UNIVERSITY OF OSLO

Introducing CuteForce –

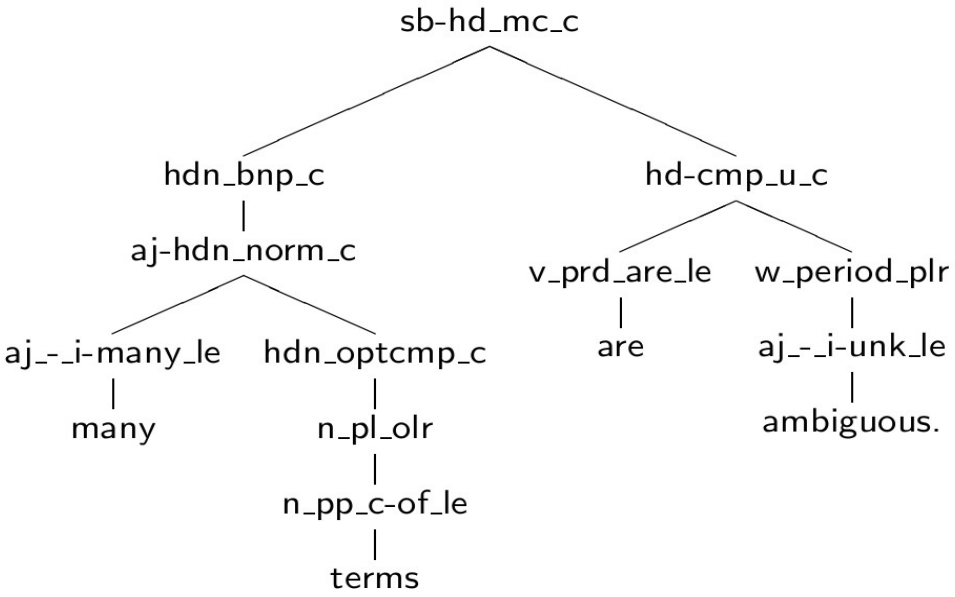
Efficient HPSG Parsing

Gisle Ytrestøl – University of Oslo
gisley@ifi.uio.no

UNIVERSITY OF OSLO

Goal

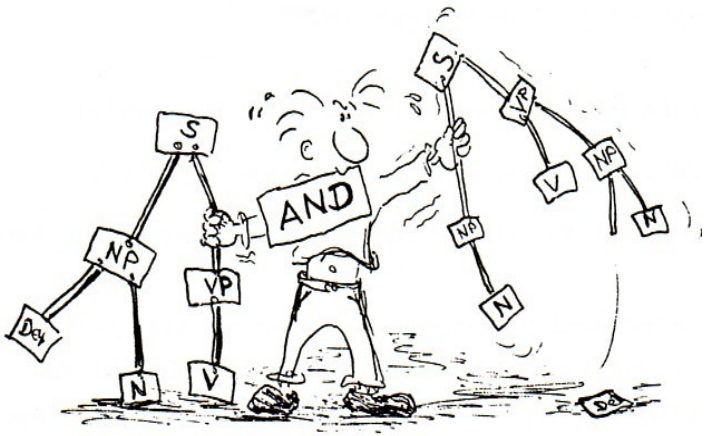
- Produce derivations like this:



Signature (unit, name, etc.)

Why? We already have PET/ERG!

- PET/ERG offers:
  - linguistically well-founded approach
  - coverage ~80 – 90% across domains: Wikipedia, GENIA, WSJ, et al
  - 500+ number of parses for long/ medium length sentences
  - fully correct tree ranked best in ~55% of cases
  - an average ~5 sec wait for each sentence

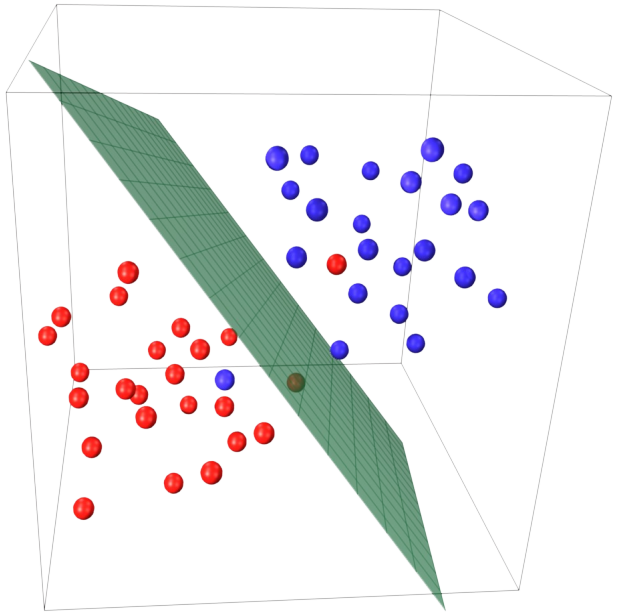


UNIVERSITY OF OSLO

CuteForce - A Cuter Approach

- Derives only one parse per sentence
- Knows very little about linguistics
- Fast (<5 ms per sentence – would theoretically parse Wikipedia in 120 cpu hours, not 14 years)
- Written in Java
- Light-weight

CuteForce Parsing



- Builds the HPSG representation incrementally in a shift-reduce fashion
- Inspired by the MaltParser (Nivre, 2007)
- Any HPSG structure can be built through CuteForce transitions
- Transitions are selected through SVM models

Signature (unit, name, etc.)

