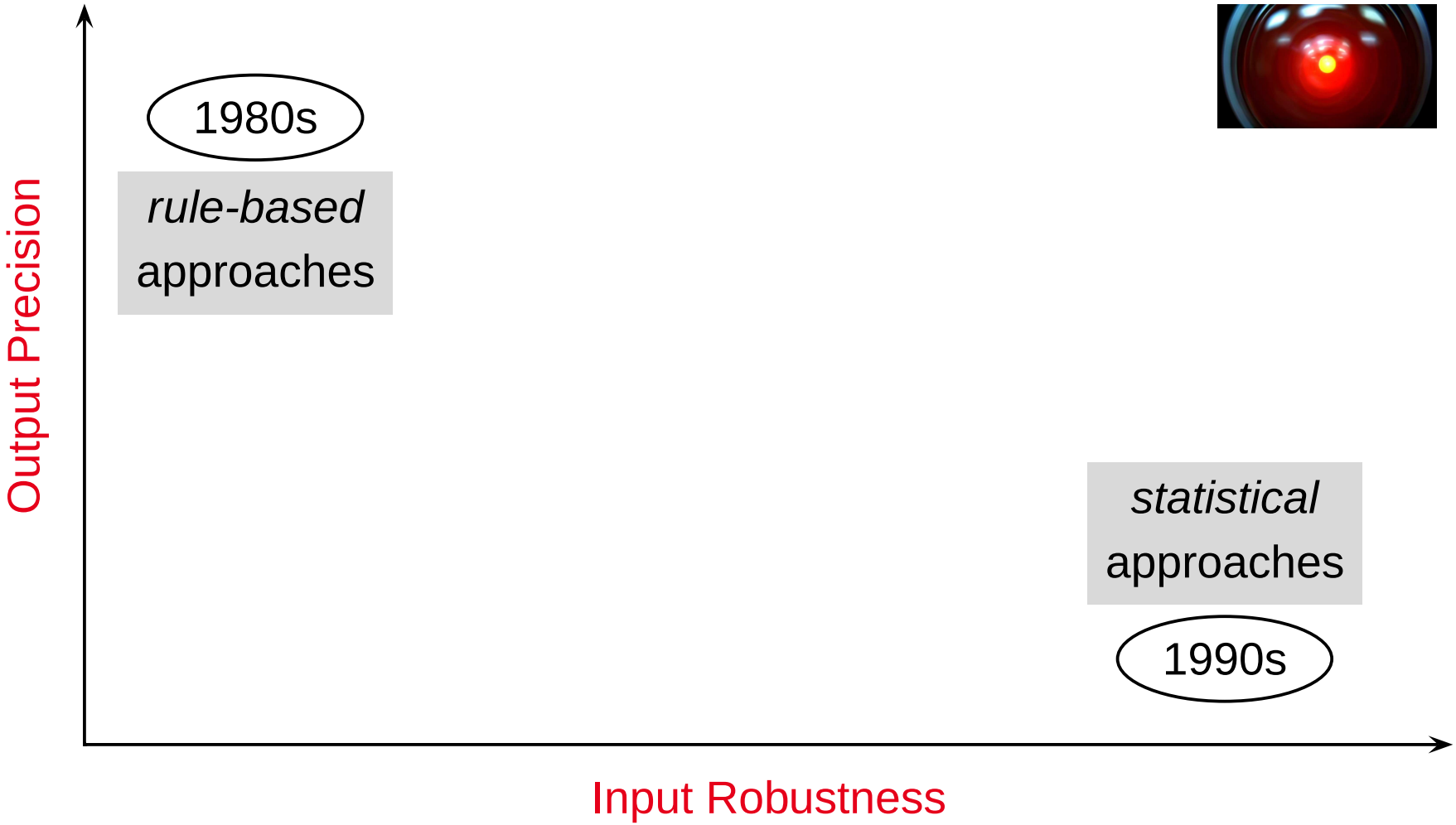
The Holy Grail: Balancing Precision and Robustness



