Formal versus statistical enrichment of grammars

Andreas van Cranenburgh

Huygens ING
Royal Netherlands Academy of Arts and Sciences

Institute for Logic, Language and Computation
University of Amsterdam

November 1, 2016

Rich Parsing Workshop, Amsterdam, 2016

Computational linguistics

. . . driven by empirical evaluation, benchmarks

Progress means . . .

- Improve on benchmark
- Challenge benchmark

Statistical Parsing

Many `inconvenient' aspects of treebank annotation typically ignored:

- non-local relations
- function tags
- morphology
- multiple parents

Statistical Parsing

Many `inconvenient' aspects of treebank annotation typically ignored:

- non-local relations
- function tags
- morphology
- multiple parents

Goal

given a treebank, fully reproduce its annotations with an automatically-induced statistical parser

Treebank annotation

PTB:

```
(S (NP-SBJ-1 (DT A) (NN record) (NN date) )
  (VP (VBZ has) (RB n't) (VP (VBN been)
    (VP (VBN set) (NP (-NONE- *-1) )))) (. .) )
```

Treebank annotation

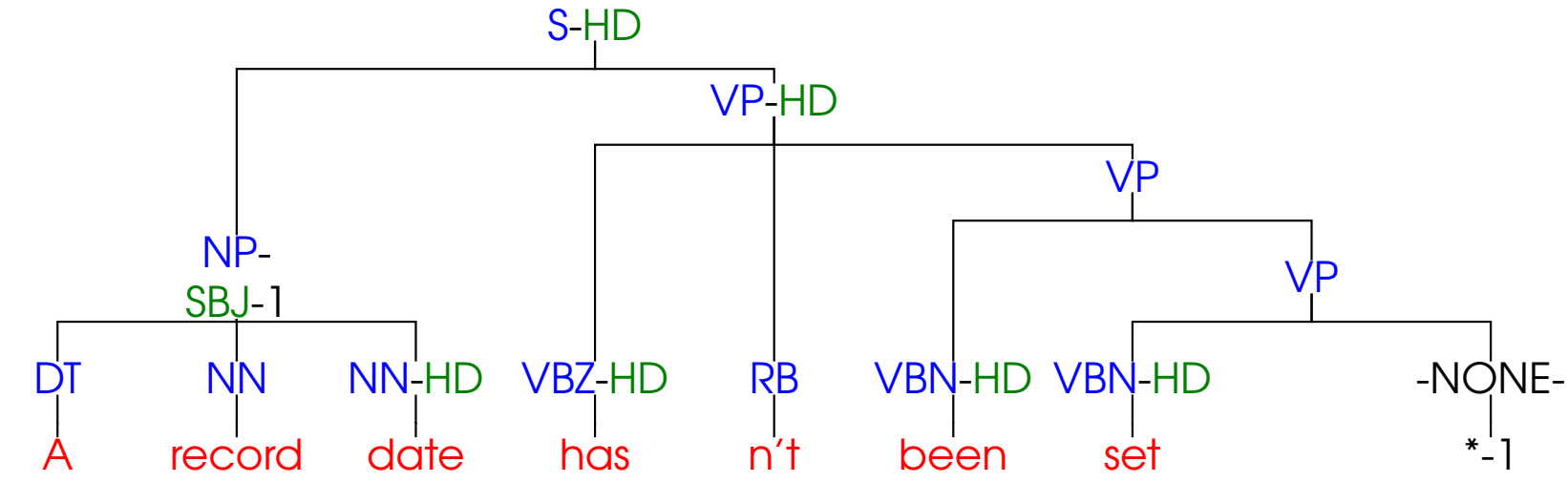
PTB:

