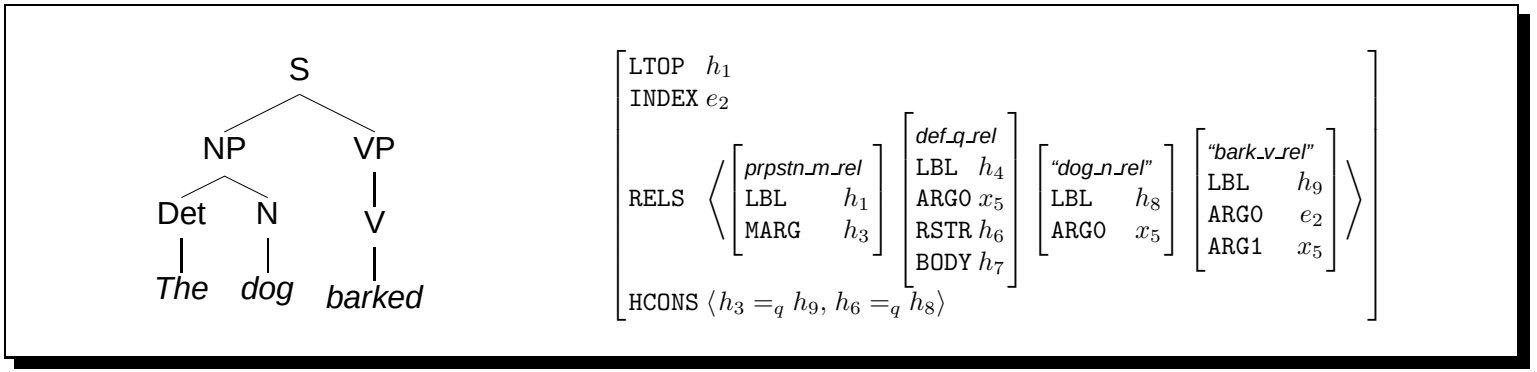


Computational Linguistics (INF2820 — Overview)