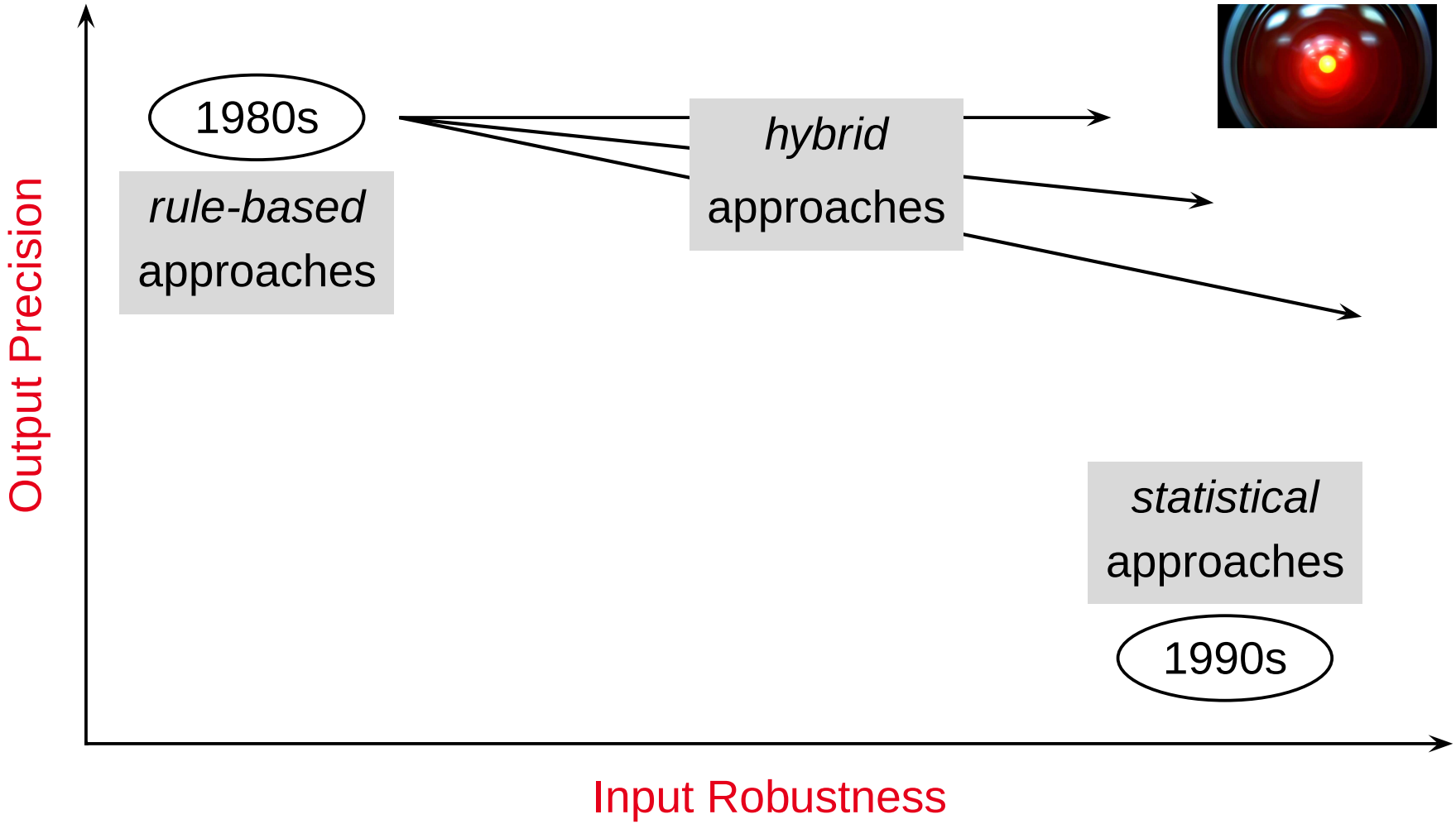
Output Precision

Input Robustness

The Holy Grail: Balancing Precision and Robustness



The Holy Grail: Balancing Precision and Robustness



Linguistics, Statistics, 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]

Linguistics, Statistics, 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]

Linguistics, Statistics, 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.