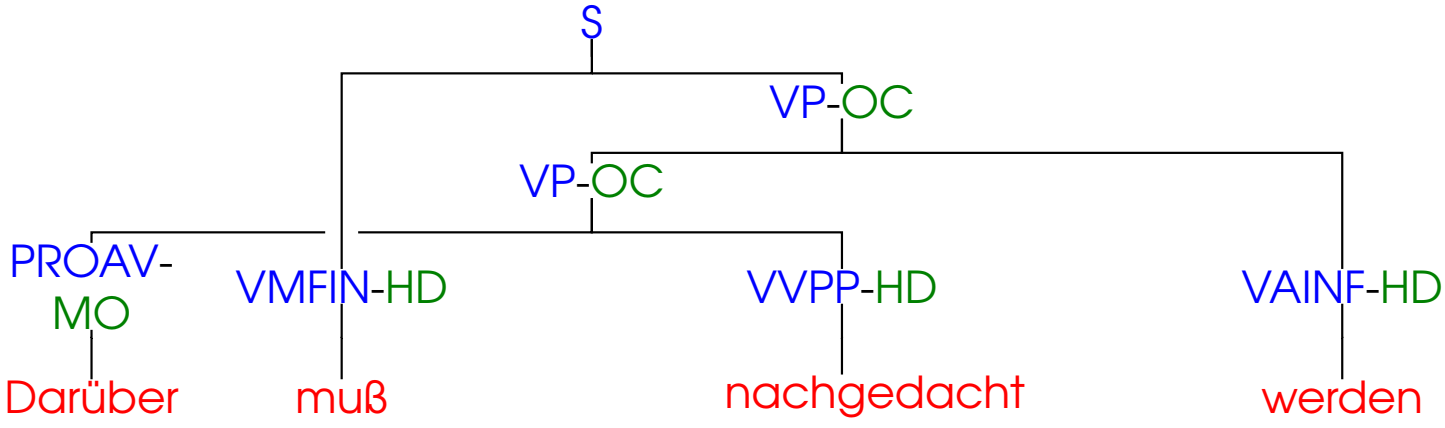
```
(S (NP-SBJ-1 (DT A) (NN record) (NN date) )
  (VP (VBZ has) (RB n't) (VP (VBN been)
    (VP (VBN set) (NP (-NONE- *-1) )))) (. .) )
```

Negra:

%% word	cat	morph	func	parent
Darüber	PROAV	--	MO	500
muß	VMFIN	3.Sg.Pres.Ind	HD	502
nachgedacht	VVPP	--	HD	500
werden	VAINF	--	HD	501
.	$.	--	--	0
#500	VP	--	OC	501
#501	VP	--	OC	502
#502	S	--	--	0

Treebank annotation





Grammar enrichment

Formal

A grammar formalism that

- precisely matches the generative capacity of natural language
- is specifically designed to produce the desired linguistic analyses

Grammar enrichment

Formal

A grammar formalism that

- precisely matches the generative capacity of natural language
- is specifically designed to produce the desired linguistic analyses

Statistical

Heuristic approach:

- add extra information by augmenting labels
- apply pre- and postprocessing steps
- exploit regularities in corpus, e.g. co-occurrence of elements, automatic state splits, &c.

Chomsky (1965):

Competence system of rules describing idealized knowledge of language

Performance language behavior affected by ambiguity, errors, reaction times, frequency effects

Chomsky (1965):

Competence system of rules describing idealized knowledge of language

Performance language behavior affected by ambiguity, errors, reaction times, frequency effects

Scha (1990):

- Difficult to write descriptively adequate grammar by hand.
- Problem of ambiguity; need to know relative plausibility of analyses.

Ergo, we need
"performance-models of language (...), which take into account statistical properties of actual language use."

Traditional parsing approach

1. Pick a grammar with the right linguistic & computational properties (competence)
2. Apply pruning if necessary (performance)
3. Add a probabilistic disambiguation component (performance)
4. Evaluate quality of model (performance)

Formal language theory

Definition

A *formal grammar* characterizes a language as a set of sentences and their structures.

