INF5820

Natural Language Processing - NLP

H2009

Jan Tore Lønning

jtl@ifi.uio.no

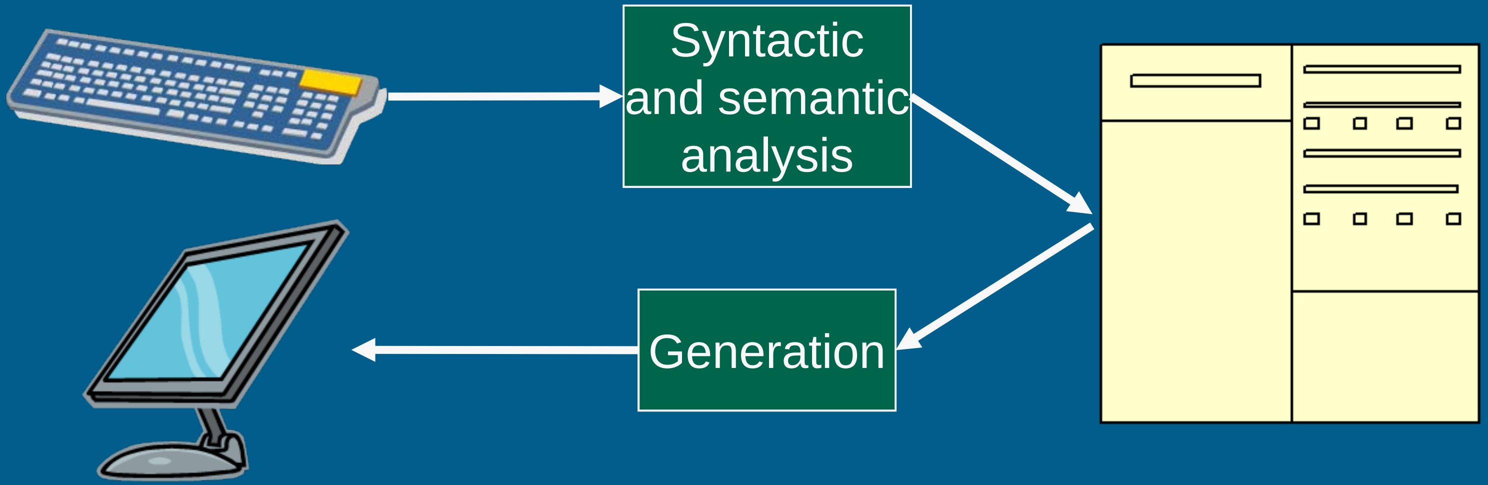
Today

- Overiew: course content
- Practicalities
- Beginning tagging

NLP applications - examples

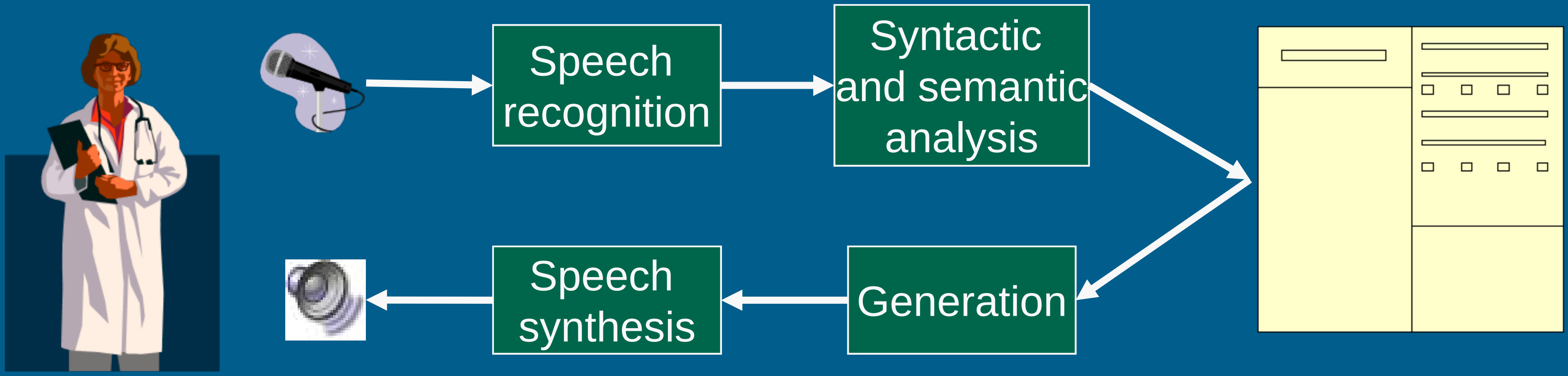
1. General:
   1. Translation
   2. Dialogue
   3. Information processing
2. Speech
   1. Speech ←→ text
   2. Voice control
3. Language support

Communicating with the computer



- The model of the computer as communicatior:
  - Analysis
  - Process
  - Generate/synthesis