Linguistics, Statistics, 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;

$\longrightarrow$ 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

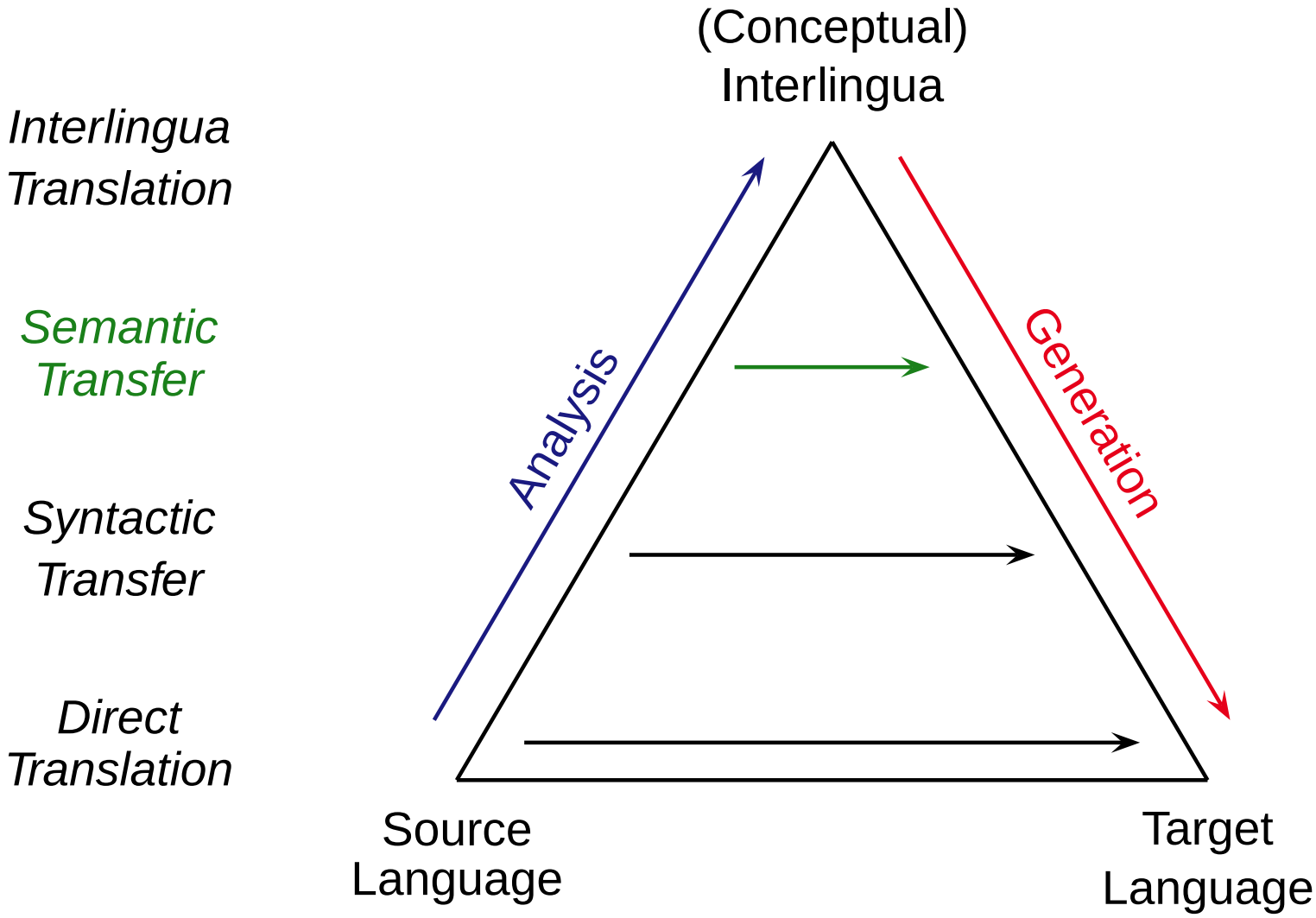
(New York Times, January 8, 1954)

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.