Chomsky hierarchy:

Type 0: Unrestricted: Model-Theoretic Syntax, e.g., HPSG

Type 1: Context-Sensitive: Mildly Context-Sensitive, e.g., TAG, CCG, LCFRS

Type 2: Context-Free: PCFG, proj. dependency grammar

Type 3: Regular: finite-state technology

Domain of locality

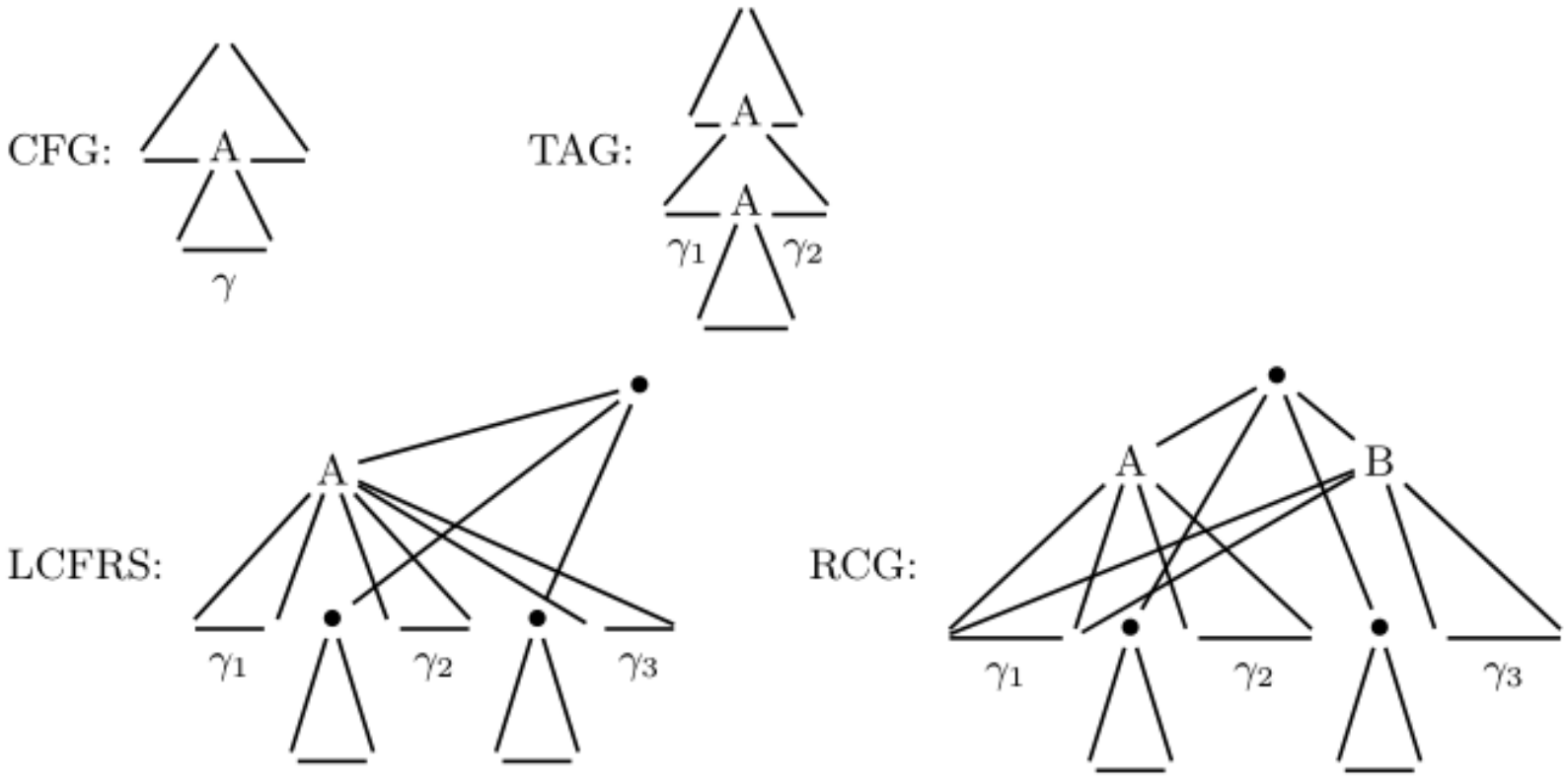
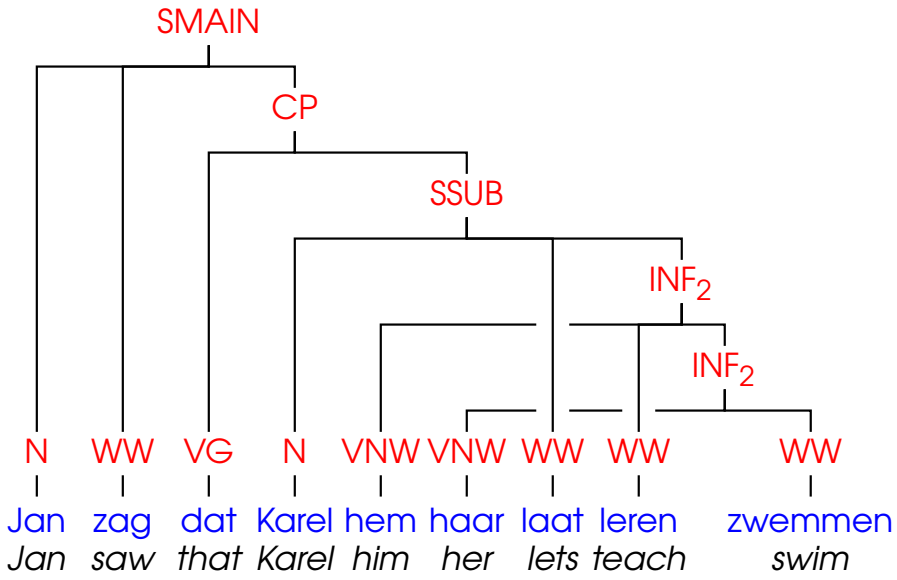