The Second Steep Road Against Bergen is a Card

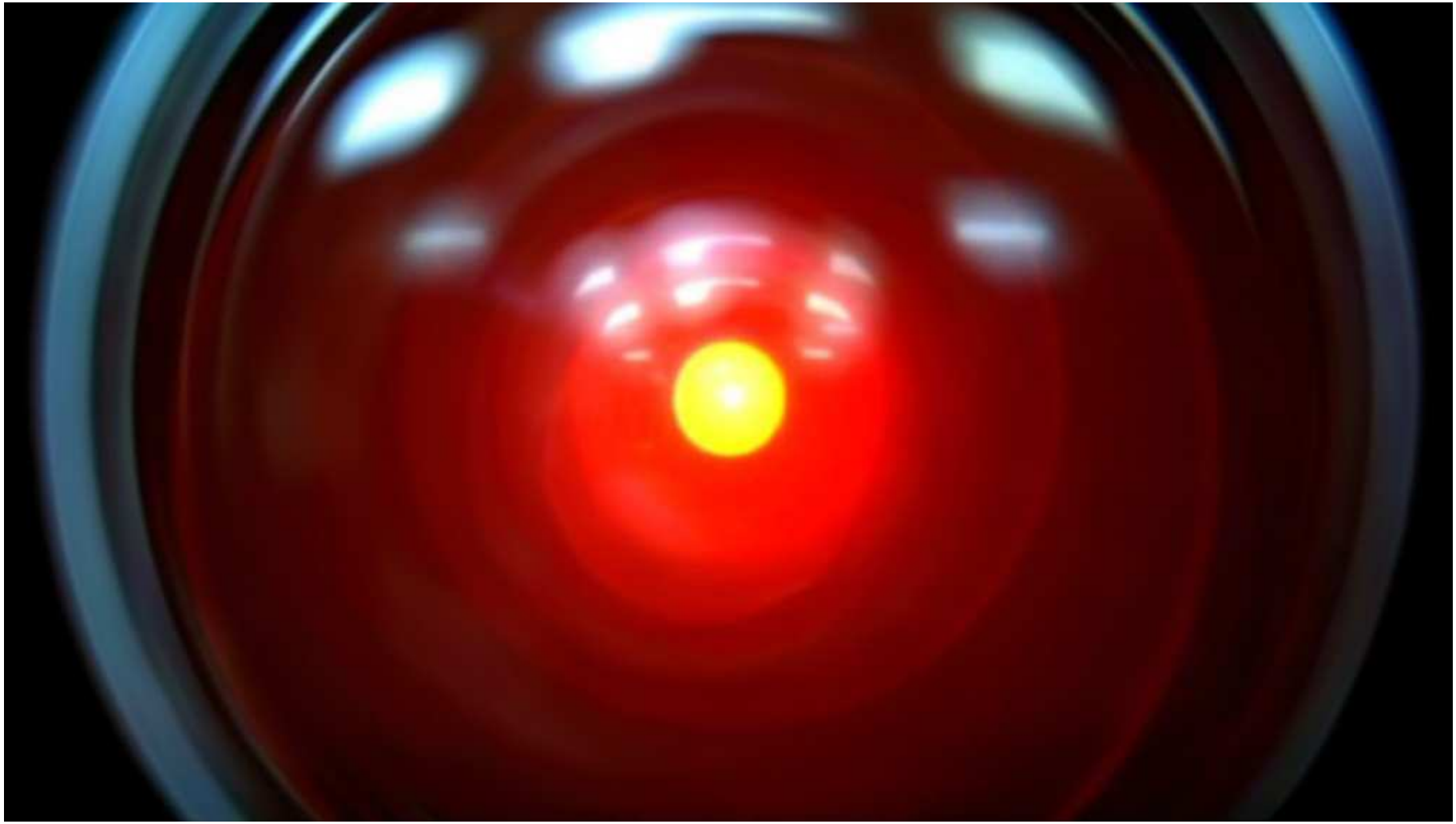
Stephan Oepen

Universitetet i Oslo & CSLI Stanford

oe@ifi.uio.no

So, What Actually is Computational Linguistics?

So, What Actually is Computational Linguistics?



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

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. (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. (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 Professor, 1990s)

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. (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 Professor, 1990s)

... a sub-discipline of our Artificial Intelligence programme.

(MIT Computer Science Professor, 1980s)

Yes, Great, But Why Should Anyone Care?

In the next three to five years, voice over IP and mobile devices [...] will become prevalent. [...] Desired technologies will soon replace menus and graphic user interfaces with natural-language interfaces. — People so much want to speak English to their computer. (Steve Ballmer, December 2005)

Yes, Great, But Why Should Anyone Care?

In the next three to five years, voice over IP and mobile devices [...] will become prevalent. [...] Desired technologies will soon replace menus and graphic user interfaces with natural-language interfaces. — People so much want to speak English to their computer. (Steve Ballmer, December 2005)

FRAMTIDSFORSKERNES DØDSLISTE [...] Datamaskinen vil mer og mer bli noe vi snakker med. Tastaturet vil nok ikke forsvinne helt, men vi vil definitivt bruke det mindre enn i dag. (Dagsavisen, January 2006)

Yes, Great, But Why Should Anyone Care?

In the next three to five years, voice over IP and mobile devices [...] will become prevalent. [...] Desired technologies will soon replace menus and graphic user interfaces with natural-language interfaces. — People so much want to speak English to their computer. (Steve Ballmer, December 2005)

FRAMTIDSFORSKERNES DØDSLISTE [...] Datamaskinen vil mer og mer bli noe vi snakker med. Tastaturet vil nok ikke forsvinne helt, men vi vil definitivt bruke det mindre enn i dag. (Dagsavisen, January 2006)