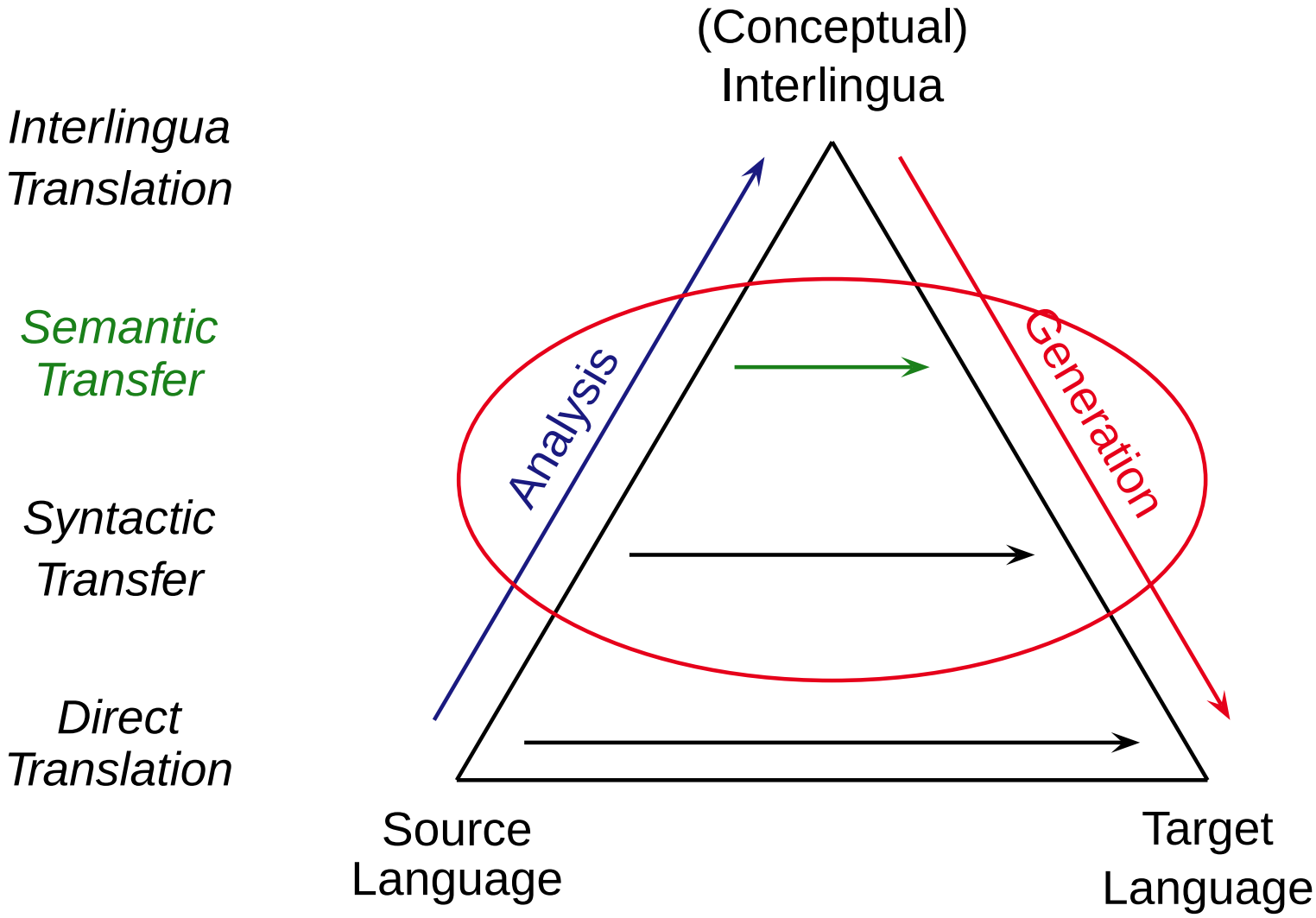
(Journal of Franklin Institute, March 1954)

- 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)



Dimensions of Machine Translation (Vauquois, 1968)



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.

Looking Back: 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

The Grar[illegible]lenges

Broad Progress

→ MT research raised found[illegible] 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