Fig. 1.1. Yields of non-terminals in different formalisms

Figure: Kallmeyer (2010): Parsing beyond CFG, p. 3

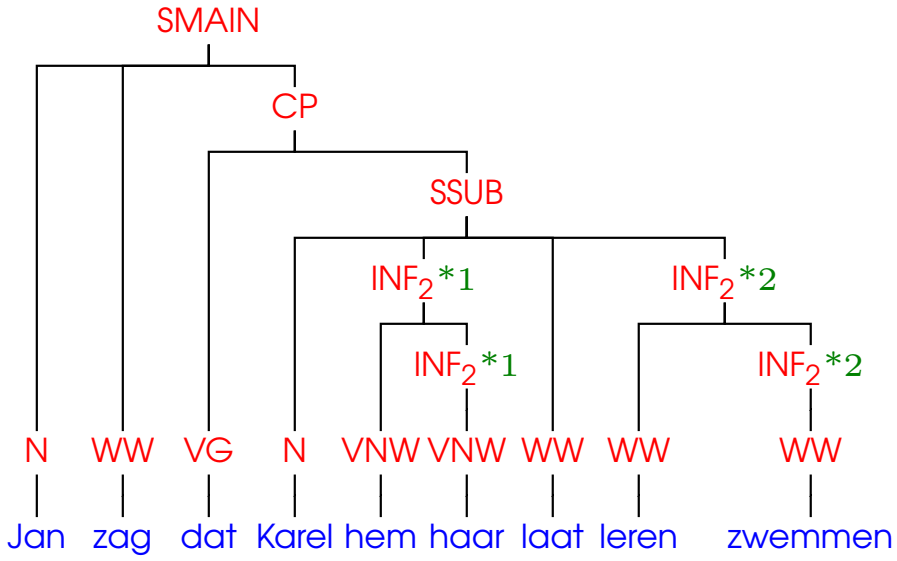
Long-Distance Dependencies



"Jan saw that Karel lets him teach her to swim."