Computational Linguistics

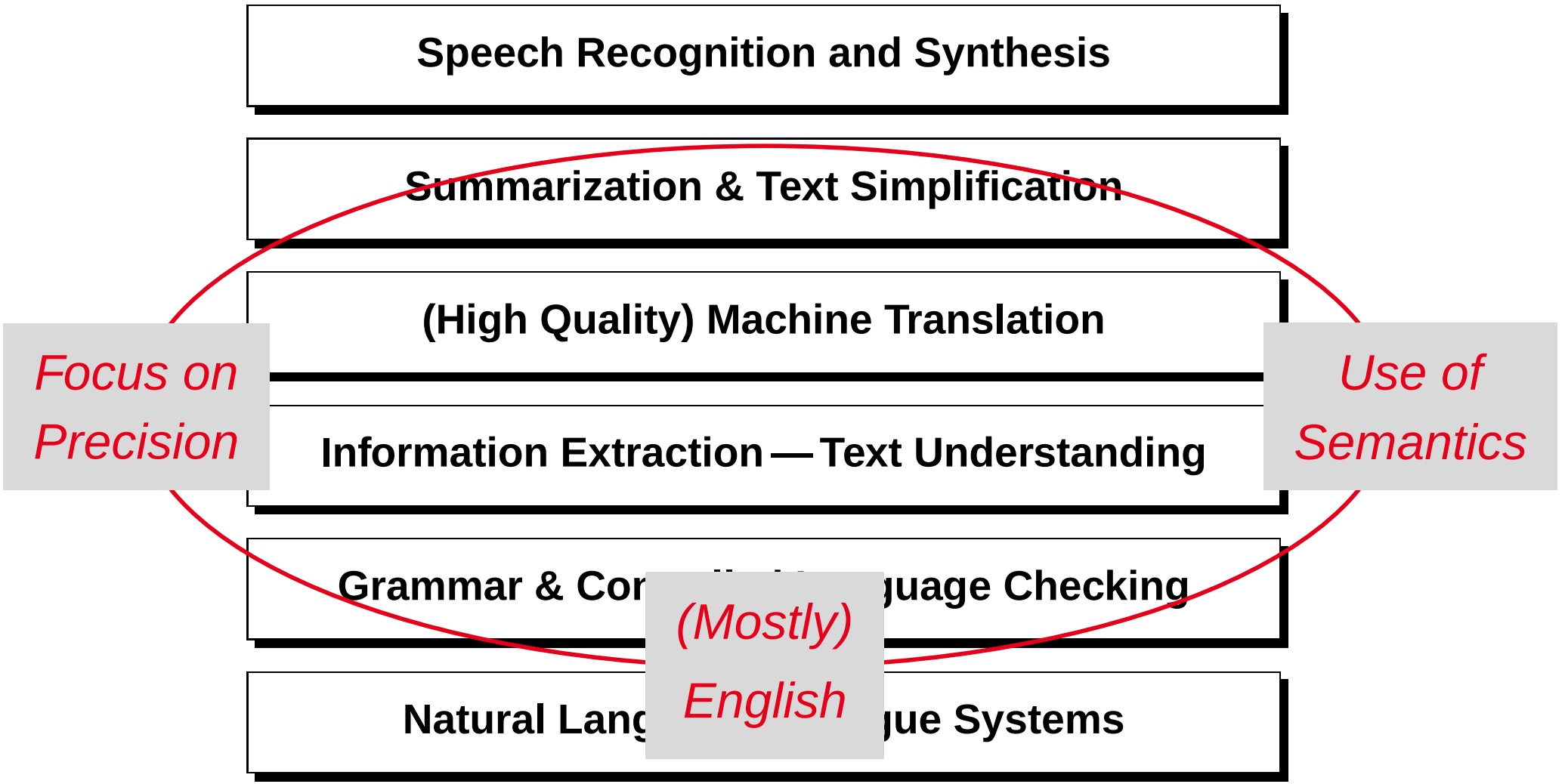
→ (young) interdisciplinary science: language, cognition, computation;

→ (once again) commercial growth potential due to 'knowledge society'.

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.

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

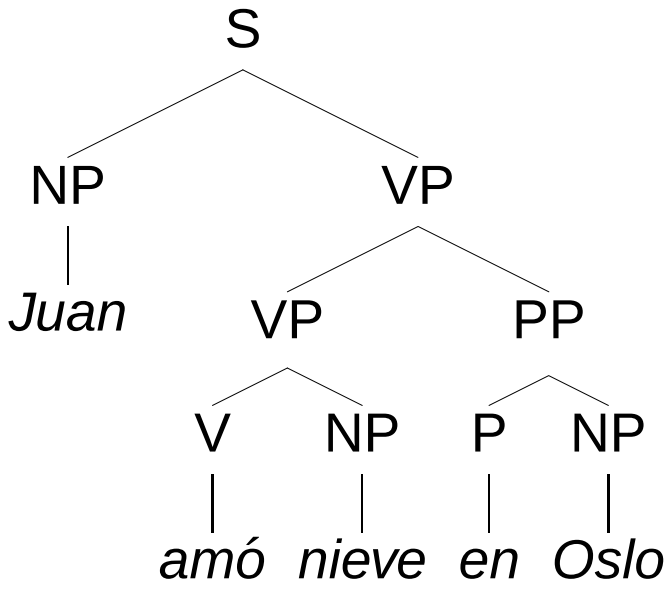
NP → “nieve”

NP → “Juan”

NP → “Oslo”

V → “amó”

P → “en”

