

Computational Linguistics (INF2820 — Outlook)

The Second Steep Road Against Bergen is a Card

Stephan Oepen

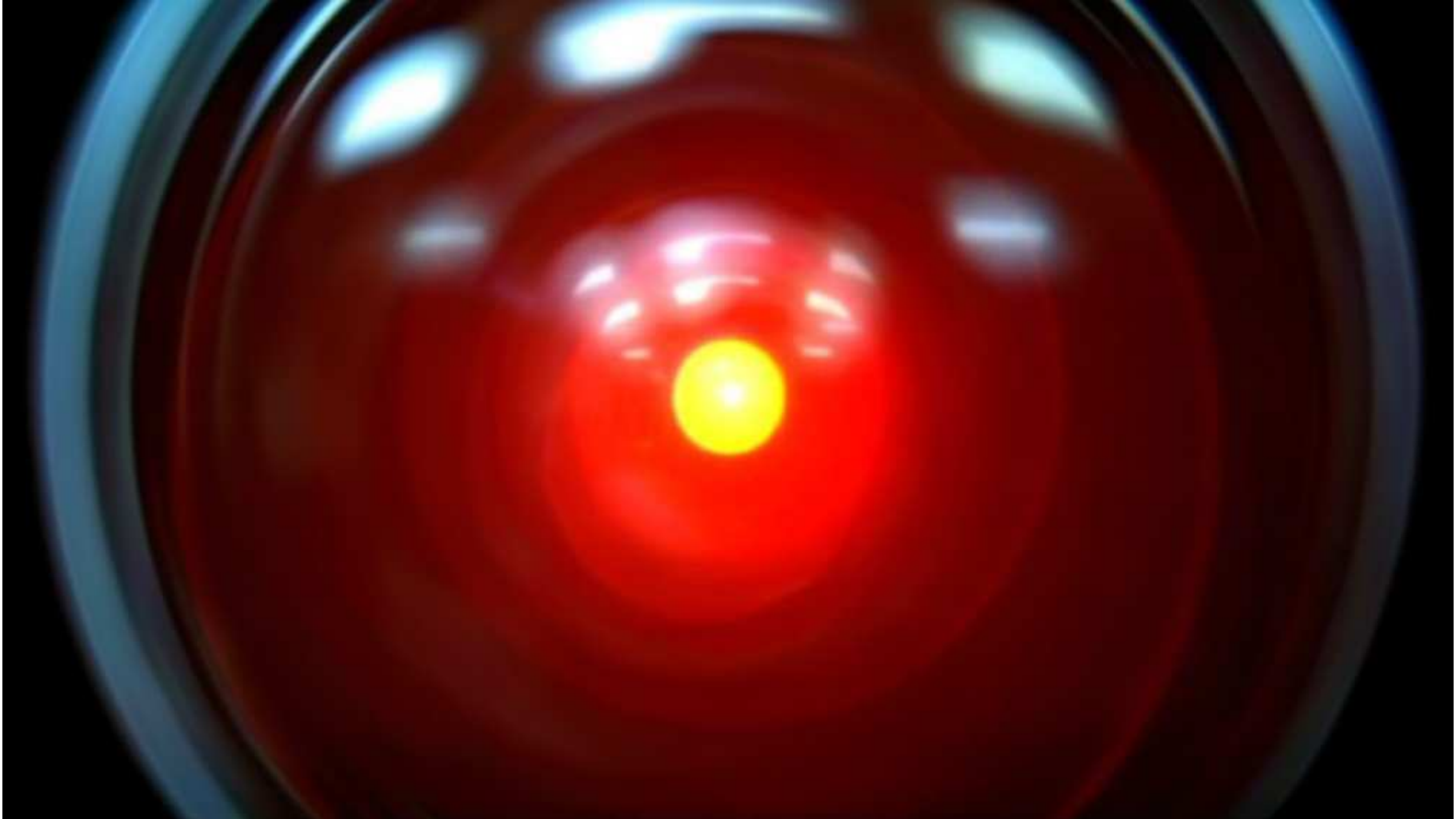
Universitetet i Oslo

oe@ifi.uio.no

So, What Actually is Computational Linguistics?



So, What Actually is Computational Linguistics?



(2001: A Space Odyssey; HAL 9000; 1968)



INF2820 — 27-MAY-10 (oe@ifi.uio.no)

Introduction to Computational Linguistics (2)

No, Really, What is Computational Linguistics?

... teaching computers our language. (Alien Researcher, 2000)



No, Really, What is Computational Linguistics?

... teaching computers our language. (Alien Researcher, 2000)

We Understand™. Unlike other solutions based on keyword or phrase recognition, YY Software's product actually understands customer e-mails and Web interaction. By automatically and accurately answering e-mail and Web requests, YY Software's flagship product [...] can produce high-benefit value proposition that increases customer satisfaction. (Start-Up Marketing Blurb, 2000)



No, Really, What is Computational Linguistics?

... teaching computers our language. (Alien Researcher, 2000)

We Understand™. Unlike other solutions based on keyword or phrase recognition, YY Software's product actually understands customer e-mails and Web interaction. By automatically and accurately answering e-mail and Web requests, YY Software's flagship product [...] can produce high-benefit value proposition that increases customer satisfaction. (Start-Up Marketing Blurb, 2000)

... the scientific study of human language—specifically of the system of rules and the ways in which they are used in communication—using mathematical models and formal procedures that can be realized and validated using computers; a cross-over of many disciplines. (Stanford Linguistics Researcher, 2003)