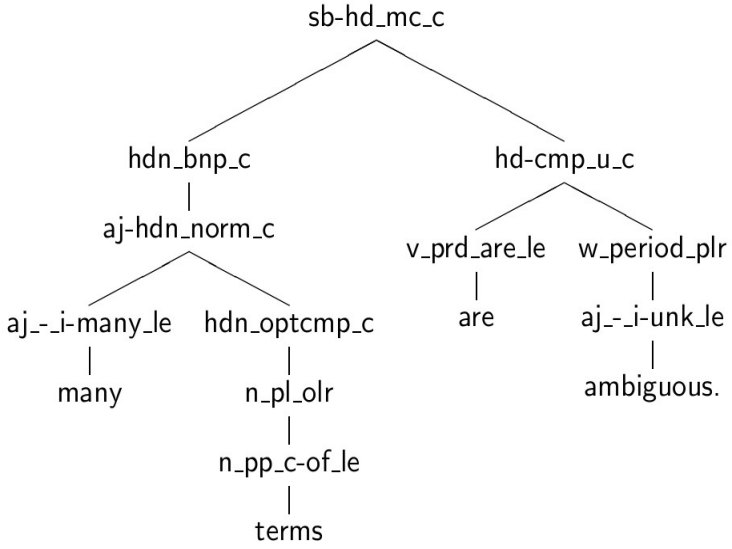
UNIVERSITY OF OSLO

CuteForce Parsing

- Learns this:

```
ARG #1=\\"defined\\" +PRED predsort +CLASS alphabe
 #1 ]")))))) (175324 hdn-aj_redrel_c 8.49123 15 20
nstruction_n1 -0.230751 15 16 ("instructions" 274
 #1=\\"instructions\\" +PRED predsort +CLASS alpha
RM #1 ]")))) (175323 hd-cmp_u_c 6.12964 16 20 (666
 \\"91\\" +ID *diff-list* +CARG #1=\\"for\\" +PRED
TRAIT native_trait +FORM #1 ])) (175322 hdn_bnp-v
.39161 17 20 (175316 v_prp_olr 0.299762 17 18 (675
\\" +FROM \\"95\\" +ID *diff-list* +CARG #1=\\"com
alized+lower ] +TRAIT native_trait +FORM #1 ])))
 [ +TNT null_tnt +TO \\"107\\" +FROM \\"106\\" +ID
L - +CASE non_capitalized+lower ] +TRAIT native_tr
79986 19 20 (698 task_n1 0 19 20 ("task" 282 "toke
=\\"task\\" +PRED predsort +CLASS alphabetic [ +IN
)))))))))))) (175930 hd-aj_scp-pr_c 39.3194 20 41
```

- Produces this:



Signature (unit, name, etc.)

Data Structure

- Input buffer β:
  - Tuples of word forms/PTB tags/lexical types:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
- Position index *i:*
  - The position index for β, starting at 1 for the first input triple.
- Active edges α:
  - Stack of active edges
  - Binary Constituents whose right-branched daughter/mother is not yet assigned
  - Empty upon termination
- Passive edges π:
  - Stack of passive edges
  - Represents the partial HPSG structure during parsing, and the full HPSG representation upon termination

UNIVERSITY OF OSLO

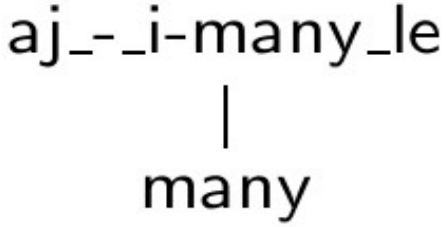
Transition system

- ACTIVE (adds active edge to α, adds lexical type β(*i+1*) to π, and increments *i*
- PASSIVE (pops α and adds binary passive edge to π)
- UNIT (adds unary passive edge to π)
- ACCEPT (Terminates the parse of the sentence. π represents the syntactic derivation of the sentence)

UNIVERSITY OF OSLO

Initial Parse State

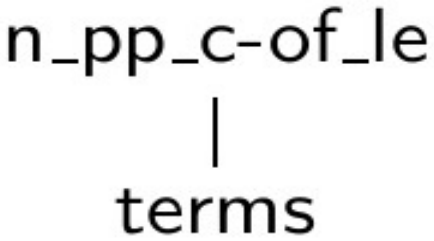
- π:



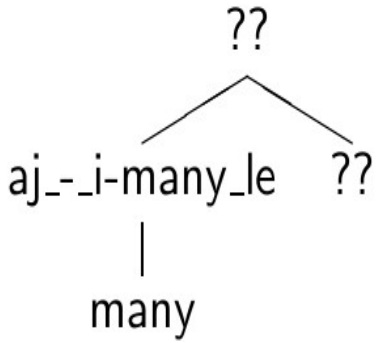
- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 1
- α: = e

Active

- π:



- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 2
- α:

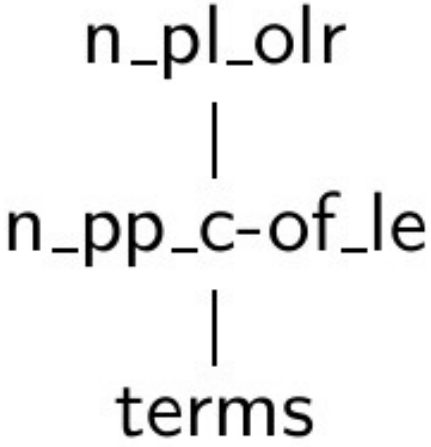


Signature (unit, name, etc.)

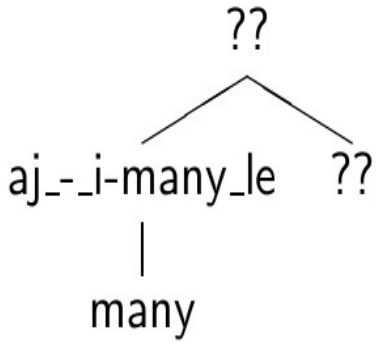
UNIVERSITY OF OSLO

Unit - (n_pl_olr)

•π:



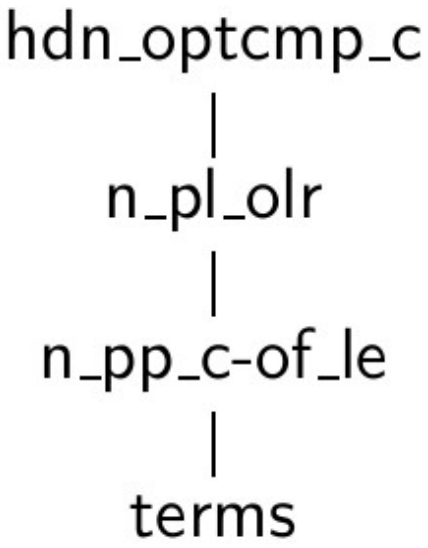
- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 2
- α:



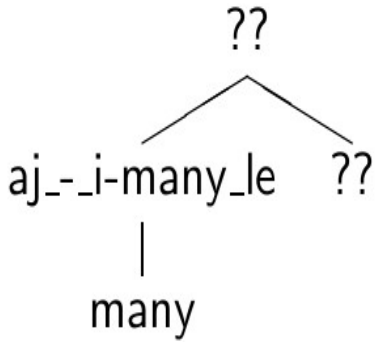
Signature (unit, name, etc.)

Unit - (hdn_optcmp_c)

- π:



- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 2
- α:

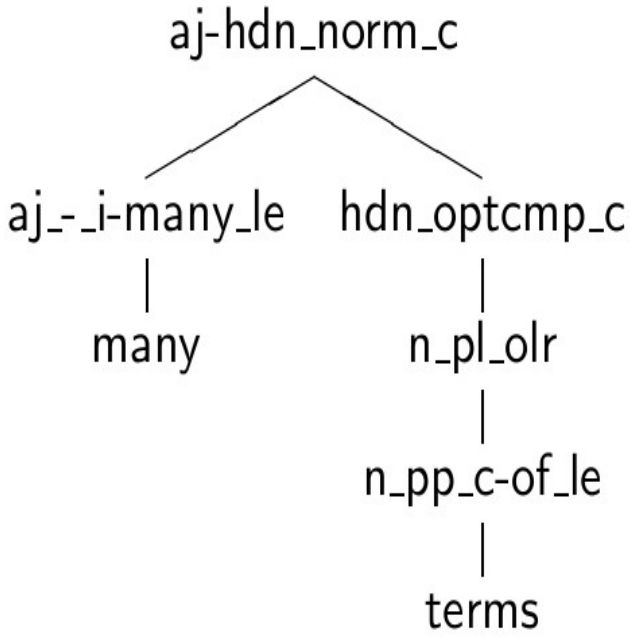


Signature (unit, name, etc.)