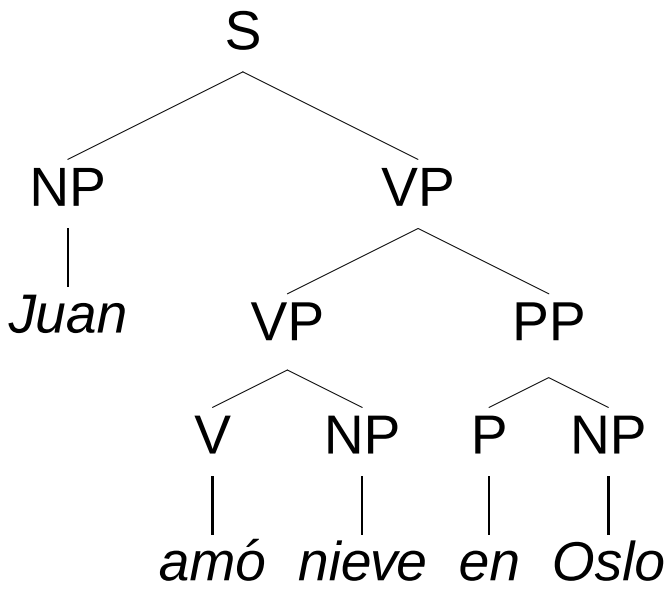


Juan amó nieve en Oslo

A Grossly Simplified Example

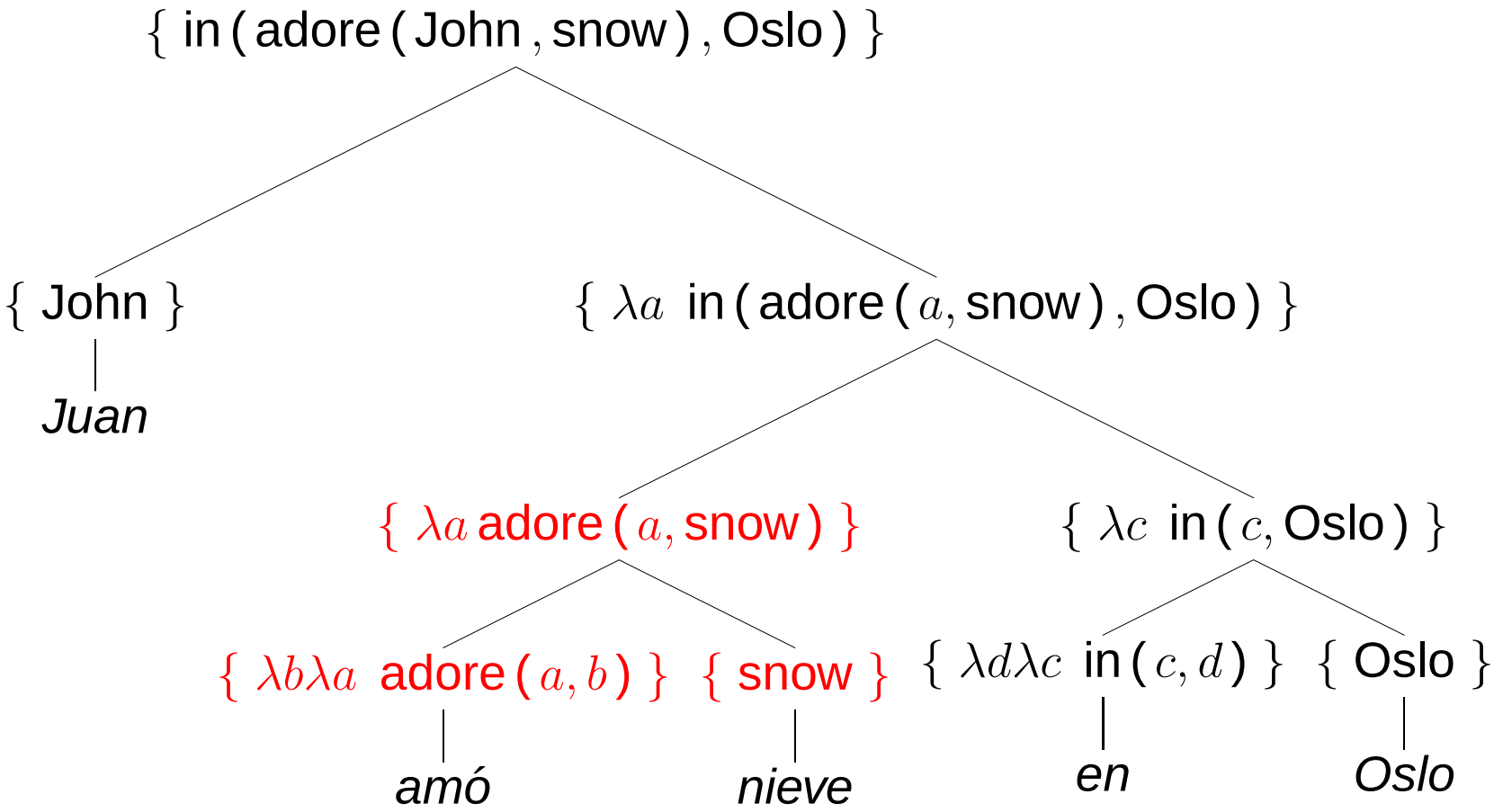
The Grammar of Spanish

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



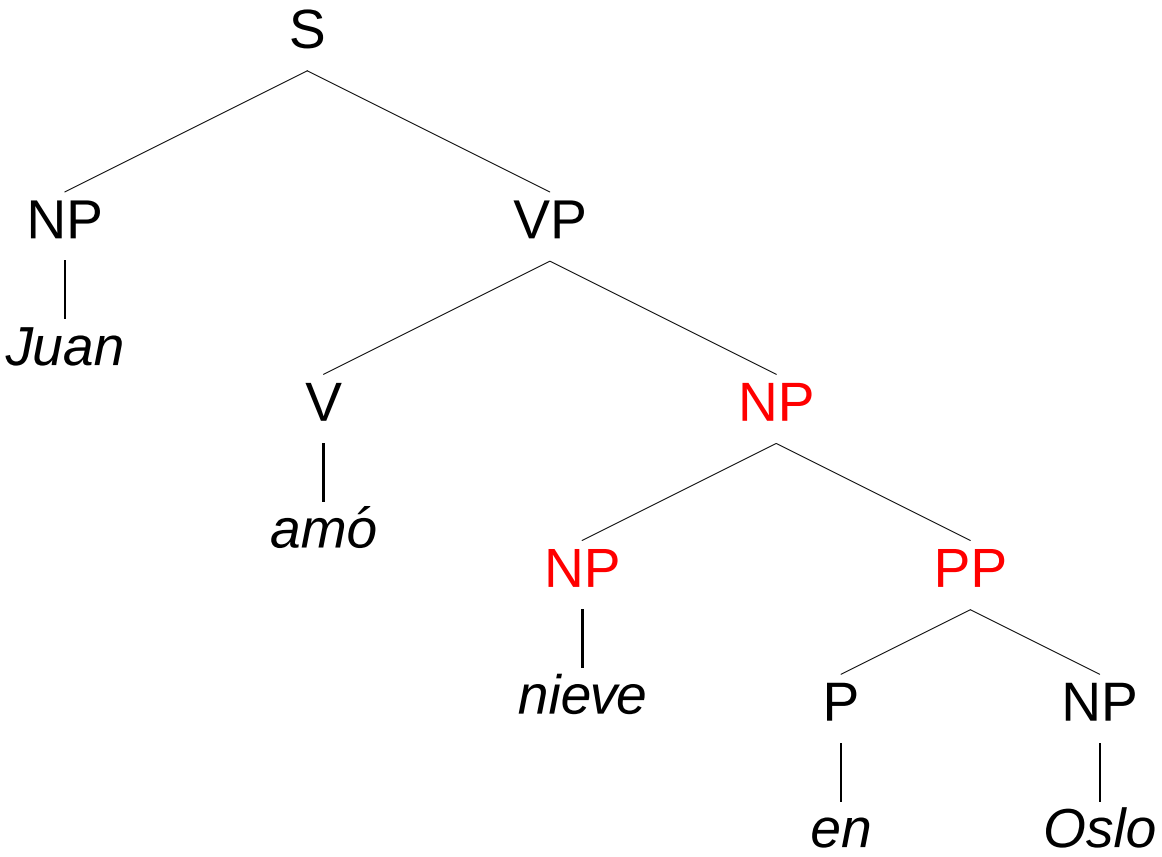
Juan amó nieve en Oslo

Meaning Composition (Grossly Simplified, Still)



VP → V NP { V (NP) }

Another Interpretation — Structural Ambiguity



NP → NP PP { PP (NP) }

An Outlook — Context-Free Grammars

- Formally, a *context-free grammar* (CFG) is a quadruple: $\langle C, \Sigma, P, S \rangle$
- C is the set of categories (aka *non-terminals*), e.g. $\{\text{S}, \text{NP}, \text{VP}, \text{V}\}$;
- Σ is the vocabulary (aka *terminals*), e.g. $\{\text{Juan}, \text{nieve}, \text{amó}, \text{en}\}$;
- P is a set of category rewrite rules (aka *productions*), e.g.

$$\begin{aligned} \text{S} &\rightarrow \text{NP VP} \\ \text{VP} &\rightarrow \text{V NP} \\ \text{NP} &\rightarrow \text{Juan} \\ \text{NP} &\rightarrow \text{nieve} \\ \text{V} &\rightarrow \text{amó} \end{aligned}$$