Oral communication



Speech recognition

Syntactic and semantic analysis

Generation

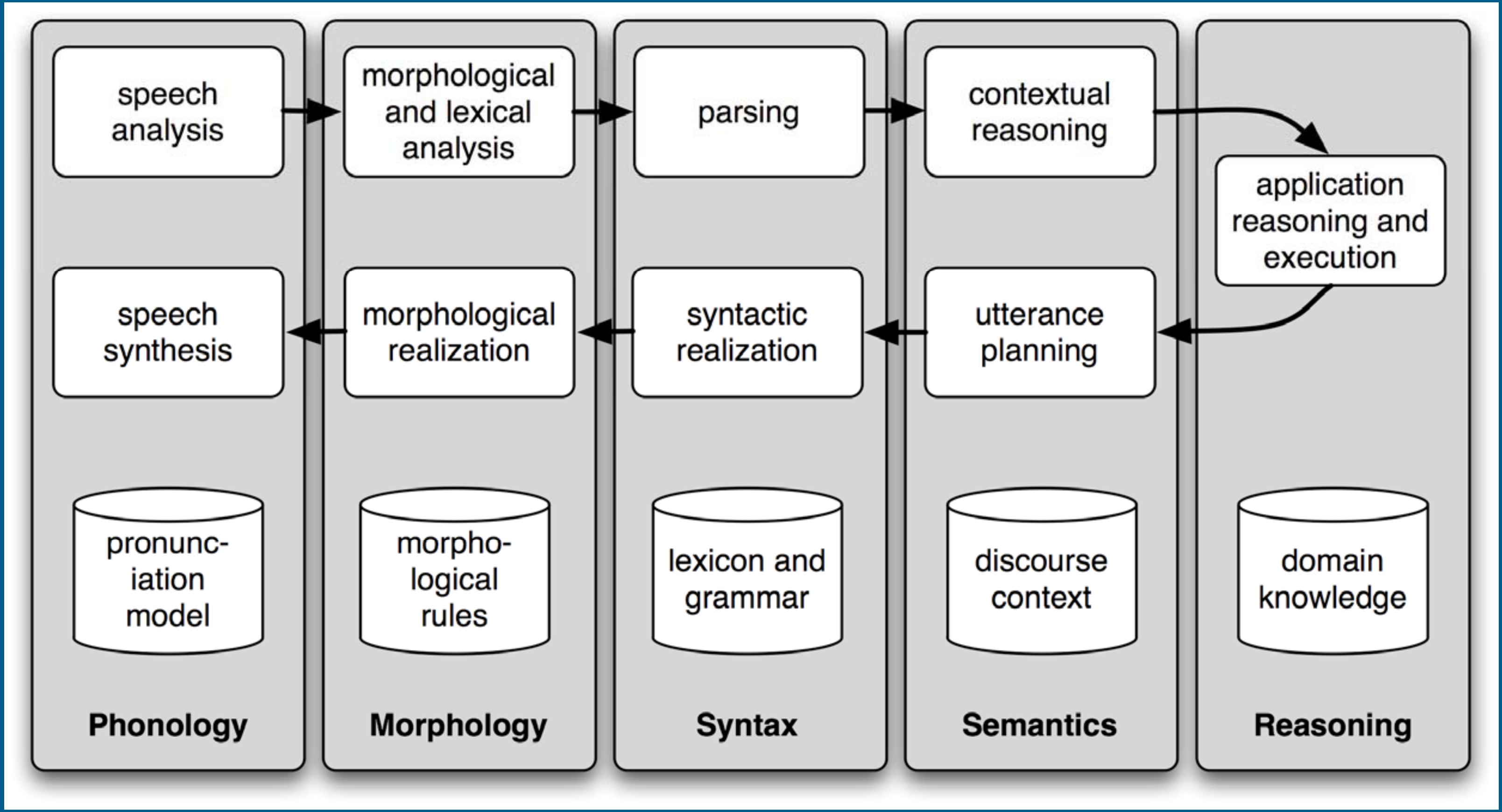
Speech synthesis

- The model of the computer as communicatior:
  - Analysis: speech, grammar, semantics, pragmatics
  - Process
  - Generate/synthesis: content, grammar, speech

The communicating computer

- This model fits many applications
  - Translation
  - Dialogue
  - Information processing
  - (with or without speech)
- The processing step varies:
  - Translation
  - Find an answer
  - Carry out an order

From NLTK


speech analysis
morphological and lexical analysis
parsing
contextual reasoning
application reasoning and execution
speech synthesis
morphological realization
syntactic realization
utterance planning
pronunc-iation model
morpho-logical rules
lexicon and grammar
discourse context
domain knowledge
Phonology
Morphology
Syntax
Semantics
Reasoning

Analysis: two approaches

- Theoretical, formal
  - Build a declarative model using
    - Linguistics
    - Logic
  - Algorithms
  - How does it fit data?
- Empirical
  - Start with naturally occurring text
  - What information can we get?