Families of Language Processing Tasks

Speech Recognition and Synthesis

Summarization & Text Simplification

(High Quality) Machine Translation

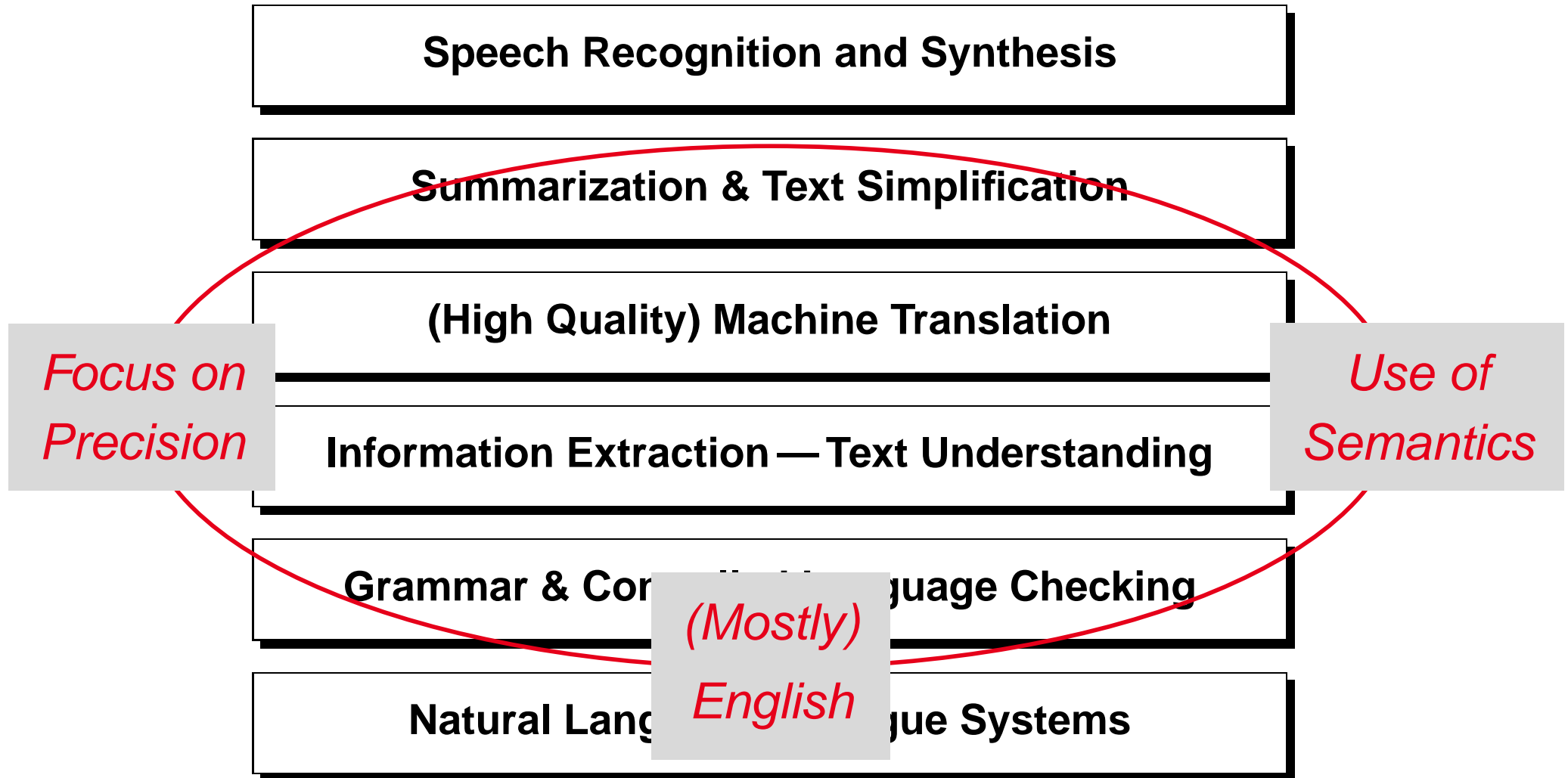
Information Extraction — Text Understanding

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).

Scraped Off the Internet

- .. The other way towards Bergen is short.
- > Sh the road to the other bergen is short .
- > Sh Den other roads against Boron Gene are short.
- .. Other one autobahn against Mountains am abrupt.



A Tool Towards Understanding: (Formal) Grammar

Wellformedness

- *Kim was happy because _____ passed the exam.*
- *Kim was happy because _____ final grade was an A.*
- *Kim was happy when she saw _____ on television.*



A Tool Towards Understanding: (Formal) Grammar

Wellformedness

- *Kim was happy because _____ passed the exam.*
- *Kim was happy because _____ final grade was an A.*
- *Kim was happy when she saw _____ on television.*

Meaning

- *Kim gave Sandy the book.*
- *Kim gave the book to Sandy.*
- *Sandy was given the book by Kim.*



A Tool Towards Understanding: (Formal) Grammar

Wellformedness

- *Kim was happy because _____ passed the exam.*
- *Kim was happy because _____ final grade was an A.*
- *Kim was happy when she saw _____ on television.*

Meaning

- *Kim gave Sandy the book.*
- *Kim gave the book to Sandy.*
- *Sandy was given the book by Kim.*

Ambiguity

- *Kim saw the astronomer with the telescope.*
- *Have her report on my desk by Friday!*



A Grossly Simplified Example

The Grammar of Spanish

S → NP VP

VP → V NP

VP → VP PP

PP → P NP

NP → “nieve”

NP → “Juan”

NP → “Oslo”

V → “amó”

P → “en”

Juan amó nieve en Oslo



A Grossly Simplified Example

The Grammar of Spanish

S → NP VP

VP → V NP

VP → VP PP

PP → P NP

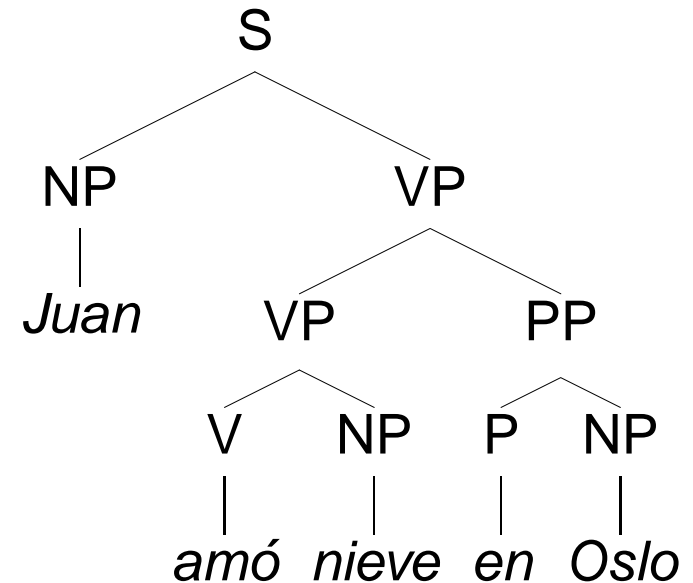
NP → “nieve”

NP → “Juan”

NP → “Oslo”

V → “amó”

P → “en”