UNIVERSITY OF OSLO

Passive - (aj-hdn_norm_c)

- π:



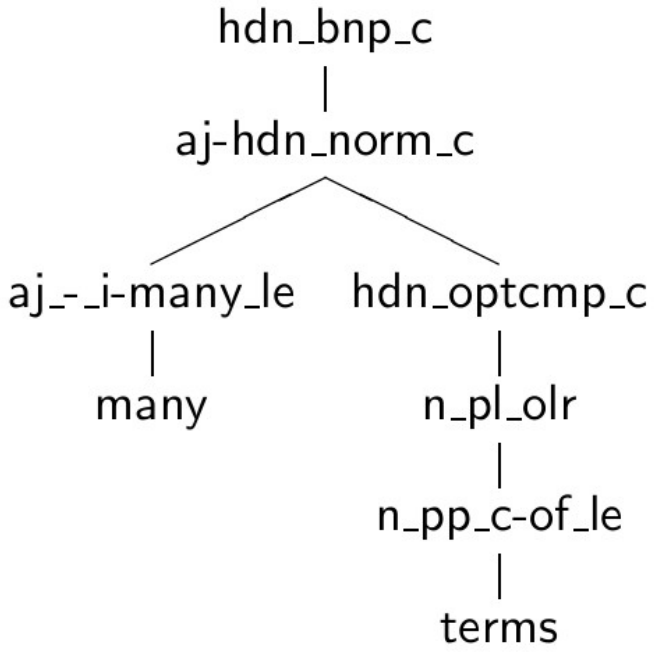
- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 2
- α:

Signature (unit, name, etc.)

UNIVERSITY OF OSLO

Unit - (hdn_bnp_c)

- π:



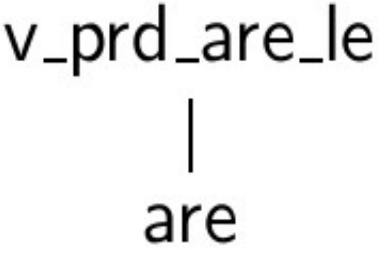
- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 2
- α:

Signature (unit, name, etc.)

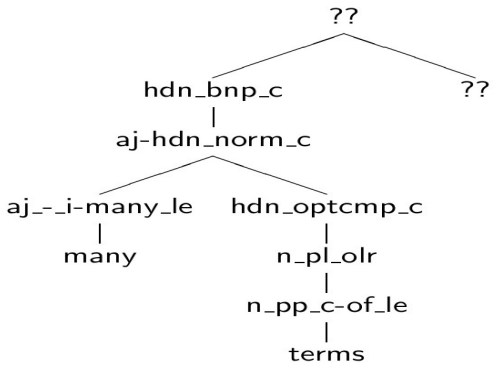
UNIVERSITY OF OSLO

Active

- π:



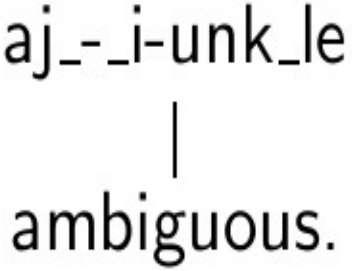
- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 3
- α:



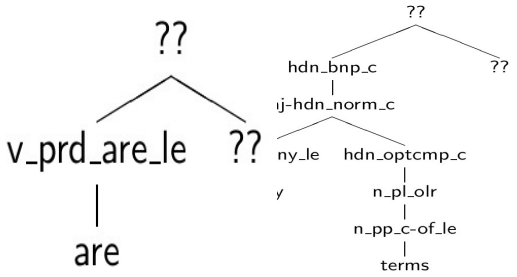
Signature (unit, name, etc.)

Active

- π:

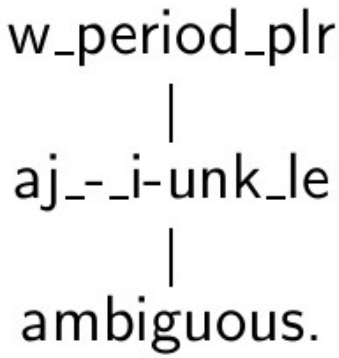


- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 4
- α:

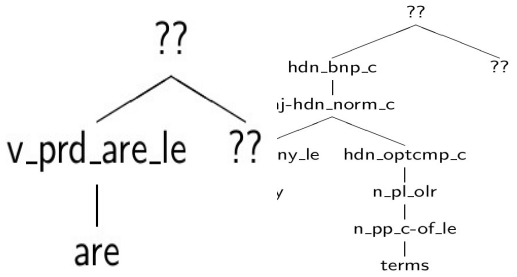


Unit - (w_period_plr)

- π:



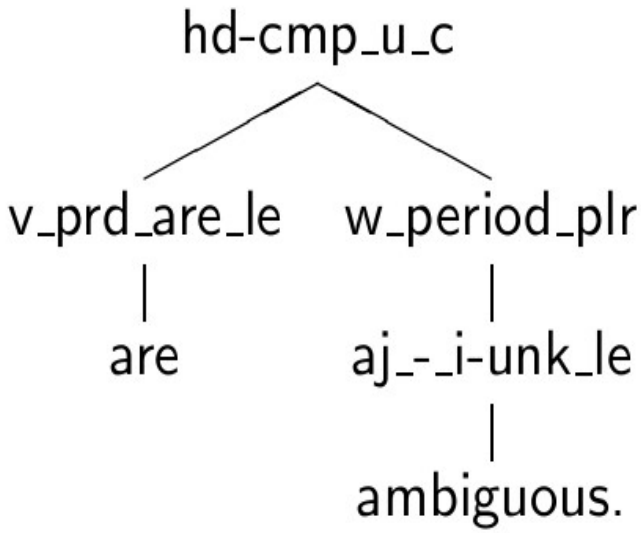
- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 4
- α:



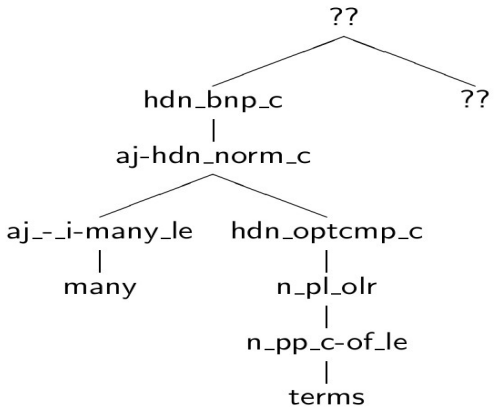
Signature (unit, name, etc.)

Passive - (hd-cmp_u_c)

- π:



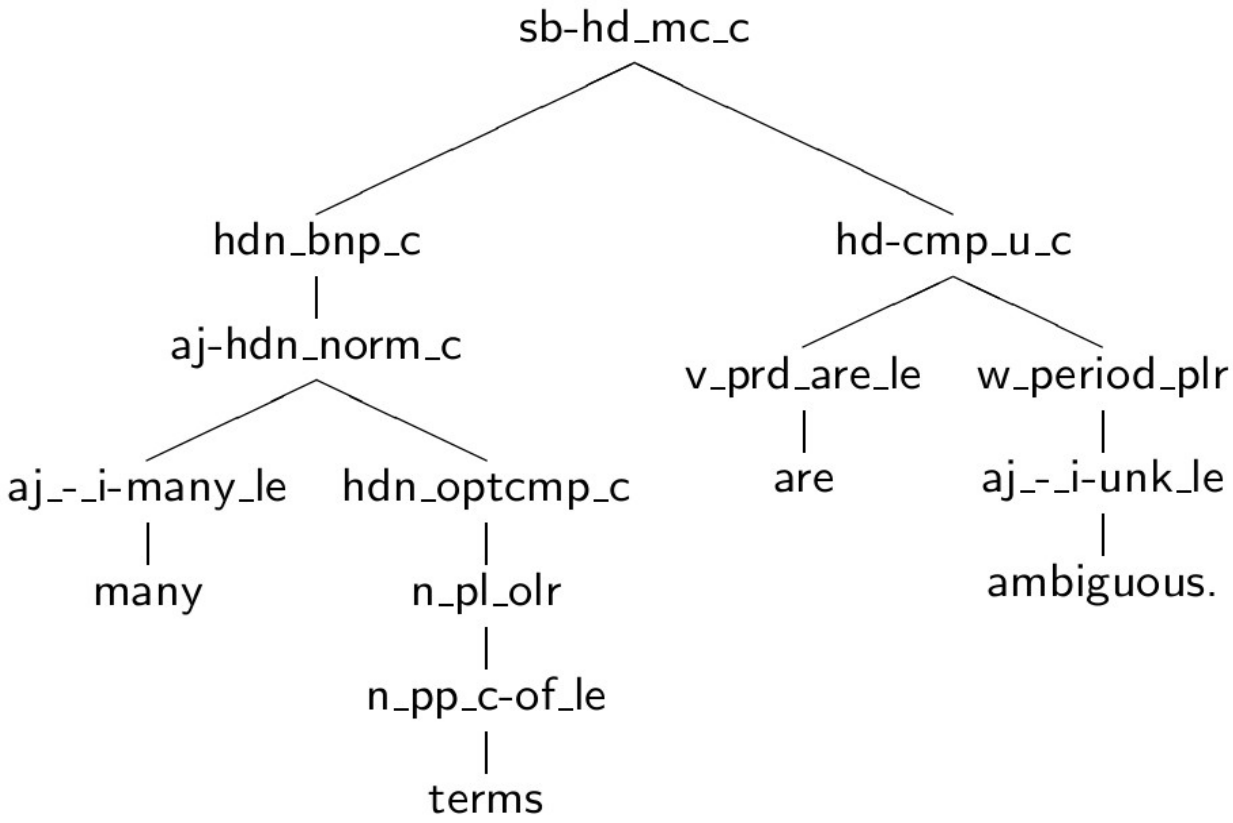
- Input buffer β:
  - (many|JJ|aj_-_i-many_le), (terms|NNS|n_pp_c-of_le), (are|VBP|v_prd_are_le), (ambiguous.|JJ|aj_-_i-nk_le)
  - i = 4
- α:



UNIVERSITY OF OSLO

Passive - (sb-hd_mc_c)

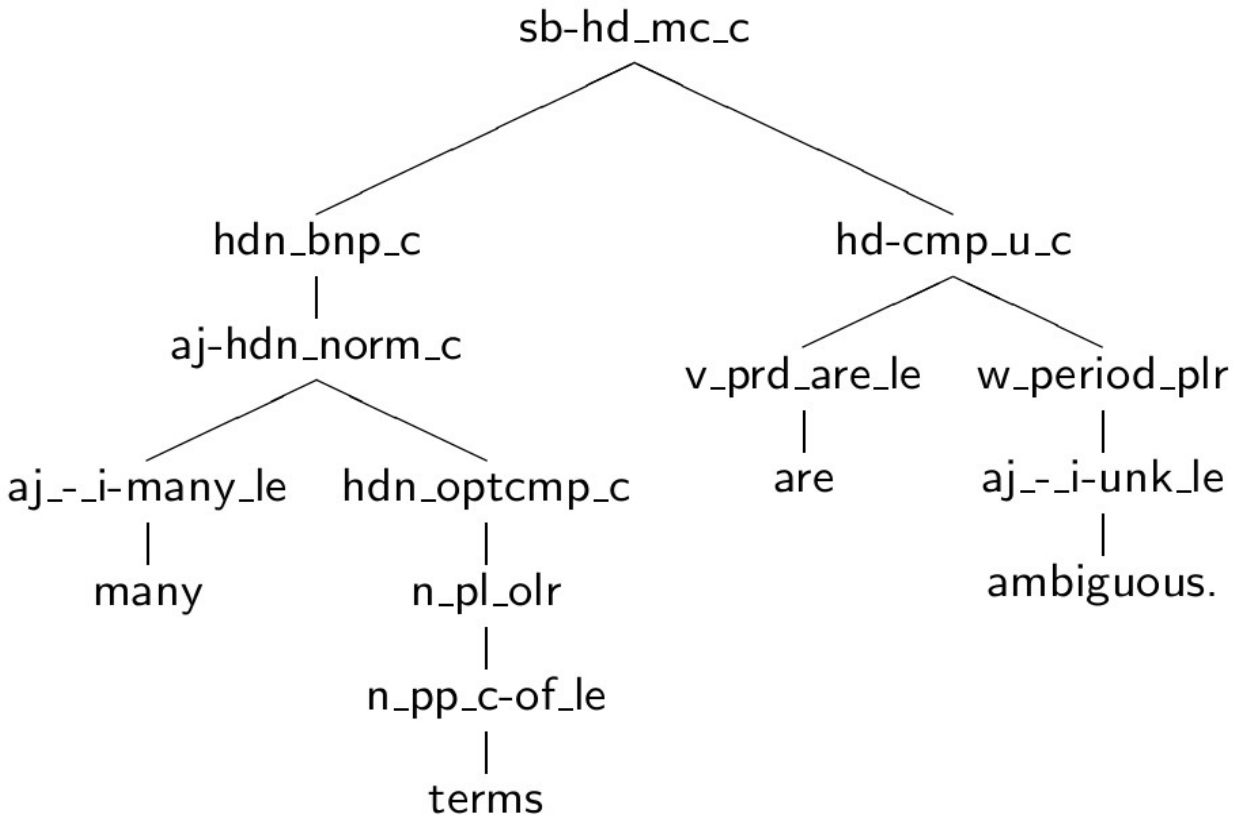
- π:



Signature (unit, name, etc.)