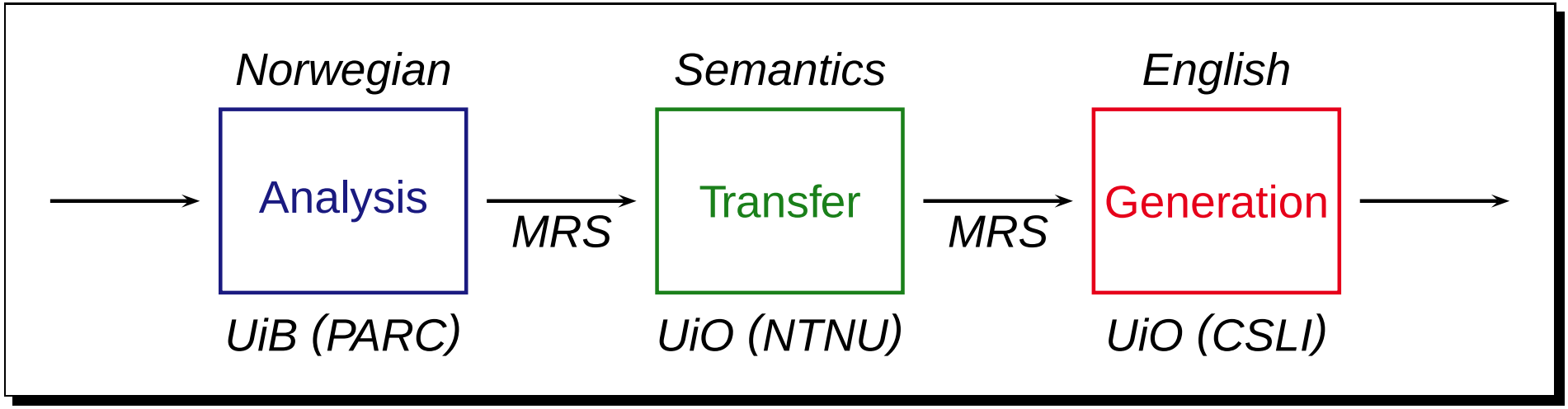
The Grand Challenges

→ MT research raised foundational questions for language processing:

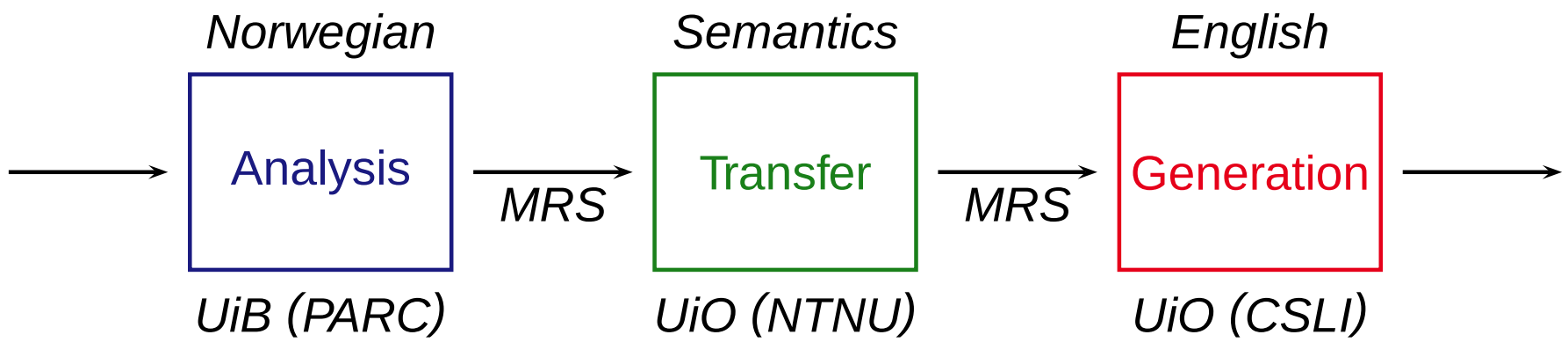
+ **representation** formalizing and encoding of linguistic knowledge;
+ **declara** mation;
+ **reversa** neration;
+ **compu** e data;
+ **re-usa** nt tools;
+ **sustai** oration.