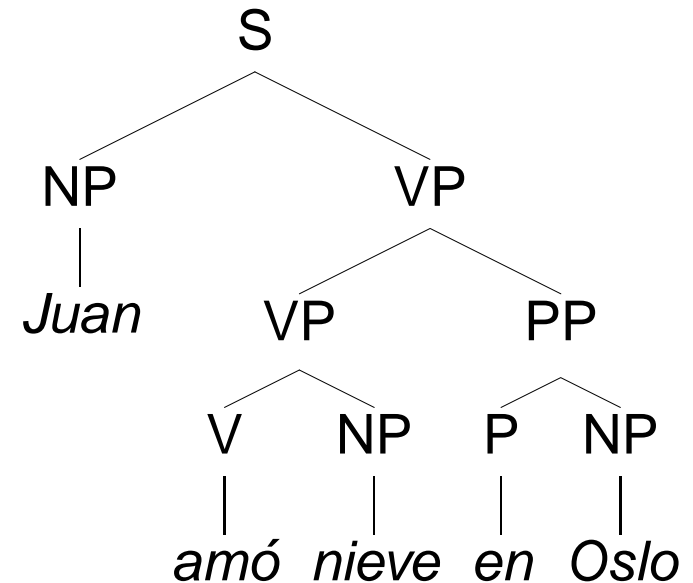
Juan amó nieve en Oslo



A Grossly Simplified Example

The Grammar of Spanish

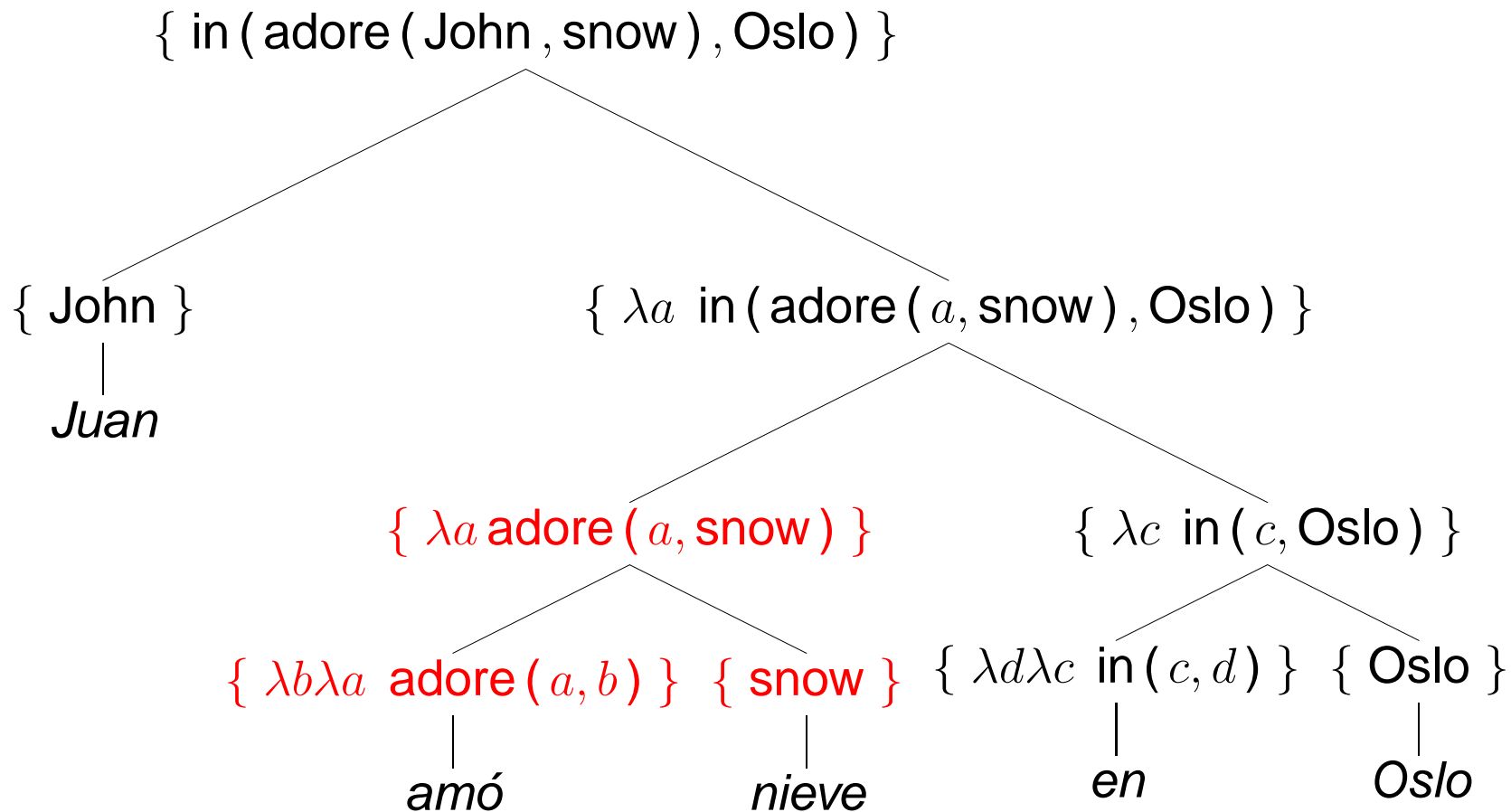
$S \rightarrow NP VP$	$\{ VP (NP) \}$
$VP \rightarrow V NP$	$\{ V (NP) \}$
$VP \rightarrow VP PP$	$\{ PP (VP) \}$
$PP \rightarrow P NP$	$\{ P (NP) \}$
$NP \rightarrow \text{"nieve"}$	$\{ \text{snow} \}$
$NP \rightarrow \text{"Juan"}$	$\{ \text{John} \}$
$NP \rightarrow \text{"Oslo"}$	$\{ \text{Oslo} \}$
$V \rightarrow \text{"amó"}$	$\{ \lambda b \lambda a \text{ adore } (a, b) \}$
$P \rightarrow \text{"en"}$	$\{ \lambda d \lambda c \text{ in } (c, d) \}$