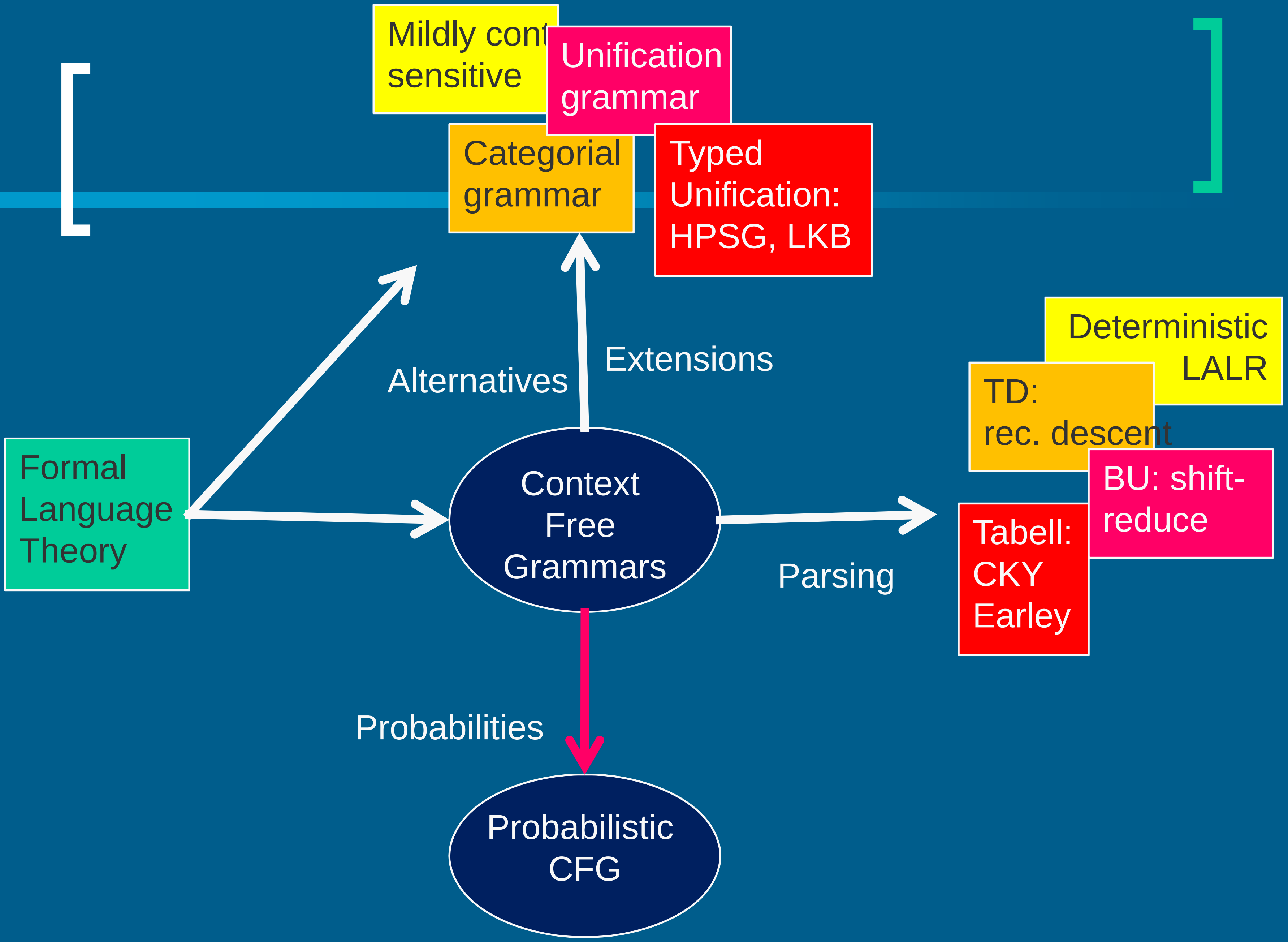
Grammars (formal approach)

Context Free Phrase-Structure Grammar (CF P-SG)

S → NP VP
NP → DET N
VP → IV
VP → TV NP
NP → NP som VP
NP → NP PP
PP → P NP
NP → kari | ola
N → barn | by | mann

BNF (Backus-Naur Form)

S ::= NP VP
NP ::= DET N | NP som VP |
NP PP | kari | ola
VP ::= IV | TV NP
PP ::= P NP
N ::= barn | by | mann



