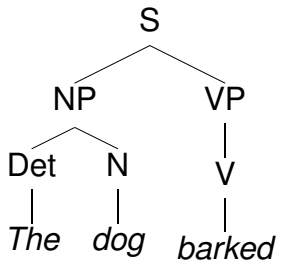


$$\begin{bmatrix} \texttt{LTOP} & h_1 \\ \texttt{INDEX} & e_2 \\ \texttt{RELS} & \left\langle \begin{bmatrix} \textit{prpstn\_m\_rel} \\ \texttt{LBL} & h_1 \\ \texttt{MARG} & h_3 \end{bmatrix} \begin{bmatrix} \textit{def\_q\_rel} \\ \texttt{LBL} & h_4 \\ \texttt{ARG0} & x_5 \\ \texttt{RSTR} & h_6 \\ \texttt{BODY} & h_7 \end{bmatrix} \begin{bmatrix} \textit{“dog\_n\_rel”} \\ \texttt{LBL} & h_8 \\ \texttt{ARG0} & x_5 \end{bmatrix} \begin{bmatrix} \textit{“bark\_v\_rel”} \\ \texttt{LBL} & h_9 \\ \texttt{ARG0} & e_2 \\ \texttt{ARG1} & x_5 \end{bmatrix} \right\rangle \\ \texttt{HCONS} & \langle h_3 =_q h_9,\ h_6 =_q h_8 \rangle \end{bmatrix}$$

An Overview of Natural Language Understanding

(In Three Quarters of an Hour)

Stephan Oepen

Universitetet i Oslo & CSLI Stanford

oe@ifi.uio.no

So, What Exactly is Natural Language Understanding?

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

So, What Exactly is Natural Language Understanding?

... 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. (Marketing Blurb, 2000)

So, What Exactly is Natural Language Understanding?

... 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. (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, 1980s)

So, What Exactly is Natural Language Understanding?

... 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. (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, 1980s)

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

(MIT CS Professor, 1970s)

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)

Natural Language Understanding

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

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

Some Traditional Applications of NLU

Machine Translation

- Traditional: analyse source to some degree, transfer, generate target.

Information Extraction & Text 'Understanding'

- Email auto- (or assisted) response: interpret customer requests;
- Semantic Web: annotate WWW with structured, conceptual data.

(Spoken) Dialogue Systems

Grammar & Controlled Language Checking

Summarization & Text Simplification

What Makes Natural Language a Hard Problem?

```
|< |Den andre veien mot Bergen er kort.| --- 16 x 52 x 112 = 112
|> |That other path against Bergen is short.| [0.70] <0.03> (0:1:0).
|> |That other path towards Bergen is short.| [0.70] <0.03> (0:0:0).
|> |That second path against Bergen is short.| [0.65] <0.03> (2:1:0).
|> |That second path towards Bergen is short.| [0.65] <0.03> (2:0:0).
|> |That other road against Bergen is short.| [0.62] <0.03> (0:3:0).
|> |That other road towards Bergen is short.| [0.62] <0.03> (0:2:0).
...
|> |The second path against Bergen is short.| [0.18] <0.03> (3:1:0).
|> |The second path towards Bergen is short.| [0.18] <0.03> (3:0:0).
|> |That second path against Bergen is a card.| [0.17] <0.02> (8:1:0).
|> |That second path towards Bergen is a card.| [0.17] <0.02> (8:0:0).
|> |That other path against Bergen is cards.| [0.17] <0.03> (5:1:0).
|> |That other path towards Bergen is cards.| [0.17] <0.03> (5:0:0).
...
|> |Short is that other road, against Bergen.| [-0.37] <0.03> (0:3:2).
|> |Short is that second road, towards Bergen.| [-0.42] <0.03> (2:2:2).
```