Juan amó nieve en Oslo



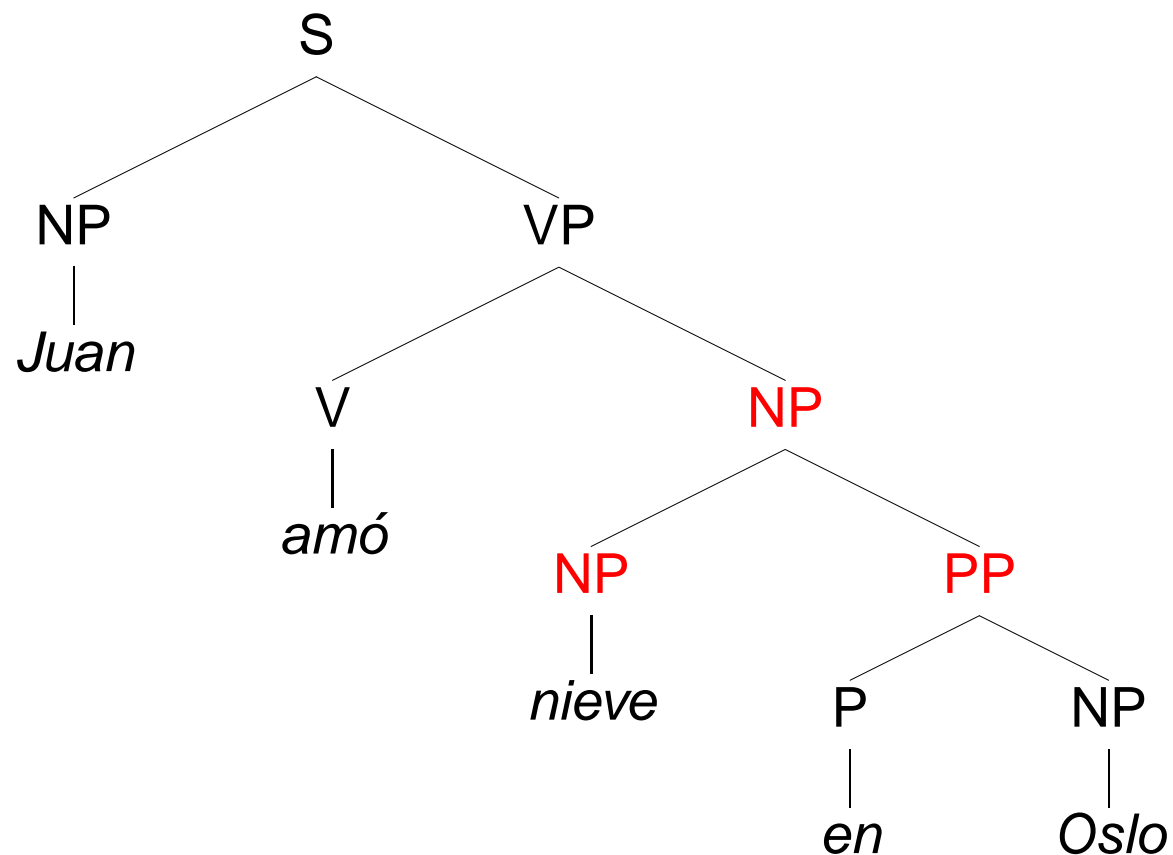
Meaning Composition (Grossly Simplified, Still)



$\text{VP} \rightarrow \text{V NP} \quad \{ \text{V} (\text{NP}) \}$



Another Interpretation — Structural Ambiguity



$NP \rightarrow NP PP \quad \{ PP(NP) \}$



Some Areas of Descriptive Grammar

Phonetics *The study of speech signals.*

Phonology *The study of sound systems.*

Morphology *The study of word structure.*

Syntax *The study of sentence structure.*

Semantics *The study of language meaning.*

Pragmatics *The study of language use.*



Some Areas of Descriptive Grammar

Phonetics *The study of speech signals.*

Phonology *The study of sound systems.*

Morphology *The study of word structure.*

Syntax *The study of sentence structure.*

Semantics *The study of language meaning.*

Pragmatics *The study of language use.*



More, and More, and More Ambiguity

Speech Recognition

<i>its</i>	<i>hard</i>	<i>to</i>	<i>wreck</i>	<i>a</i>	<i>nice</i>	<i>beach</i>
<i>it</i>	<i>'s</i>	<i>hard</i>	<i>to</i>	<i>recognize</i>		<i>speech</i>

Morphology

- *fisker* *fisk*_N + plural vs. *fiske*_V + present vs. *fisker*_N + singular;
- *brus-automat* vs. *bru-sau-tomat*; *vinduene* vs. *vin-duene*; et al.

Semantics

- *All Norwegians speak two languages.* $\exists l_1, l_2 \forall n \dots$ vs. $\forall n \exists l_1, l_2 \dots$



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.*

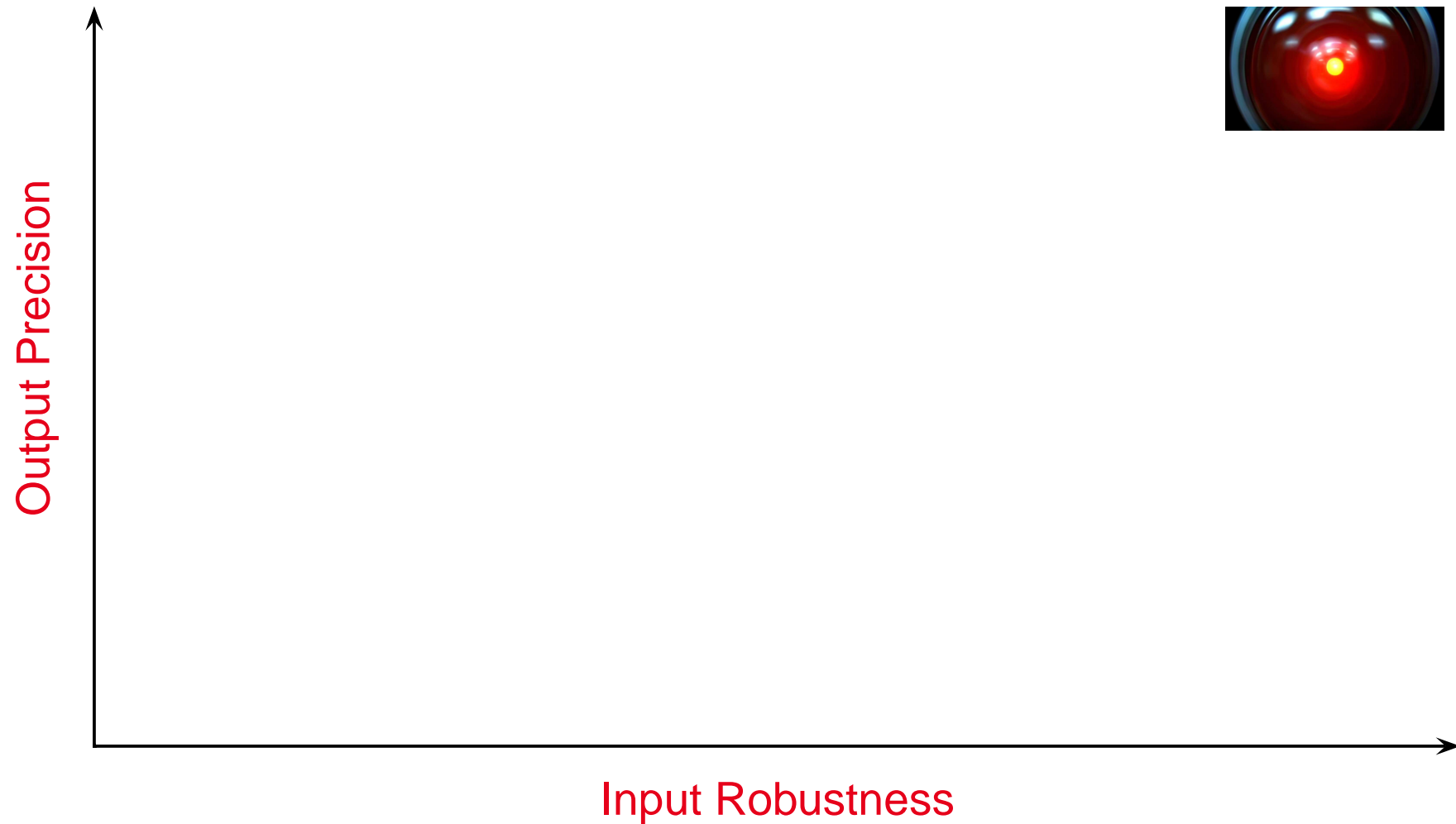
[Fred Jelinek, 1980s]