Formal Language Theory

Alternatives

Mildly cont sensitive

Unification grammar

Categorial grammar

Typed Unification: HPSG, LKB

Extensions

Context Free Grammars

Parsing

Deterministic LALR

TD: rec. descent

BU: shift-reduce

Tabell: CKY Earley

Probabilities

Probabilistic CFG

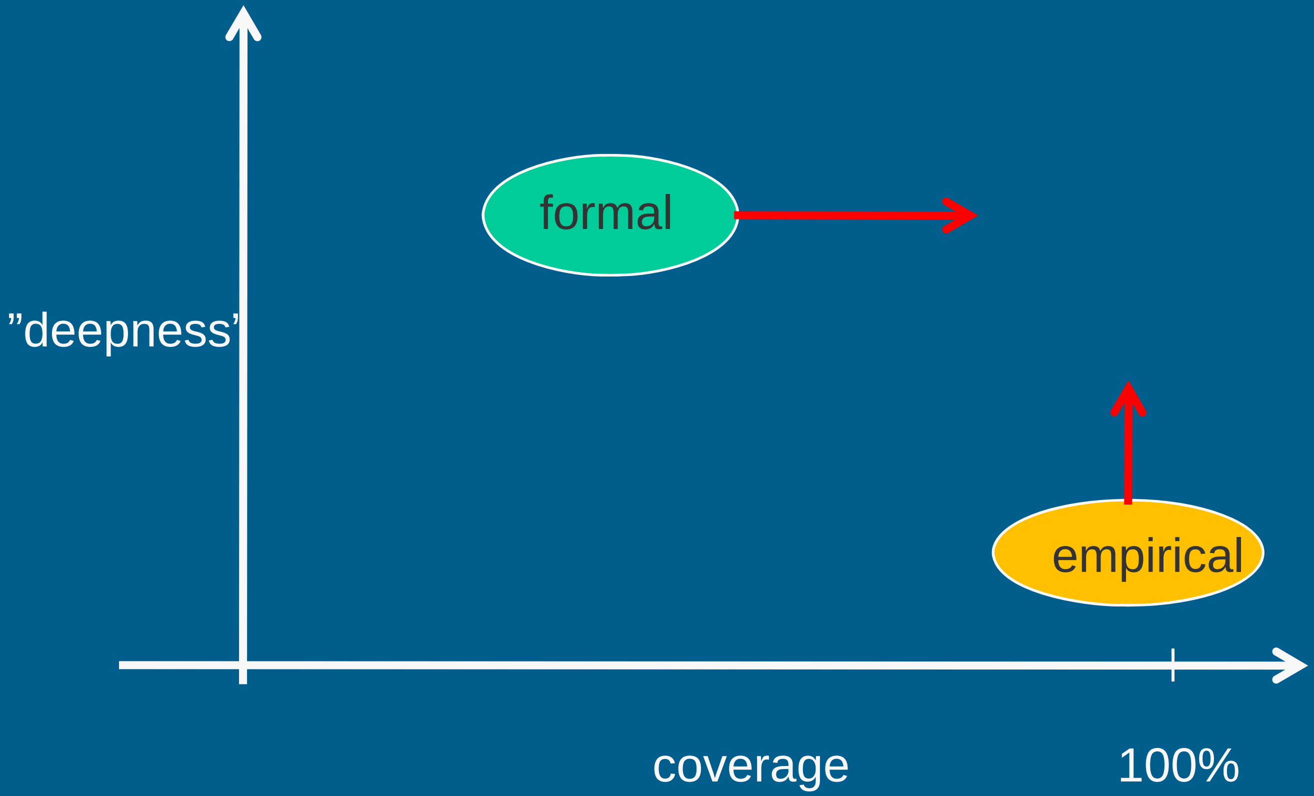
Formal approach: challenges

- Coverage
  - Ca 80%
  - The grammar isn’t complete
  - The text isn’t grammatical
- Ambiguities
  - Sentences are ambiguous
  - Long sentences may get many parses (in the thousands)
- Larger coverage → more rules → more ambiguities
- Efficiency

Empirical methods

- Examples:
  - Tagging
  - Speech recognition
  - Statistical MT
- Learn from examples: generalize
- Stochastic methods: probabilities
- Challenge for analysis:
  - Input to compositional semantics

Two approaches



From formal towards hybrid

- Coverage:
  - Supply with simpler methods where the formal method fails
  - Challenge: compatible output
- Ambiguities
  - Stochastic methods

A decisive difference

- Formal methods:
  - A clearcut division between
    - Grammatical – ungrammatical
    - Possible analysis – impossible
  - Choosing the most probable between the grammatical ones
- Empirical, stochastic approach
  - Choose the ”best” (most probable)
  - No division between possible and impossible