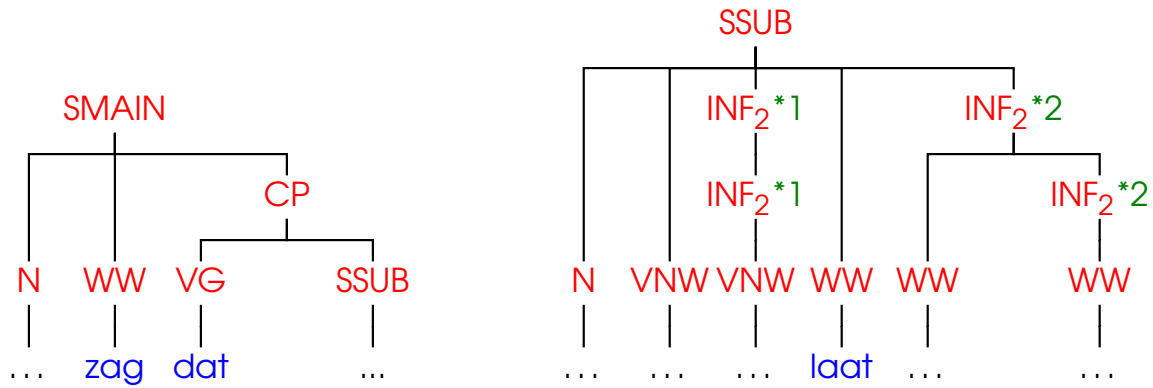
- Cross-serial dependencies are beyond context-free
- Can be captured by mildly context-sensitive grammars

CFG approximation



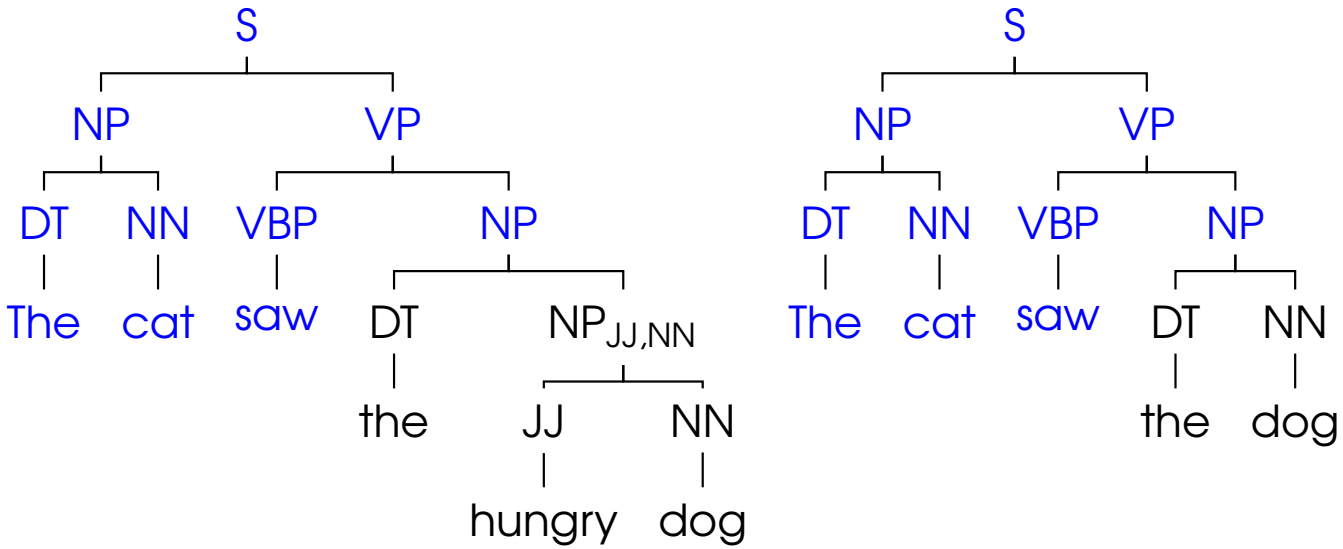
- Alternatively, long-distance dependencies can be encoded in the labels

Non-locality w/DOP fragments



- With DOP tree fragments,
  complex linguistic phenomena can be captured
  statistically instead of formally

Grammar induction: 2DOP



- Induce a Tree-Substitution Grammar from treebank
- Heuristic: recurring tree fragments are building blocks
- Compare pairs of trees
  and extract common fragments

Sangati & Zuidema (2011). Accurate parsing w/compact TSGs: Double-DOP

Function labels

Syntactic categories (form): NP, VP, S, ...