Competition of Paradigms

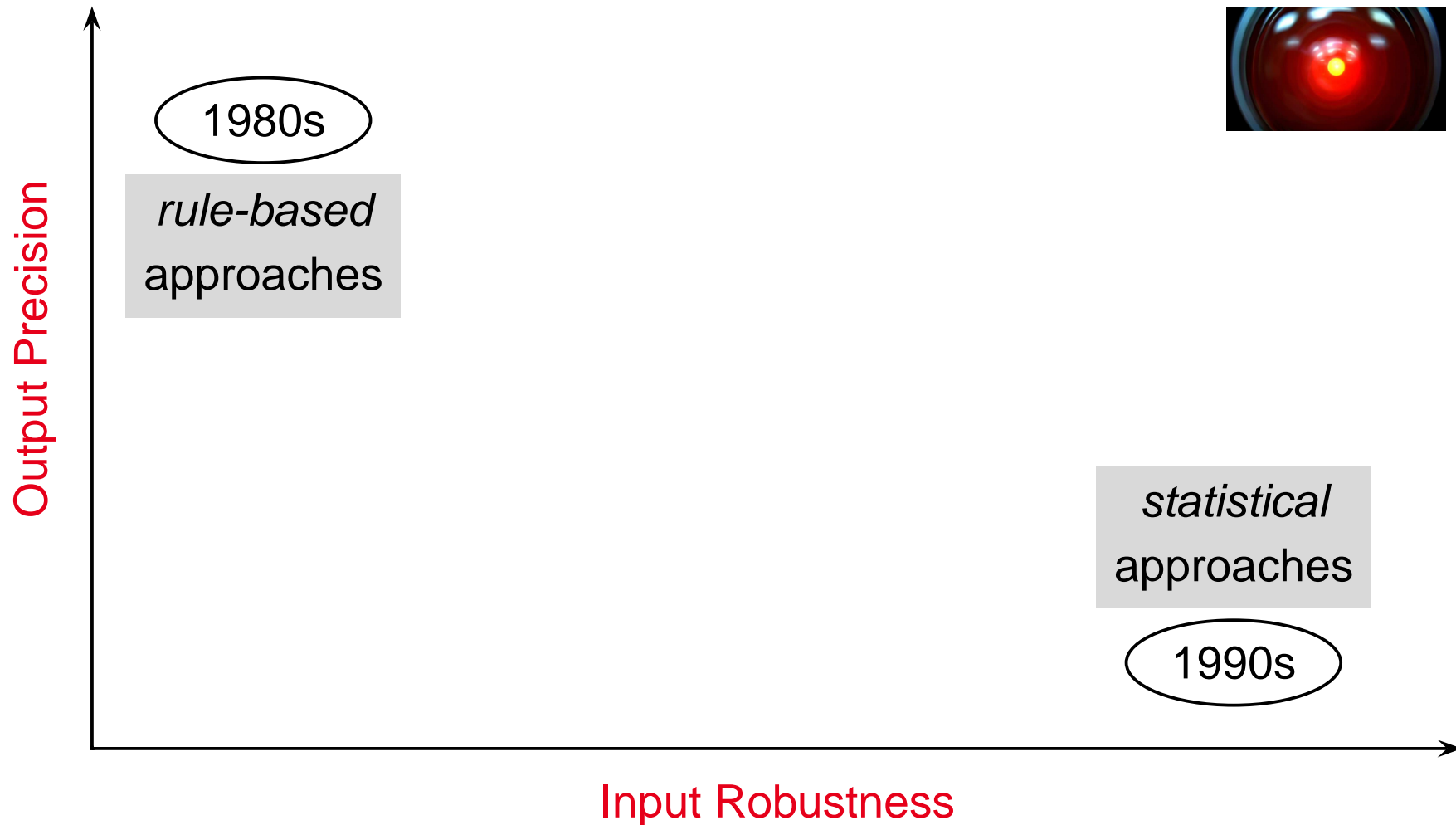
- Rationalist: formally encode linguistic and extra-linguistic knowledge;
 - empiricist: statistical models trained on distributional data (corpora);
 - older and wiser Jelinek today: *Some of my best friends are linguists.*
- hybrids: combination of approaches required for long-term success.



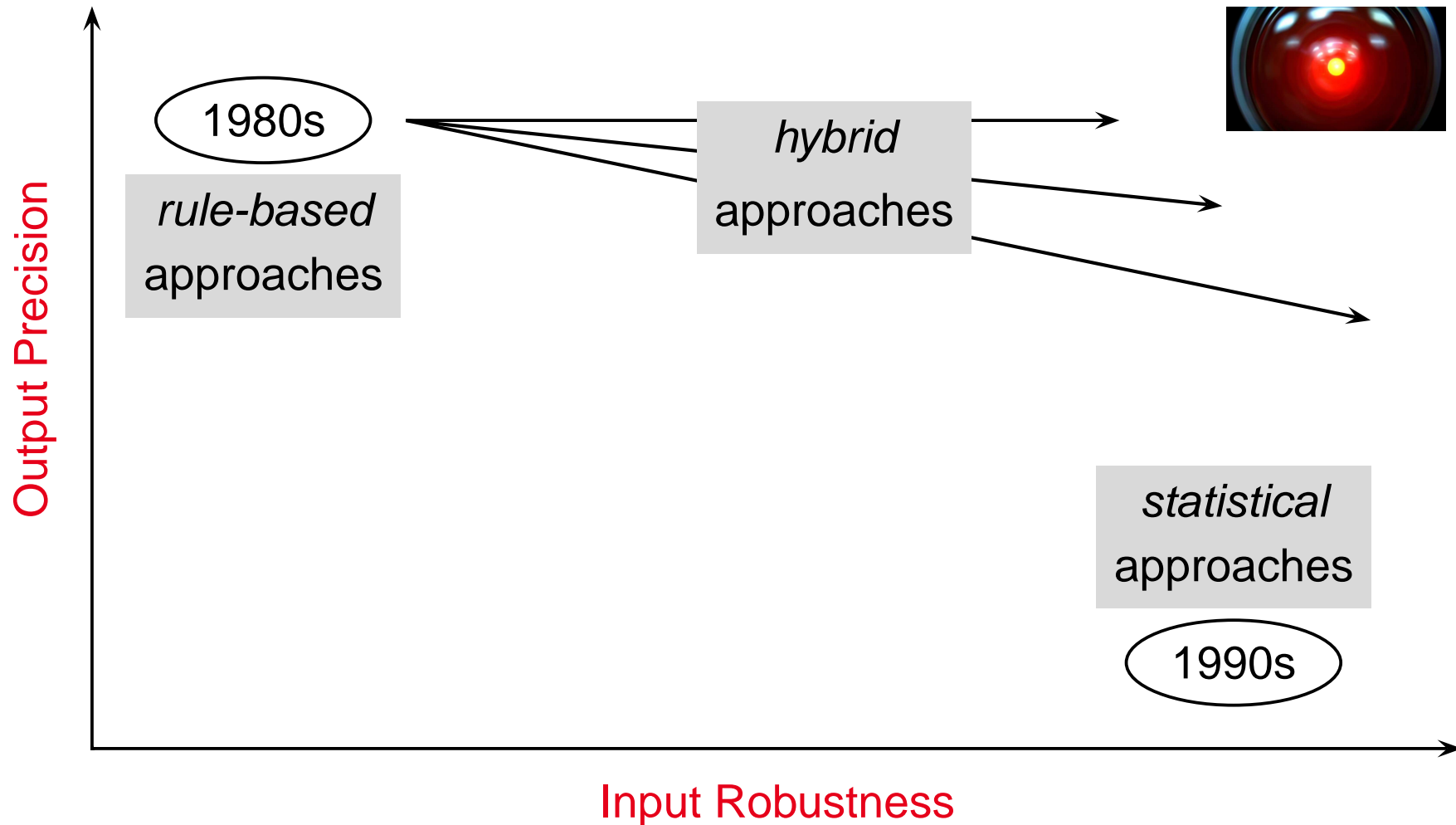
The Holy Grail: Balancing Precision and Robustness