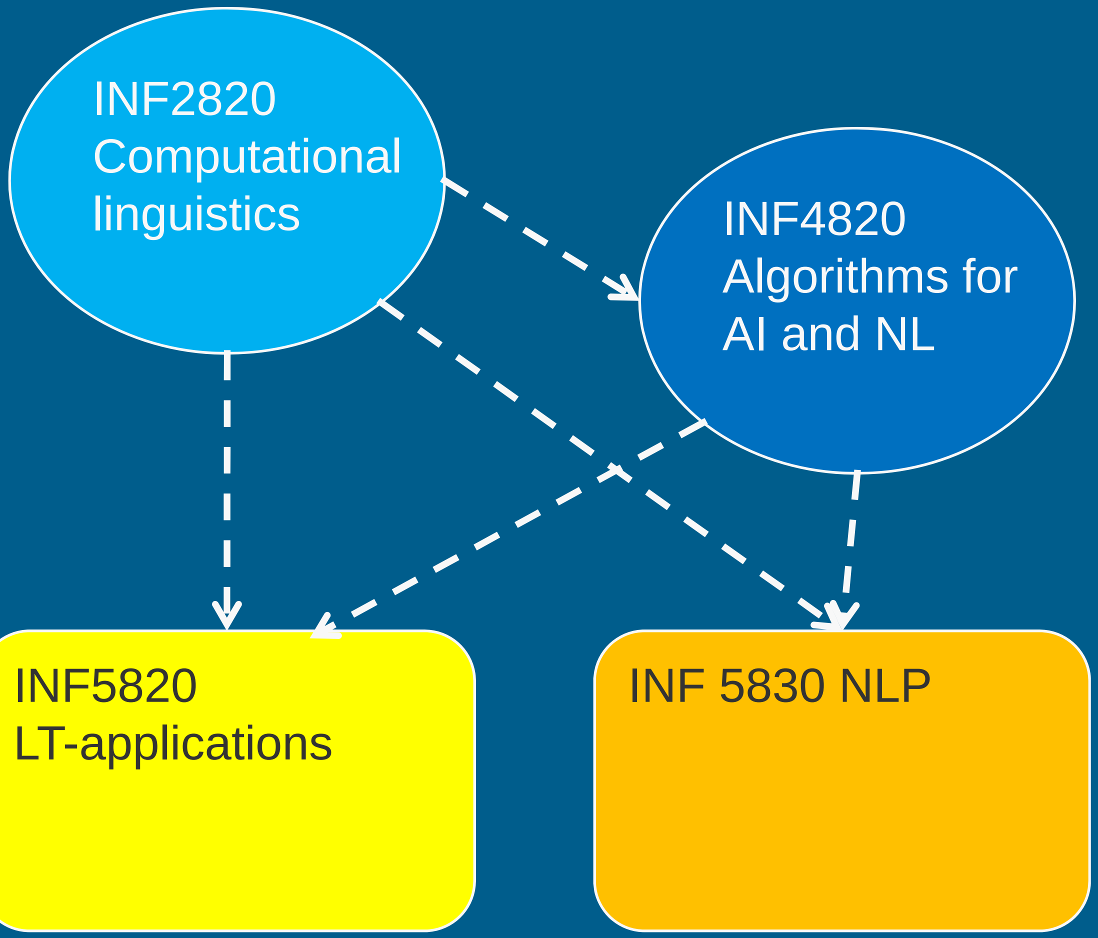
INF5830

- http://www.uio.no/studier/emner/matnat/ifi/INF580/index.xml
- Bygger på INF4820 (kan tas samtidig)
- Alternerer med INF5820 Language technological applications

Mixed audience

- Challenge:
  - Participants have different backgrounds (e.g. INF4820, 5820)
  - Content of some courses have changed
    - E.g. HMM in INF4820
    - Probabilistic CFG in INF2820/INF4820
- Goal:
  - INF2820 or INF4820 sufficient background
  - Avoid repetition
  - Consult INF4820