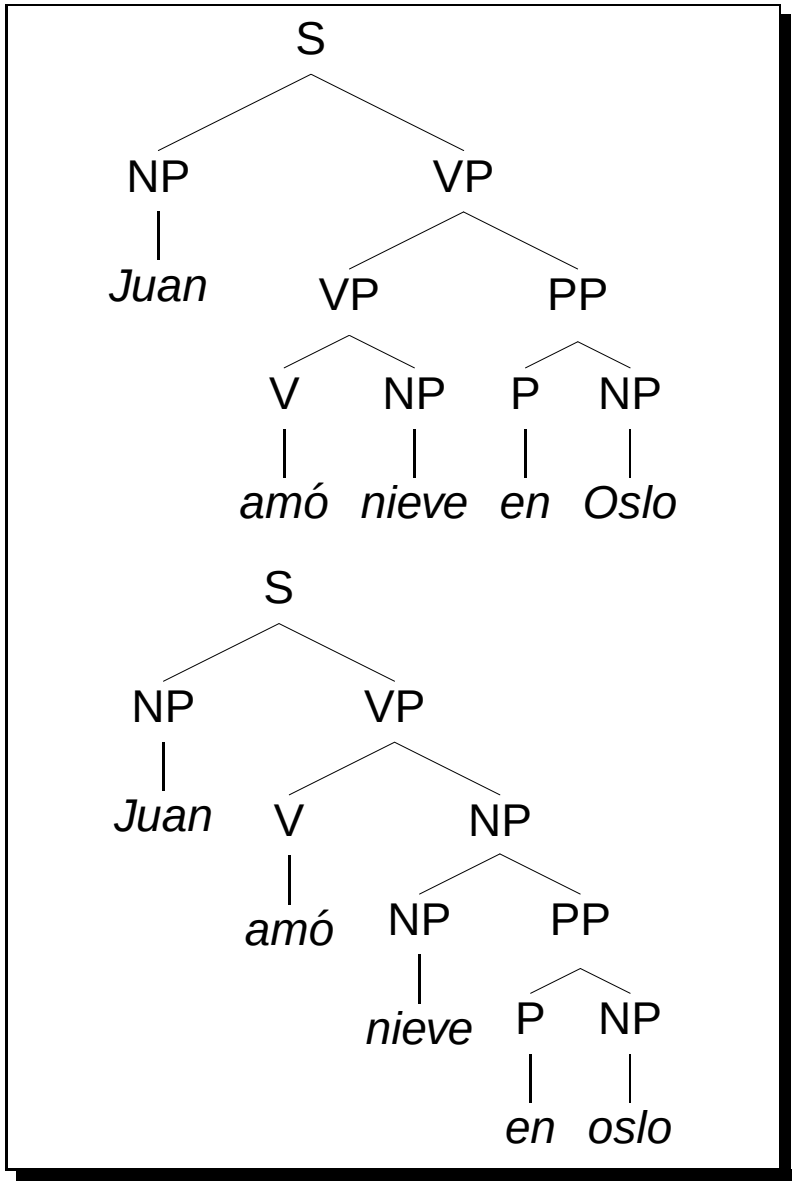
- $S \in C$ is the *start symbol*, a filter on complete ('sentential') results;
- for each rule '$\alpha \rightarrow \beta_1, \beta_2, ..., \beta_n$' $\in P$: $\alpha \in C$ and $\beta_i \in C \cup \Sigma$; $1 \leq i \leq n$.

Parsing: Recognizing the Language of a Grammar

$S \rightarrow NP\ VP$
$VP \rightarrow V\ NP$
$VP \rightarrow VP\ PP$
$NP \rightarrow NP\ PP$
$PP \rightarrow P\ NP$
$NP \rightarrow Juan \mid nieve \mid Oslo$
$V \rightarrow amó$
$P \rightarrow en$

All Complete Derivations

- are rooted in the start symbol S;
- label internal nodes with categories $\in C$, leafs with words $\in \Sigma$;
- instantiate a grammar rule $\in P$ at each local subtree of depth one.



More, and More, and More Ambiguity

Speech Recognition

its		*hard*	*to*	*wreck*	*a*	*nice*	*beach*
it	*'s*	*hard*	*to*	*recognize*			*speech*

Morphology

- *fisker* *fisk* $_{\mathsf{N}}$ + plural vs. *fiske* $_{\mathsf{V}}$ + present vs. *fisker* $_{\mathsf{N}}$ + singular;
- *brus-automat* vs. *bru-sau-tomat*; *vinduene* vs. *vin-duene*; et al.

Semantics

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

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.***

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.

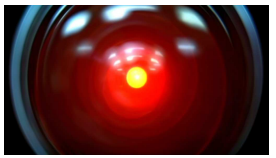
[Fred Jelinek, 1980s]