The Holy Grail: Balancing Precision and Robustness



The Holy Grail: Balancing Precision and Robustness



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]

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]

(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;
- 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 'ggeneral' 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)

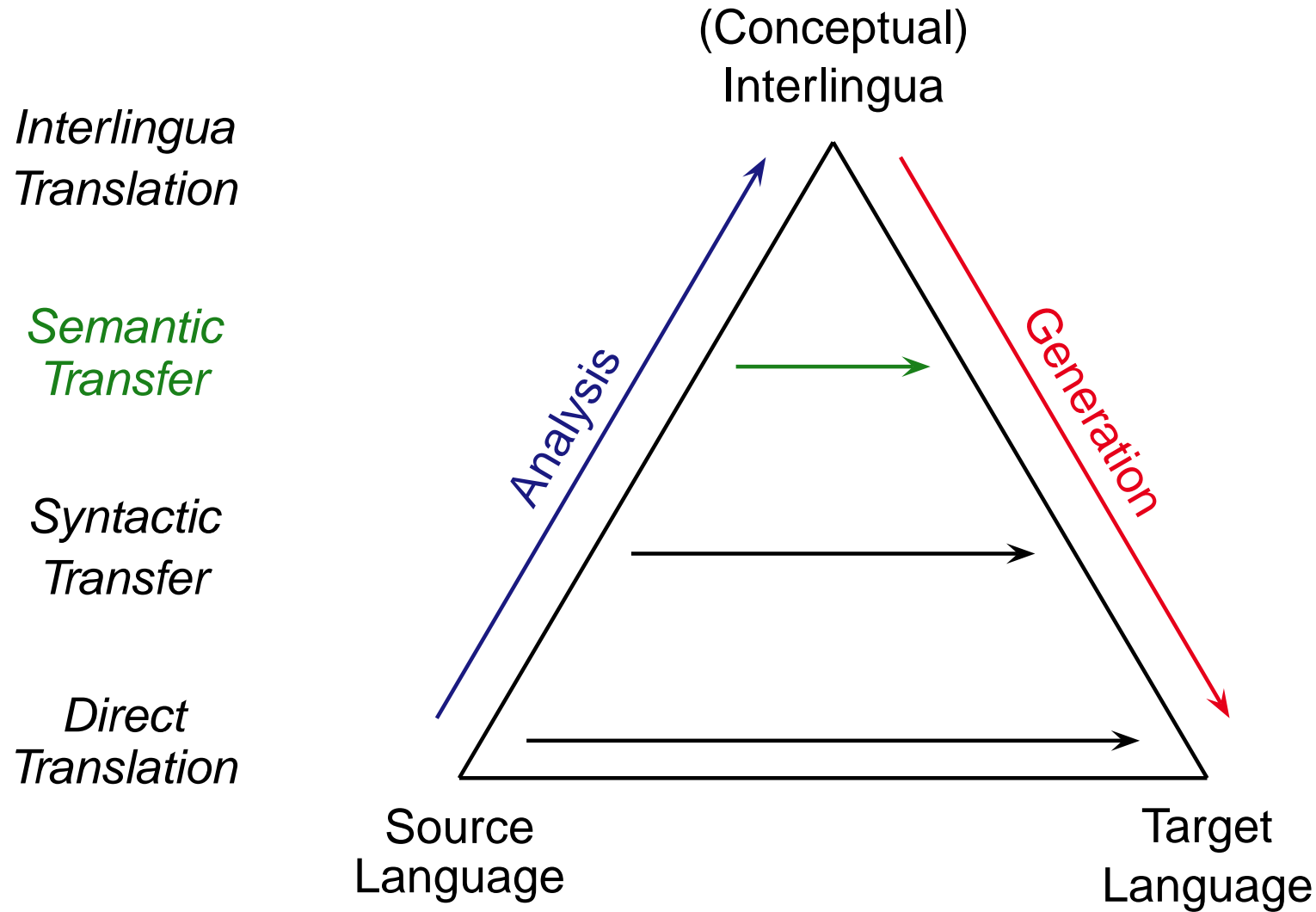
The switch is assured in advance by attaching the rule sign 21 to the Russian 'ggeneral' in the bilingual glossary which is stored in the machine, and by attaching the rule-sign 110 to the Russian 'major'. The stored instructions, along with the glossary, say whenever you read a rule sign 110 in the 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)