Function labels (function): SBJ, OBJ, TMP, LOC, ...

- Classifier:
  - Blaheta & Charniak (2000), Assigning Function Tags to Parsed Text
- Integrate in grammar:
  - Gabbard et al. (2006), Fully parsing the Penn treebank
  - Fraser et al. (2013), Knowledge sources for constituent parsing of German

Evaluation: function tag accuracy over correctly parsed labeled bracketings.

Can DOP handle discontinuity without LCFRS?

Negra dev set, gold tags:

Split-PCFG

⇓

PLCFRS

⇓

PLCFRS Double-DOP

77.7 % F1

41.5 % EX

Split-PCFG

⇓

Split-Double-DOP

Can DOP handle discontinuity without LCFRS?

Negra dev set, gold tags:

Split-PCFG

⇓

PLCFRS

⇓

PLCFRS Double-DOP

77.7 % F1

41.5 % EX

Split-PCFG

⇓

Split-Double-DOP

78.1 % F1

42.0 % EX

Answer: Yes!

Fragments can capture discontinuous contexts

Importance of probabilities

What happens when probabilities of fragments are randomly shuffled?

Negra parsing	F1
PLCFRS treebank grammar	65.9
2DOP	77.7
2DOP, shuffled probabilities	74.1

Importance of probabilities

What happens when probabilities of fragments are randomly shuffled?

Negra parsing	F1
PLCFRS treebank grammar	65.9
2DOP	77.7
2DOP, shuffled probabilities	74.1

Conclusion

co-occurrence of productions more important than frequency effects.

Parsing results

Parser	F1	EX	func
GERMAN: Tiger			
Dep: HaNi2008	75.3	32.6	
2DOP: Cr et al	**78.2**	**40.0**	93.5
Dep: FeMa2015	82.6	45.9	
ENGLISH: WSJ			
PLCFRS: EvKa2011	79.0		
2DOP: Cr et al, WSJ	**87.0**	34.4	86.3
2DOP: SaZu2011, no disc.	87.9	33.7	
DUTCH: Lassy			
2DOP: Cr et al	76.6	34.0	92.8

HaNi: Hall & Nivre (2008); SaZu: Sangati & Zuidema (EMNLP 2011);
EvKa: Evang & Kallmeyer (IWPT 2011);
FeMa: Fernández-González & Martins (ACL 2015);
Cr et al: van Cranenburgh, Scha, Bod (JLM 2016).

Conclusion

Linguistically rich: non-local relations, function tags

Efficiency: CFG base grammar, tree fragment extraction

Competence: idealized rules

Performance: actual language use

Tree fragments increase the abilities of a performance model w.r.t. discontinuous constituents, without increasing formal complexity.


KEEP
CALM
because
THIS TOO
SHALL PARSE

DEMO

https://lang.science.uva.nl/parser/

References

- Remko Scha (1990). Language theory and language technology; competence and performance, in Q.A.M. de Kort and G.L.J. Leerdam, editors, *Computertoepassingen in de Neerlandistiek*, pp. 7–22. English translation: `http://iaaa.nl/rs/LeerdamE.html`
- van Cranenburgh, Scha, Bod (2016) Data-Oriented Parsing with Discontinuous Constituents and Function Tags. *Journal of Language Modelling*, vol. 4, no. 1, pp. 57-111. `http://dx.doi.org/10.15398/jlm.v4i1.100`