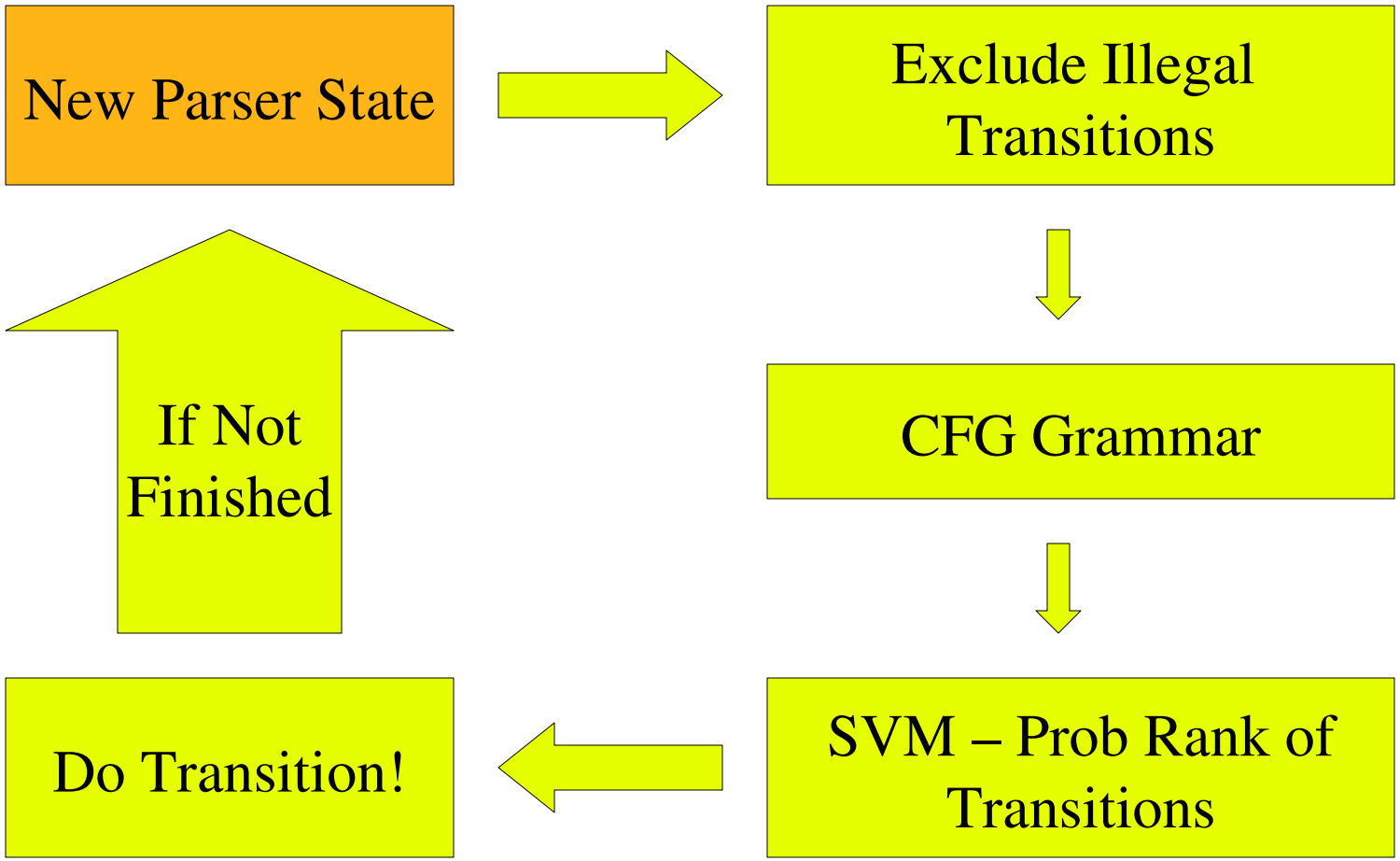
UNIVERSITY OF OSLO

Finished



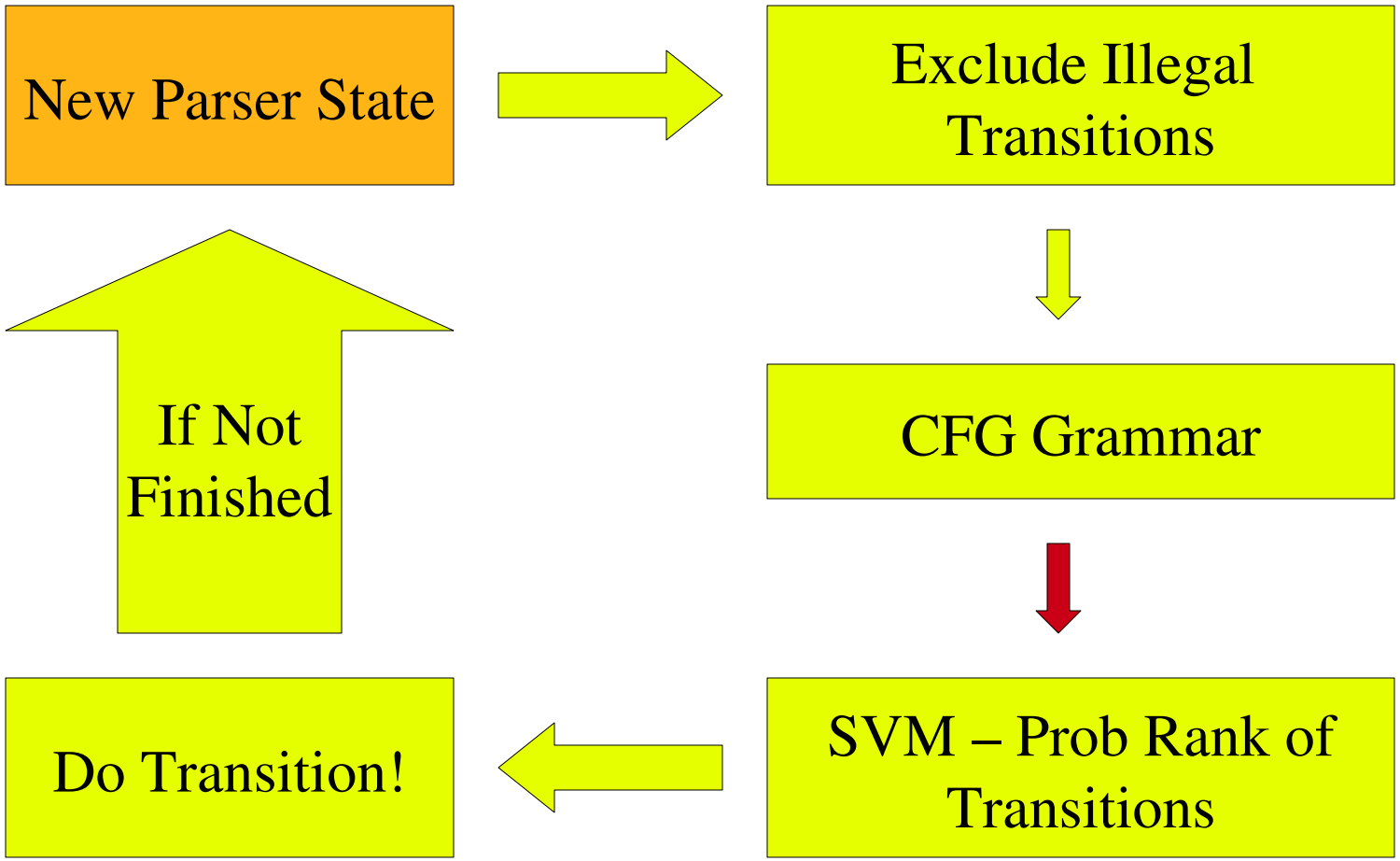
Signature (unit, name, etc.)

Oracle



Signature (unit, name, etc.)

Common Point of Failure



SuperTagging

- *Lexical Type: A feature structure containing category information for the lexical entry. Describes at least head type and valence information. (Dridan, 2009)*
- Input buffer:
  - Word form (ERG Gold Standard Tokenization)
  - PTB Tags (Training/Test data tagged by TnT)
  - Lexical Types (Gold Standard for Training, Test data tagged by C&C SuperTagger)

UNIVERSITY OF OSLO

SuperTagging

- *<part-of-speech>_<subcategorisation_<description>_le*
- Current C&C Accuracy:
  - LT Accuracy: 0.86
  - POS & SUB Accuracy: 0.93
  - POS Accuracy: 0.97

UNIVERSITY OF OSLO

Preliminary Results* / Supertagged Input

- <u>CFG Constraints:</u>
- Coverage: ~0.71
- Exact Match: ~0.39
- Parseval f-score: ~0.82
- Sentence Length:
  - Avg # of tokens: ~ 14.0
  - Parsed: ~ 11.2
  - Failed: ~ 20.9

- <u>No CFG Constraints:</u>
- Coverage: ~0.99
- Exact Match: ~0.26
- Parseval f-score: ~0.76
- Sentence Length:
  - Avg # of tokens: ~ 14.0