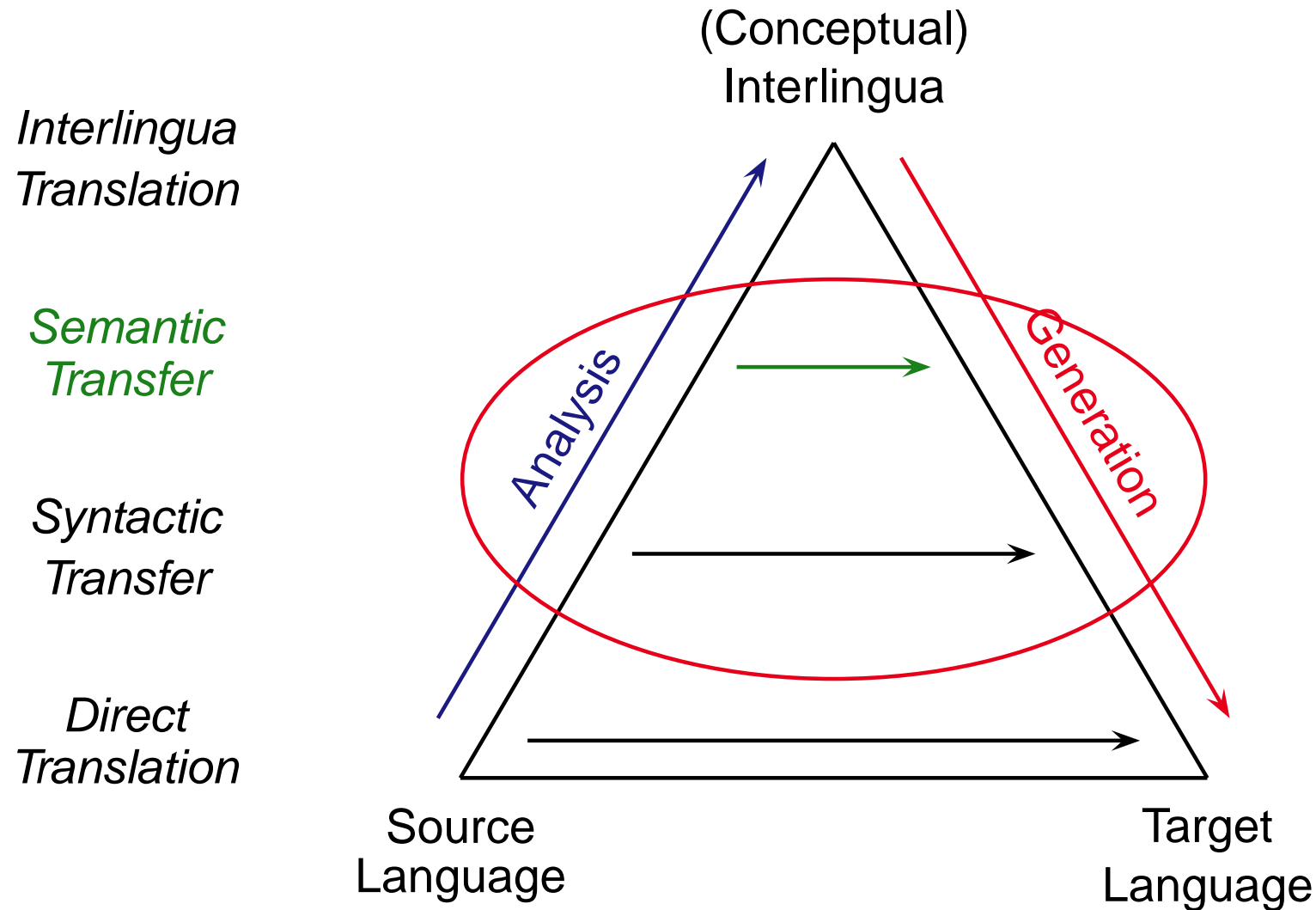
- Georgetown Experiment: first public MT demonstration (with IBM);
- minuscule scale: 250 words, six 'syntactic' rules → first MT boom.



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.



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

- ... 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 raised foundational questions for language processing:

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

+ **declarati** on;

+ **reversab** eration;

+ **computa** data;

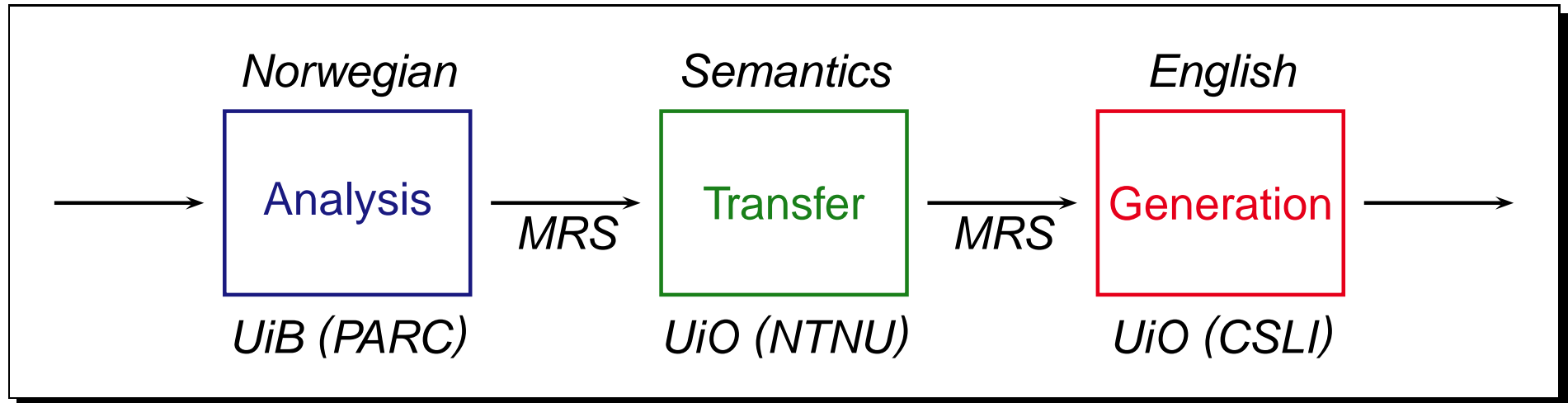
+ **re-usabi** tools;

+ **sustaina** tion.