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

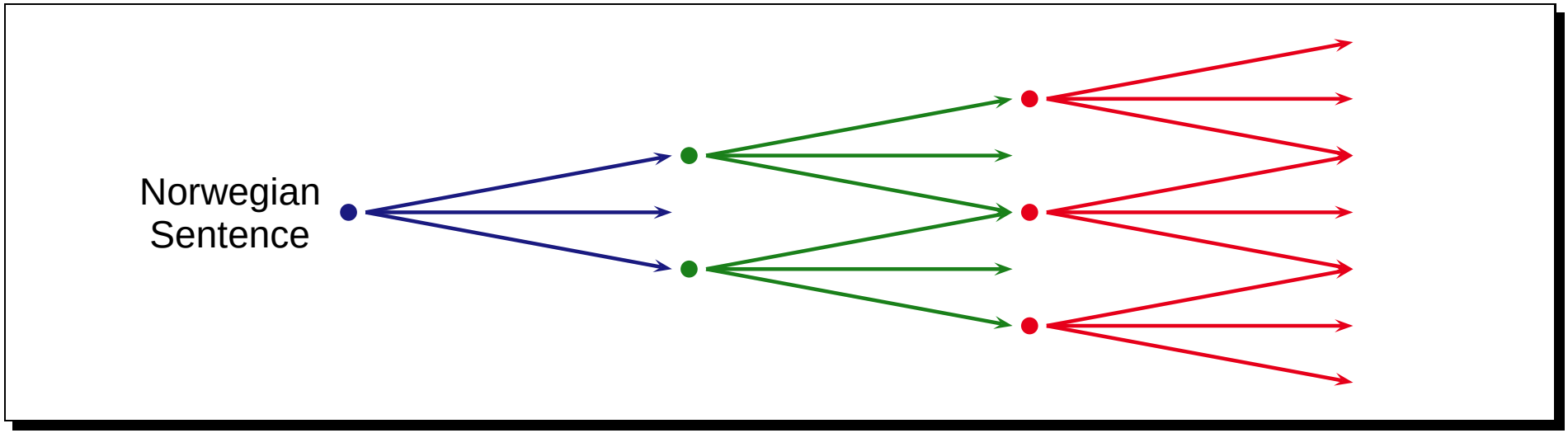


Some LOGON Highlights

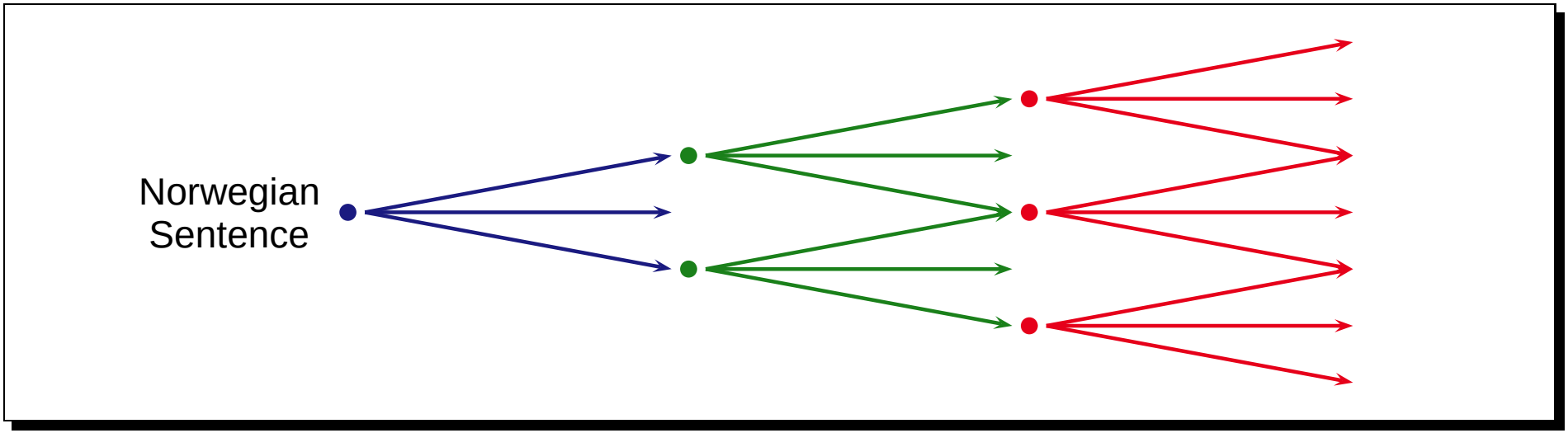
- Re-usable, mono-lingual precision grammars as linguistic back-bone;
- abstract from language-internal idiosyncrasies by semantic transfer;

$\rightarrow$ 'plug & play' of general-purpose resources for flexible MT framework.

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.

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.