Competition of Paradigms

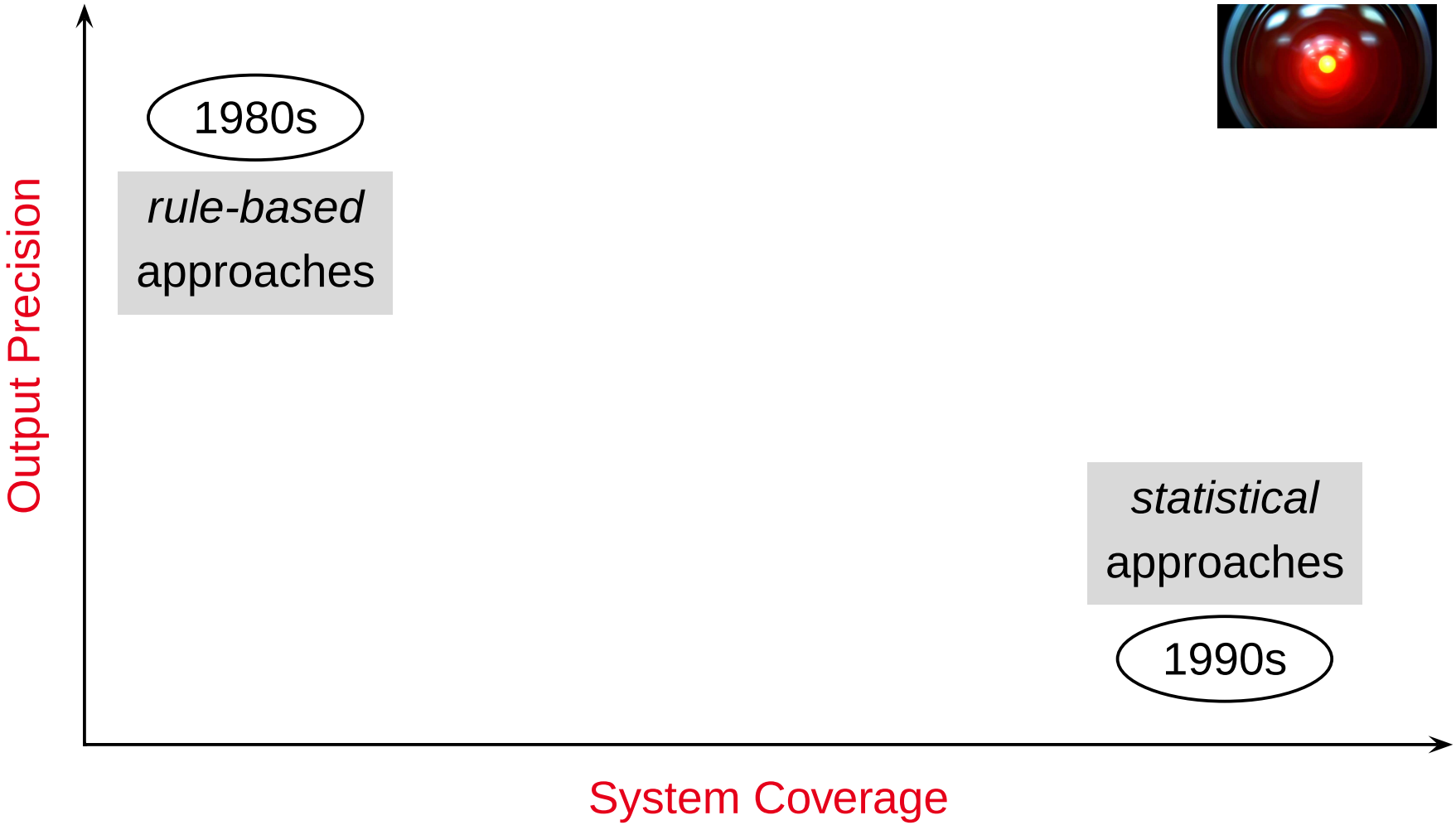
- Rationalist: formally encode linguistic and extra-linguistic knowledge;
- empiricist: statistical models trained on distributional data (corpora);
- Jelinek eventually turned off the lights — grammar research stable;