  - Parsed: ~ 14.0
  - Failed: ~ 18.0

*Lexical types not evaluated

UNIVERSITY OF OSLO

Preliminary Results* / Gold Standard Buffer Input

- <u>CFG Constraints:</u>
- Coverage: ~0.72
- Exact Match: ~0.50
- Parseval f-score: ~0.89
- Sentence Length:
  - Avg # of tokens: ~ 14.0
  - Parsed: ~ 11.2
  - Failed: ~ 21.3

- <u>No CFG Constraints:</u>
- Coverage: ~0.99
- Exact Match: ~0.35
- Parseval f-score: ~0.84
- Sentence Length:
  - Avg # of tokens: ~ 14.0
  - Parsed: ~ 14.0
  - Failed: ~ 17.8

***Lexical types not evaluated**

Further Ways of Improving CF

- Sanity testing/bug fixing
- Deterministic approaches (cheap):
  - Improve feature model
  - Further experiments on partitioning of feature model
- Non-deterministic approach (expensive):
  - Back-tracking
- Unification validation using Pet (expensive)

References

- Rebecca Dridan. Using lexical statistics to improve HPSG parsing. PhD thesis, Saarland University, 2009
- Joakim Nivre, Jens Nilsson, Johan Hall, Atanas Chanev, Gulsen Eryigit, San- dra Kubler, Svetoslav Marinov, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. Natural Language Engineering, 13(1):1-41, 2007.