Related courses



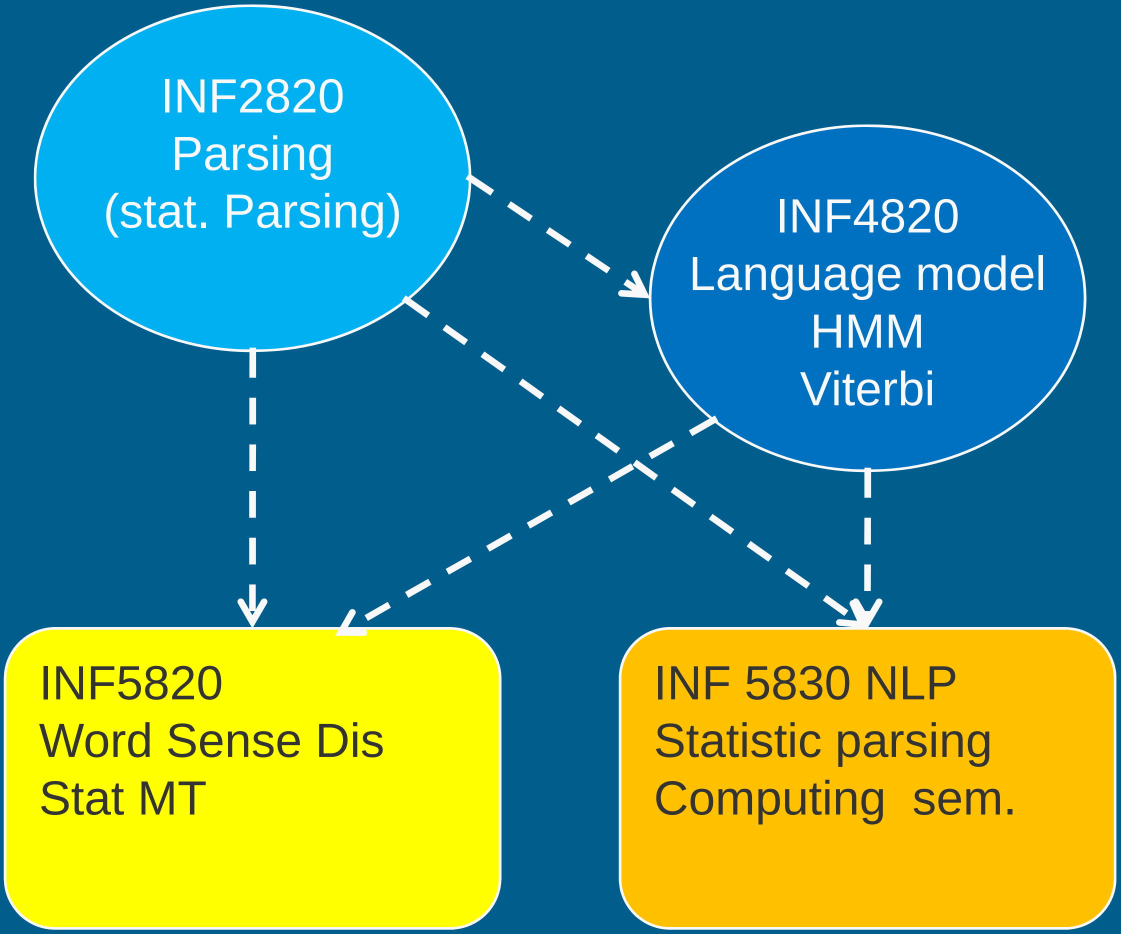
INF2820 Computational linguistics

INF4820 Algorithms for AI and NL

INF5820 LT-applications

INF 5830 NLP

Statistical NLP



INF2820
Parsing
(stat. Parsing)

INF4820
Language model
HMM
Viterbi

INF5820
Word Sense Dis
Stat MT

INF 5830 NLP
Statistic parsing
Computing sem.

-stat.
inference ?
- smoothing ?
- information
theory ?

Content

- Probabilities 28.8 (=INF4820, 5820)
- Tagging
  - CG
  - HMM, short (more in INF4820: Viterbi)
  - Max Ent
- Probabilistic CFG
  - Basic
  - CKY-parsing
  - Charniak-parser
  - Collins-parser

Content, contd.

- RASP-systemet
- Dependency parsing
- From parsing to semantics
  - PropBank, FrameNet
  - Role labeling
  - Relation detection

Schedule

- Class
  - Monday 14.15-16
  - Wednesday 10.15-12 (not every week)
- Exam
  - Dec. 10, 2:30 PM

Assignments

- 3 sets
- Familarize ourselves with techniques and tools

1. N-gram tagging
2. Prob. Parsing
3. Small group project

PhD-students

- Use code INF9830
- Supposed to do more than master students
- Class presentation

PART OF SPEECH TAGGING

Part of speech tagging

- Example: Oslo-Bergen-tagger

Parts of Speech

- 8 (ish) traditional parts of speech
  - Noun, verb, adjective, preposition, adverb, article, interjection, pronoun, conjunction, etc
  - Called: parts-of-speech, lexical categories, word classes, morphological classes, lexical tags...
  - Lots of debate within linguistics about the number, nature, and universality of these
    - We’ll completely ignore this debate.

POS examples

N	noun	*chair, bandwidth, pacing*
V	verb	*study, debate, munch*
ADJ	adjective	*purple, tall, ridiculous*
ADV	adverb	*unfortunately, slowly*
P	preposition	*of, by, to*
PRO	pronoun	*I, me, mine*
DET	determiner	*the, a, that, those*

POS Tagging

- J&M: “The process of assigning a part-of-speech or lexical class marker to each word in a collection.”

WORD	tag
the	**DET**
koala	**N**
put	**V**
the	**DET**
keys	**N**
on	**P**
the	**DET**
table	**N**

Why is POS Tagging Useful?

- First step of
  - Chunking (partial parsing)
  - Named entity recognition
  - Word sense disambiguation
- Speech synthesis
  - How to pronounce “lead”? No: “passasjer”?