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

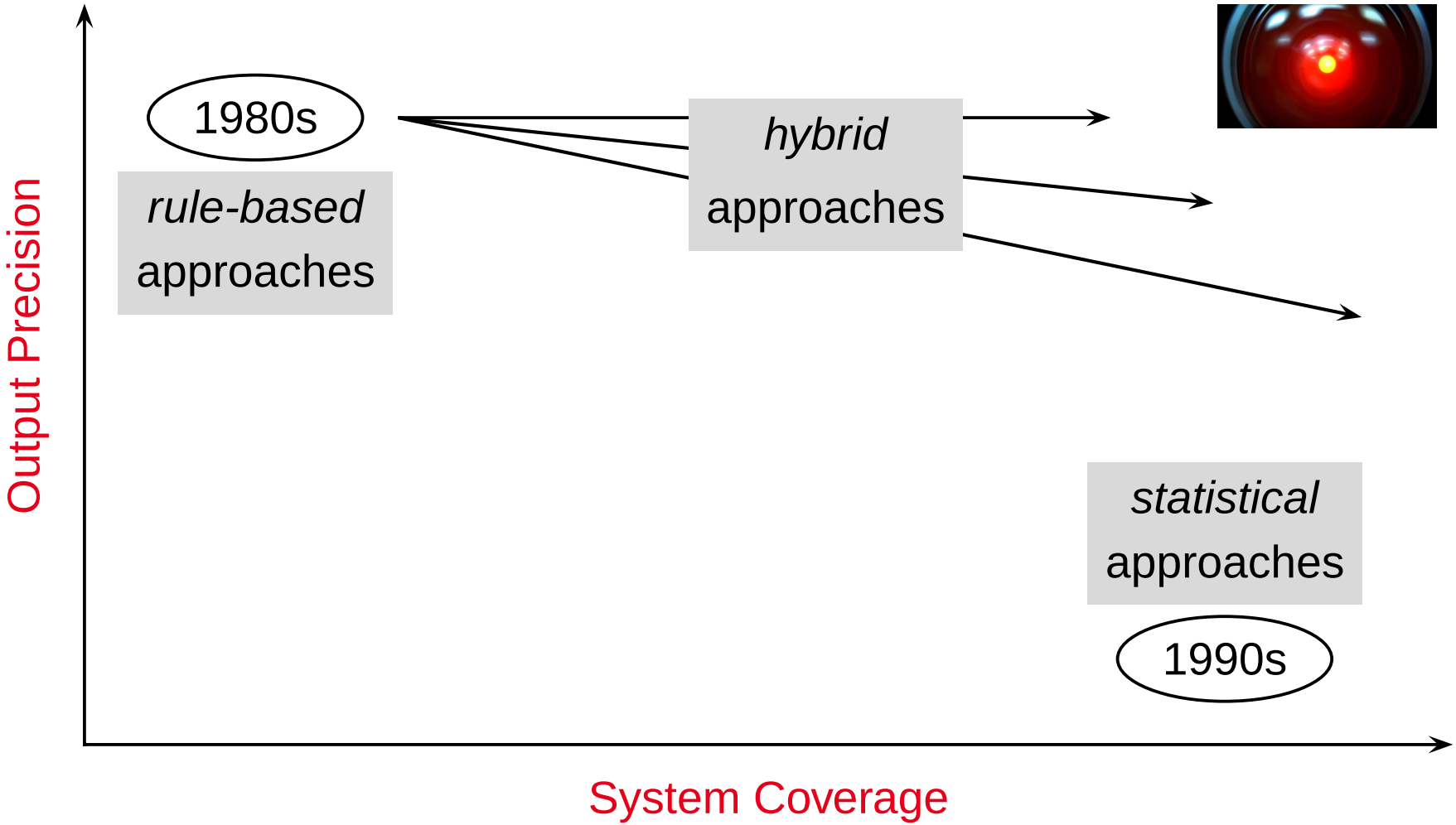
The Holy Grail: Balancing Coverage and Precision



The Holy Grail: Balancing Coverage and Precision



The Holy Grail: Balancing Coverage and Precision



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`

INF2880 — What We Are About to Do (and Why)

Course Outline

- Extend understanding of (natural) language as a system of rules;
- learn how to *formalize* grammars through typed feature structures;
- design and implement common algorithms and probabilistic models;
- solve regular exercises: immediate gratification (risk of late hours).

Three Interacting Components

- **grammar engineering** formalize linguistic theories with complex interactions of multiple phenomena; implementation and debugging;
- **processing** understand common parsing algorithms; unification of feature structures; implement an efficient unification-based parser;
- **probabilistic models** capture relative frequency of (competing) phenomena; disambiguation: pick analysis with highest probability.

Grammar Engineering from a CS Perspective

Implementation Goals

- Translate linguistic constraints into specific formalism → formal model;
- computational grammar provides mapping between form and meaning;
- assign correct analyses to grammatical, reject ungrammatical inputs;
- parsing and generation algorithms: apply mapping in either direction.

Analogy to (Object-Oriented) Programming

- Computational system with observable behavior: immediately testable;
- typed feature structures as a specialized (OO) programming language;
- make sure that all the pieces fit together; revise – test – revise – test ...

Why Common-Lisp for Implementation Exercises?

- Arguably most widely used language for ‘symbolic’ computation;
- easy to learn: extremely simple syntax; straightforward semantics;
- a rich language: multitude of built-in data types and operations;
- full standardization; Common-Lisp has been stable for a decade;
- LKB (experimentation environment) implemented in Common-Lisp;

$\rightarrow$ for our purposes, (at least) as good a choice as any other language.

$$n! \equiv \begin{cases} 1 & \text{for } n = 0 \\ n \times (n-1)! & \text{for } n > 0 \end{cases}$$

```
(defun ! (n)
  (if (= n 0)
    1
    (* n (! (- n 1)))))
```

Comments on Background Literature

Formal Grammar and General NLP

- Sag, Ivan A. Tom Wasow, and Emily M. Bender: *Syntactic Theory. A Formal Introduction (2nd Edition).* Stanford, CA: CSLI Publications (2003);
- Jurafsky, Daniel and Martin, James H.: *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd Edition).* Upper Saddle River, NJ: Prentice Hall (2008).

The Linguistic Knowledge Builder

- Copestake, Ann: *Implementing Typed Feature Structure Grammars.* Stanford, CA: CSLI Publications (2001).