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

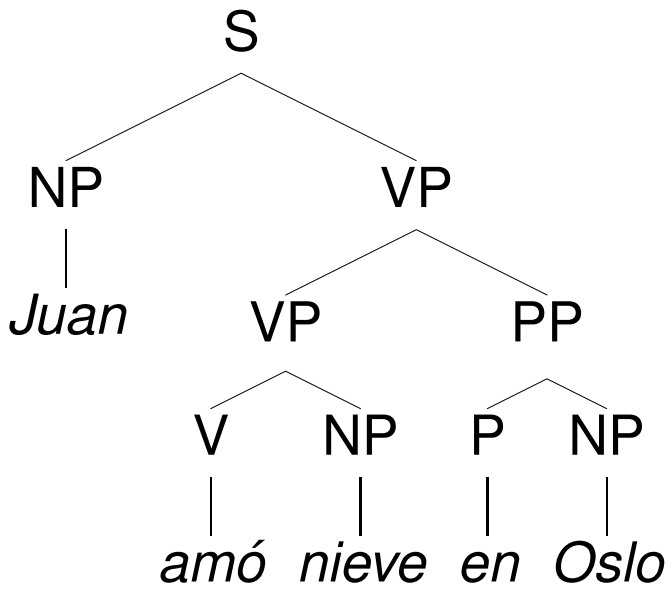
NP → “nieve”

NP → “Juan”

NP → “Oslo”

V → “amó”

P → “en”