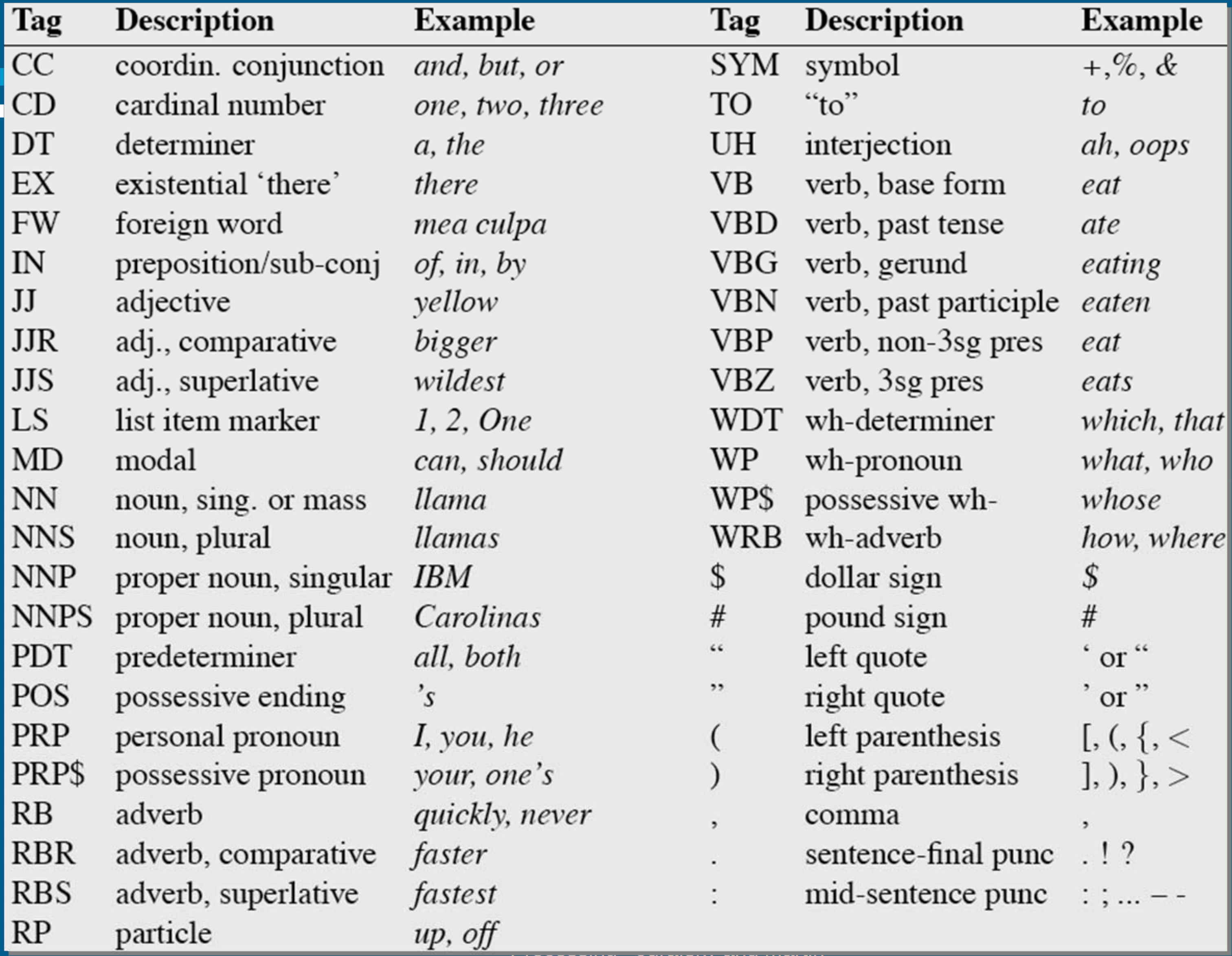
INsult	inSULT
OBject	obJECT
OVERflow	overFLOW
DIScount	disCOUNT

- Information extraction
  - Lemmatization
  - Finding names, relations, etc.
- POS brings info to neighboring words
  - Speech recognition

Choosing a Tagset

- There are so many parts of speech, potential distinctions we can draw
- To do POS tagging, we need to choose a standard set of tags to work with
- Could pick very coarse tagsets
  - N, V, Adj, Adv.
- More commonly used set is finer grained, the “Penn TreeBank tagset”, 45 tags
  - PRP$, WRB, WP$, VBG
- Even more fine-grained tagsets exist
- Tradeoff:
  - How much information is needed?
  - How difficult is the disambiguation?

Pen TreeBank POS Tagset



Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	“to”	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential ‘there’	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	“	left quote	‘ or “
POS	possessive ending	*’s*	”	right quote	’ or ”
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one’s*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