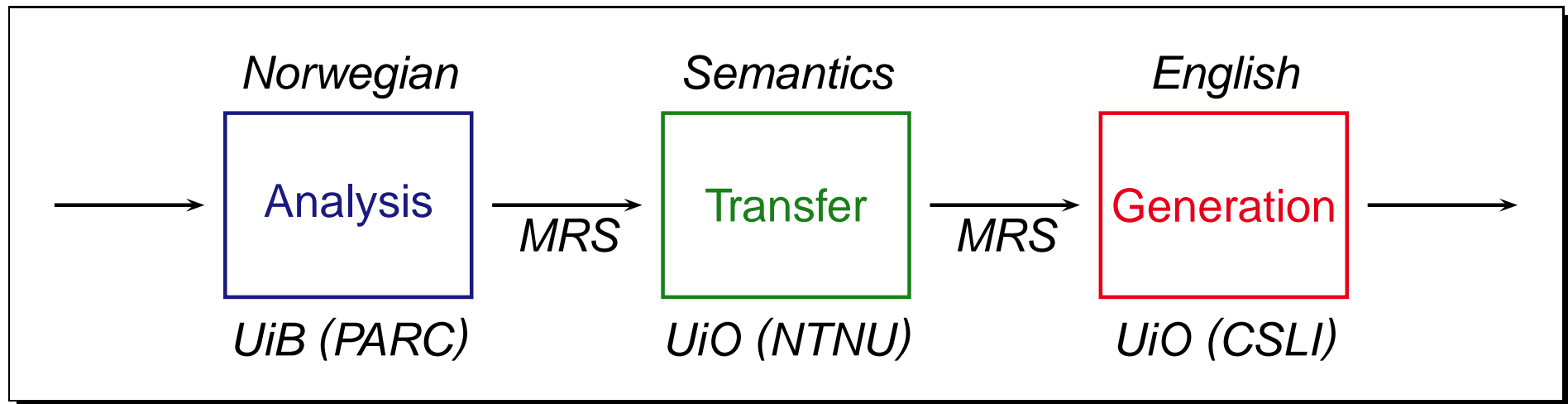
Since (around) the early 1990s, federations of computational linguists deploy advanced grammatical formalisms, high-efficiency tools, shared (interface) representations, and rigid development and evaluation methodologies— to the analysis of a growing set of languages, applied to many diverse tasks and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

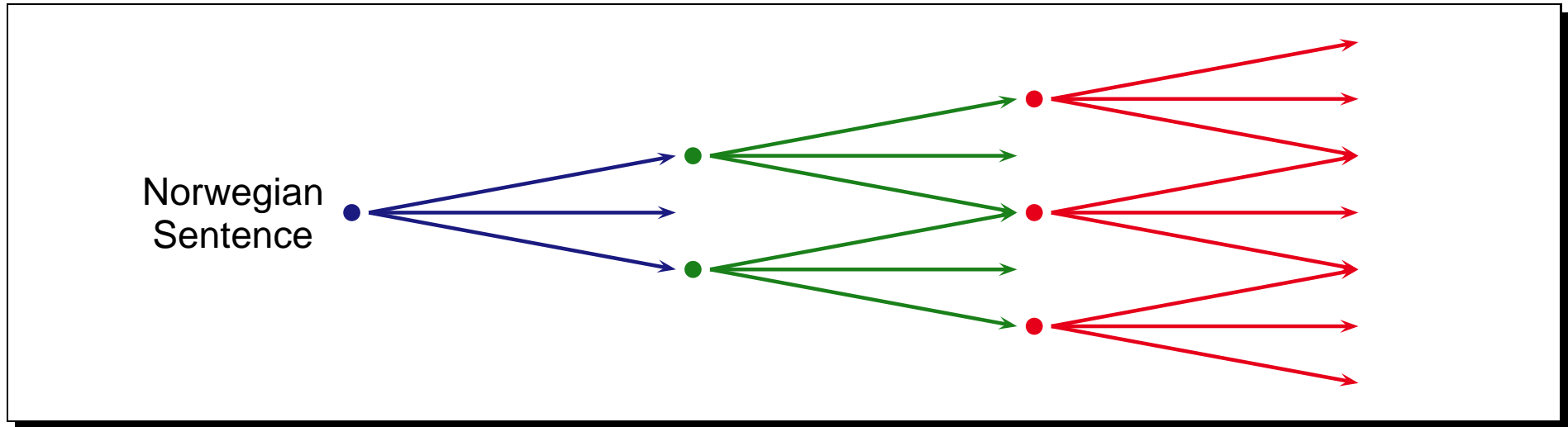


The Real Challenge — Language Ambiguity

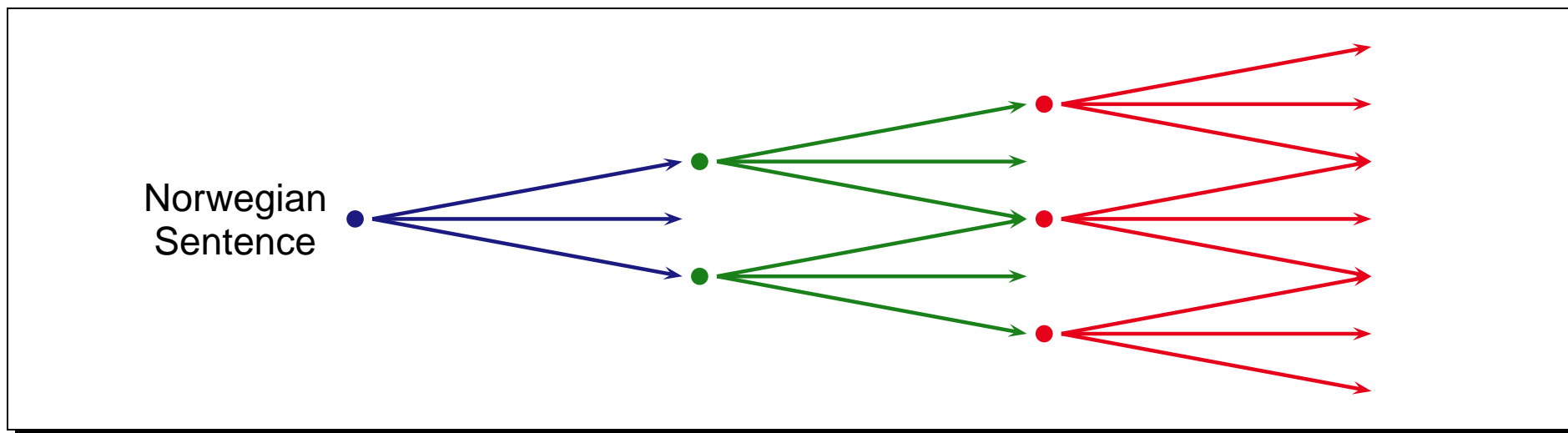
- < Den andre veien mot Bergen er kort. --- $12 \times 30 \times 25 = 25$
- > The other path towards Bergen is short. {0.58} (1:1:0).
- > The other road towards Bergen is short. {0.56} (1:0:0).
- > The second road towards Bergen is short. {0.55} (2:0:0).
- > That other path towards Bergen is a card. {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).
- ...



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;
- 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*

des

That

des

Google Translate

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

r Gjen-

r Gjen-



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;
- 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

- A **Partial coverage system potentially useful tool for translators.** 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).



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 %;
- + ann Software, Grammar Extensions, Transfer 'Ontology', et al. levels.



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.



Summary — Computational Linguistics Today

Some Lessons Learned

- Surprisingly hard problem: many unknowns in human language capacity;
 - statistical NLP can deliver robust, practical systems → limited scalability;
 - knowledge-based systems demand long-term development → re-usability;
 - limited-domain applications possible (e.g. BUSSTUC); too few end-to-end;
- empiricist vs. rationalist stand-off now largely reconciled: cross-fertilization.

Background Reading

http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