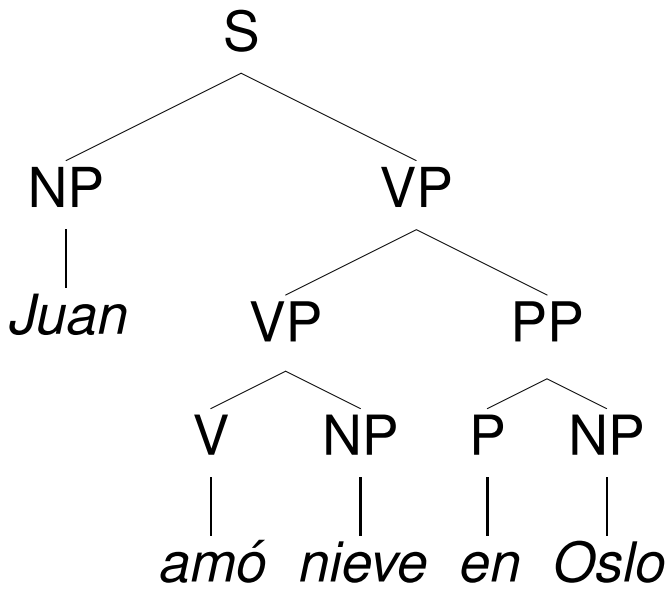


Juan amó nieve en Oslo

A Grossly Simplified Example

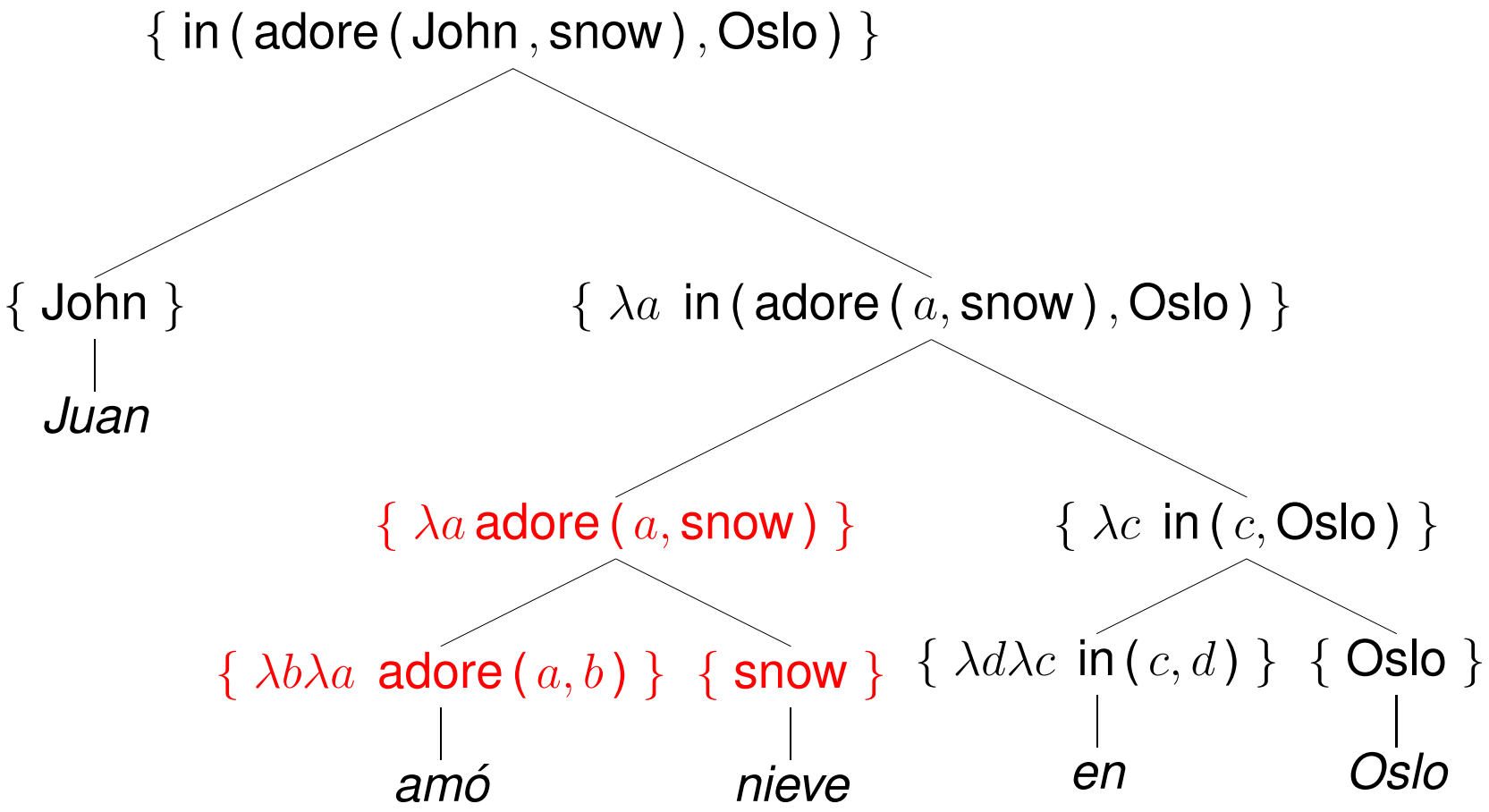
The Grammar of Spanish

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



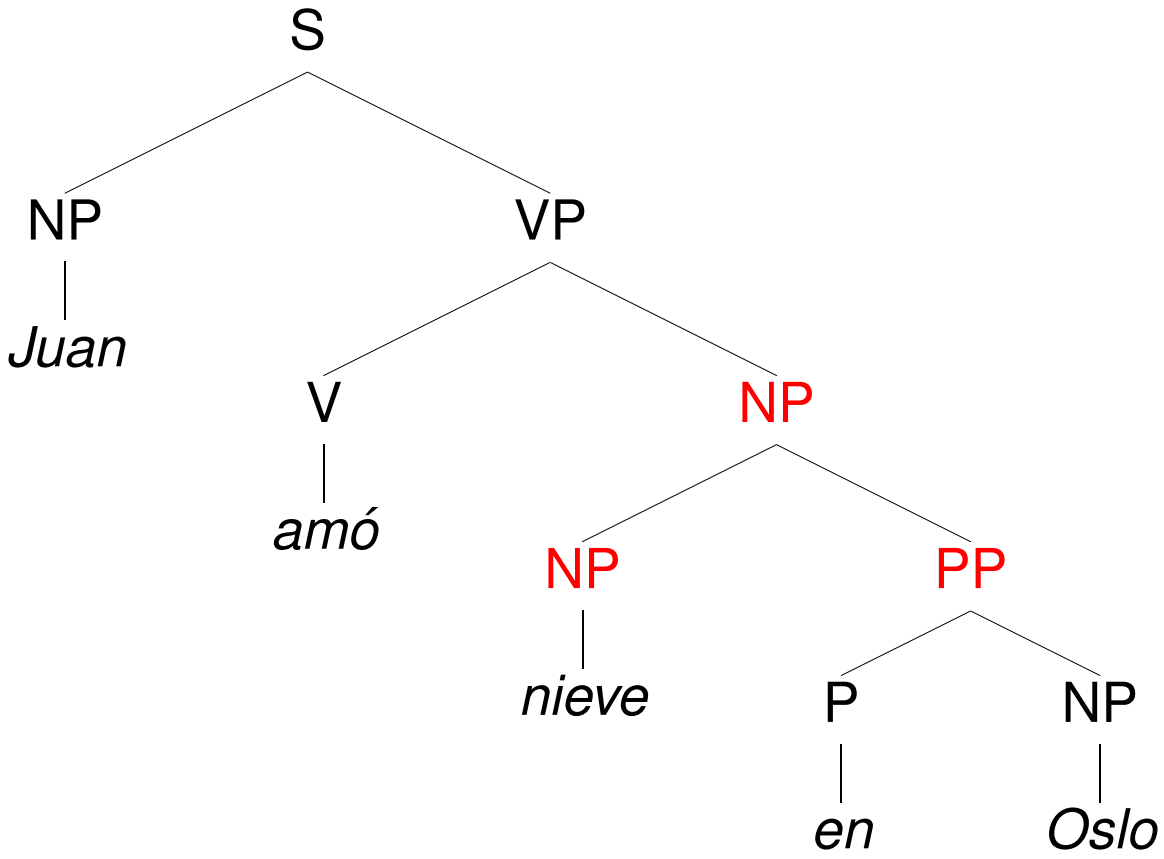
Juan amó nieve en Oslo

Meaning Composition (Grossly Simplified, Still)



VP → V NP { V (NP) }

Another Interpretation — Structural Ambiguity



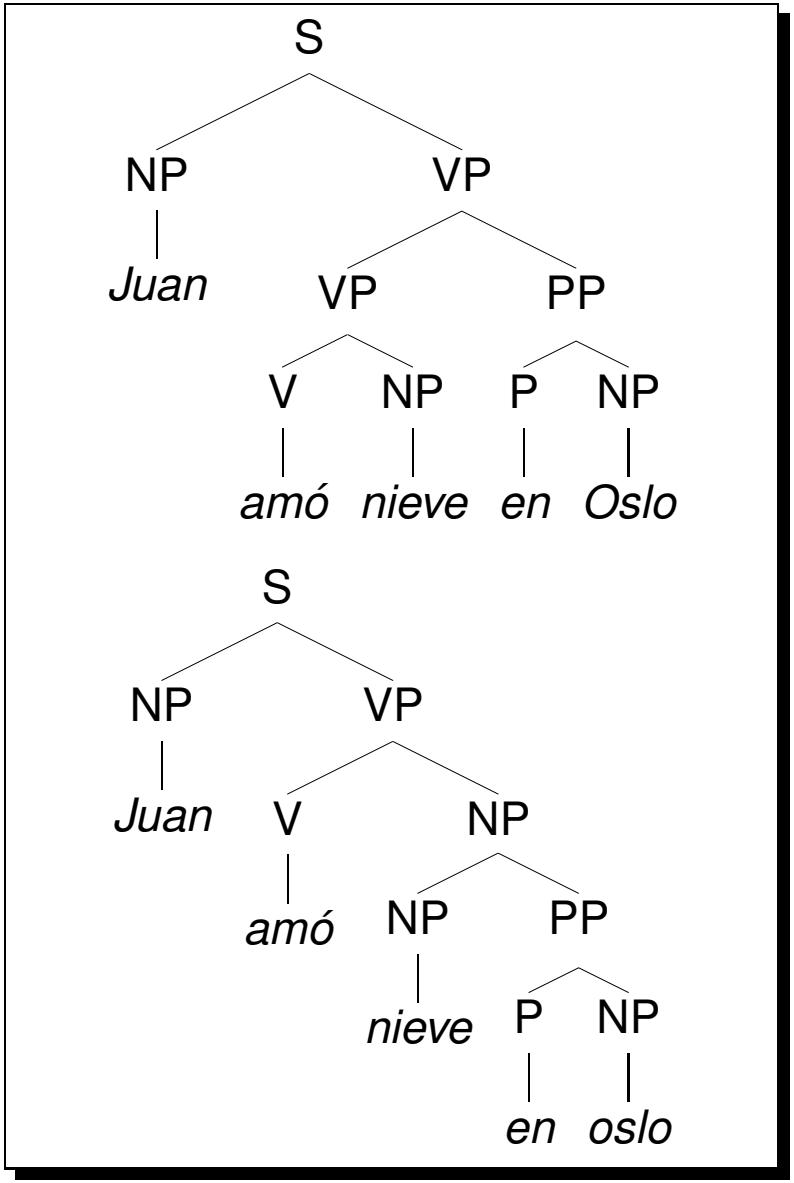
$$\text{NP} \rightarrow \text{NP PP} \quad \{\text{PP}(\text{NP})\}$$

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.



Review: The Chomsky Hierarchy of Languages

- (Formal) Languages vary in 'degree of structural complexity' exhibited;
- traditionally: a^* (iteration) vs. $a^n b^n$ (nesting) vs. $a^n b^n c^n$ ('cross-serial');
- Chomsky Hierarchy: inclusion classes of formal languages; Type 0 – 3.