Using the Penn Tagset

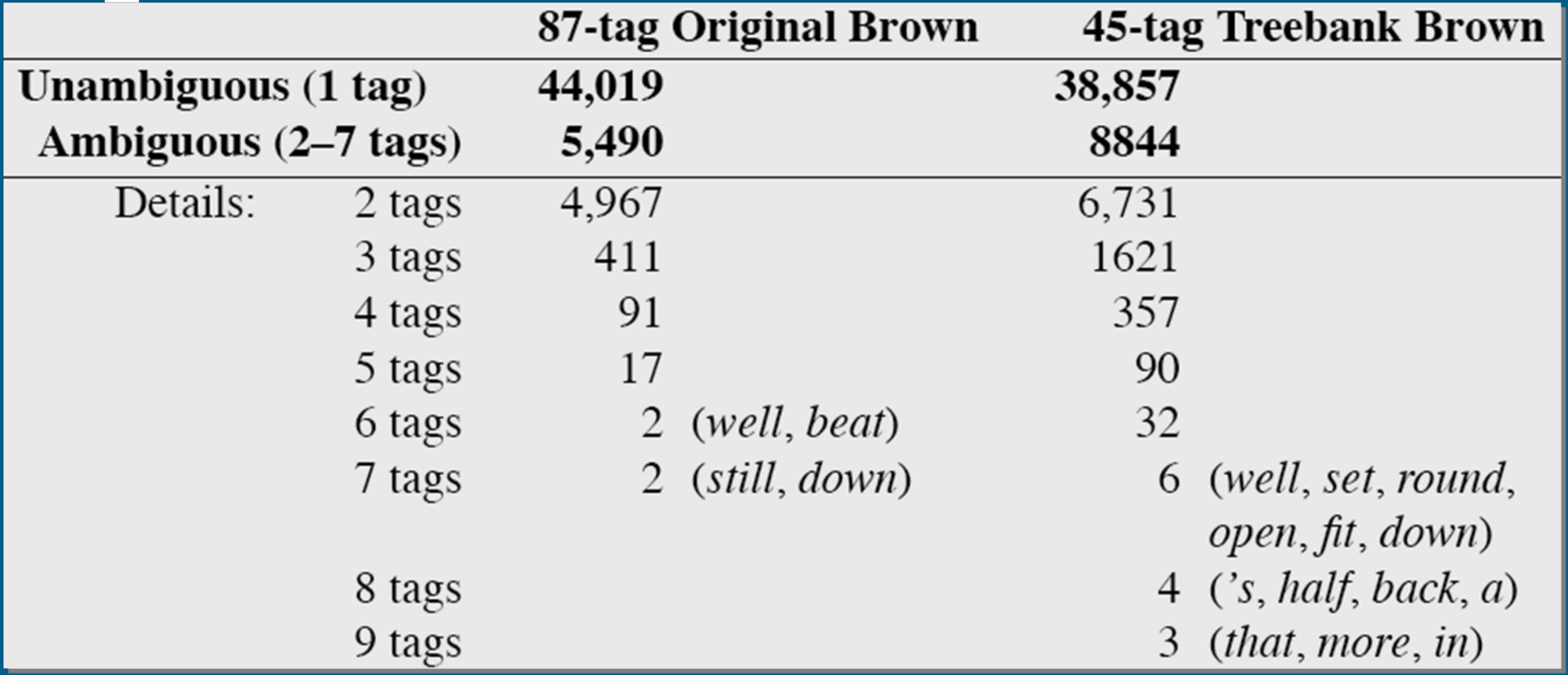
- The/DT grand/JJ jury/NN commmented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
- Prepositions and subordinating conjunctions marked IN (“although/IN I/PRP..”)
- Except the preposition/complementizer “to” is just marked “TO”.

POS Tagging

- Words often have more than one POS: *back*
  - The *back* door = JJ
  - On my *back* = NN
  - Win the voters *back* = RB
  - Promised to *back* the bill = VB
- The POS tagging problem is to determine the POS tag for a particular instance of a word.

These examples from Dekang Lin

How Hard is POS Tagging? Measuring Ambiguity



		87-tag Original Brown		45-tag Treebank Brown	
Unambiguous (1 tag)		**44,019**		**38,857**	
Ambiguous (2–7 tags)		**5,490**		**8844**	
Details:	2 tags	4,967		6,731	
	3 tags	411		1621	
	4 tags	91		357	
	5 tags	17		90	
	6 tags	2	(*well*, *beat*)	32	
	7 tags	2	(*still*, *down*)	6	(*well*, *set*, *round*, *open*, *fit*, *down*)
	8 tags			4	(*'s*, *half*, *back*, *a*)
	9 tags			3	(*that*, *more*, *in*)

Two Methods for POS Tagging

1. Rule-based tagging
   - (ENGTWOL)
2. Stochastic
   1. Probabilistic sequence models
      - HMM (Hidden Markov Model) tagging
      - MEMMs (Maximum Entropy Markov Models)

Rule-Based Tagging

- Start with a dictionary
- Assign all possible tags to words from the dictionary
- Write rules by hand to selectively remove tags
- Leaving the correct tag for each word.

Start With a Dictionary

she:	PRP
promised:	VBN,VBD
to	TO
back:	VB, JJ, RB, NN
the:	DT
bill:	NN, VB

- Etc… for the ~100,000 words of English with more than 1 tag

Assign Every Possible Tag

			NN		
			RB		
	VBN		JJ		VB
PRP	VBD	TO	VB	DT	NN
She	**promised**	**to**	**back**	**the**	**bill**

Tagging vs parsing

- A tagger faces the same two tasks as a grammar-based parser
- Ambiguity:
  - Choose the correct tag sequence between several candidates
- Coverage:
  - Assigning tags to words not in the lexicon:
    - Proper names
    - New words
    - Compounds
    - typos

Ambiguity

- How to tag genuine ambiguities?

	VB	PRP$	NN
PRP	VBD	PRP	VB
I	saw	her	duck

- Possible parses:
  - PRP VB PRP$ NN
  - PRP VBD PRP$ NN
  - PRP VBD PRP VB
- Impossible
  - PRP VBD PRP VB
  - + 4more