0	unrestricted	$\beta_1 \longrightarrow \beta_2$	Turing Machine
1	context-sensitive	$\beta_1 \alpha \beta_2 \longrightarrow \beta_1 \gamma \beta_2$	linearly-bounded automaton
2	context-free	$\alpha \longrightarrow \beta$	push-down automaton
3	regular	$\alpha \longrightarrow \delta\alpha \mid \alpha\delta$	finite-state automaton
$\alpha \in C,\ \beta_i \in (C \cup \Sigma)^*,\ \gamma \in (C \cup \Sigma)^+,\ \delta \in \Sigma^+$			

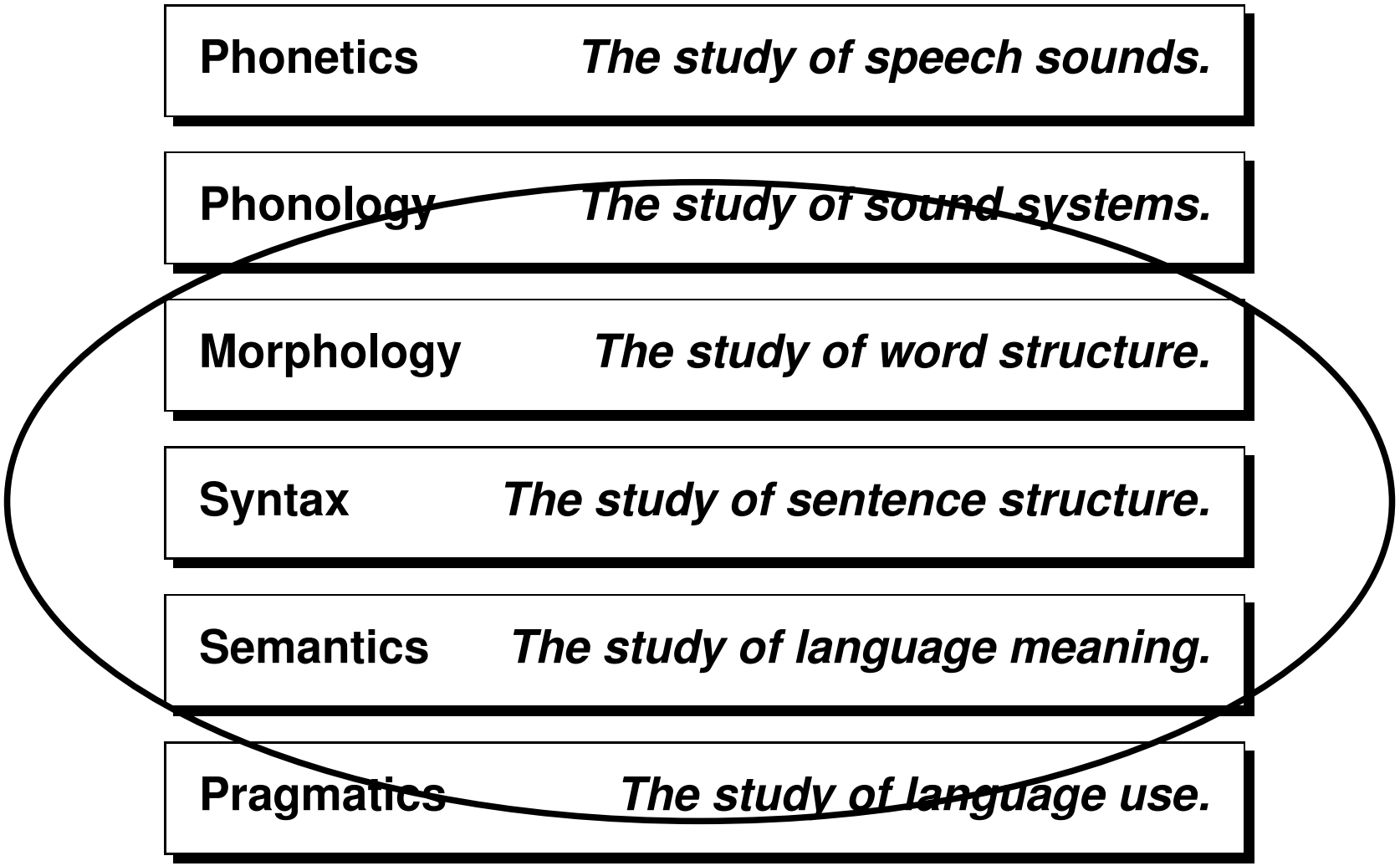
What is the Formal Complexity of Natural Languages?

- Minimally context-free (center self-embedding, e.g. in relative clauses);
- (Culy; Shieber, 1985): *not* context-free (Bambara, Swiss German);
- (Joshi, 1985): extra class of *mildly* context-sensitive languages (TAG).

Some Areas of Descriptive Grammar

Phonetics	***The study of speech sounds.***
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



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. $\forall n \exists l_1, l_2 \ldots$

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 approximate human language competence;
- Jelinek eventually turned off the lights — LFG & HPSG groups stable;

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

Competing Approaches (1 of 2)

Can you send me copies of all checks in December?

Statistical Part-of-Speech Tagging (96.7 % Accuracy)

1.0	1.0	0.98	1.0	1.0	1.0	1.0	0.93	1.0	1.0	1.0
MD	PRP	VB	PRP	NNS	IN	DT	NNS	IN	NNP	.
Can	you	send	me	copies	of	all	checks	in	December	?

Text Classification ($\sim$ 85 % Accuracy)

CheckCopyRequest 0.6934, *CheckBookRequest* 0.0247,
StatementCopyRequest 0.0066, ...

Competing Approaches (2 of 2)

$\langle\, h_1,$
$\{\, h_1$:int_m(h_2), h_3:_can_v_modal(e_4, h_5), h_7:_send_v(e_8, x_9, x_{10}, x_{11}),
h_{12}:pronoun_q(x_9, h_{13}, h_{14}), h_{15}:pron(x_9 {2nd}),
h_{16}:pronoun_q(x_{10}, h_{17}, h_{18}), h_{19}:pron(x_{10} {1sg}),
h_{20}:bare_q(x_{11}, h_{21}, h_{22}), h_{23}:_copy_n_of(x_{11} {pl}, x_{24}),
h_{25}:_all_q(x_{24}, h_{26}, h_{27}), h_{28}:_check_n(x_{24} {pl}),
h_{28}:temp_loc(_, x_{24}, x_{29}), h_{30}:proper_q(x_{29}, h_{31}, h_{32}), h_{33}:mofy(x_{29}, "DEC") $\}$,
$\{\, h_2 =_q h_3,\ h_5 =_q h_7,\ h_{13} =_q h_{15},\ h_{17} =_q h_{19},\ h_{21} =_q h_{23},\ h_{26} =_q h_{28},\ h_{31} =_q h_{33} \,\}\,\rangle$

(Truth-Conditional or) Logical-Form Semantics

+ high-level abstraction; grounded in entities and relations → inference;

− very difficult to construct (correctly, with broad-coverage) and process.

Putting Things Together: Language and the World

Discourse and Pragmatics

- h_{15}:pron(x_9 { 2nd }) $\rightarrow$ email recipient; h_{19}:pron(x_{10} { 1sg }) $\rightarrow$ email sender;
- h_{28}:temp_loc(_, x_{24}, x_{29}), h_{33}:mofy(x_{29}, "DEC") $\rightarrow$ 2005 (but maybe 2006);
- h_1:int_m(h_2), h_3:_can_v_modal(e_4, h_5), h_7:_send_v(e_8, x_9, x_{10}, x_{11}) $\rightarrow$ request.

World Knowledge (Plus Back-End Databases)

- 'all checks in December 2005'
  $\rightarrow \{ x \mid x \; isa$ check $\wedge$ 20051201 $\leq$ date(x) $\leq$ 20051231 $\}$
- request h_7:_send_v(e_8, x_9, x_{10}, x_{11}), h_{23}:_copy_n_of(x_{11}, x_{24}), h_{28}:_check_n(x_{24})
  $\rightarrow$ `<CheckCopyRequest from="26046712345" ...> ... </>`

Summary — Natural Language Understanding 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

- general: `http://www.coli.uni-saarland.de/~hansu/what_is_cl.html`;
- Jurafsky, Daniel and Martin, James H.: *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ (2000).

Based on Research and Contributions of

Dan Flickinger

Tim Baldwin, John Beavers
Emily M. Bender, John Carroll, Ann Copestake,
Jan Tore Lønning, Rob Malouf, Ivan A. Sag,
Hans Uszkoreit, Tom Wasow, and others.