

HPSG Tutorial, Part 2

Grammar Implementation

Dan Flickinger

CSLI, Stanford

`lingo.stanford.edu/erg`
`moin.delph-in.net/ErgTop`

Desiderata for a linguistic theory

- General principles that hold true of human language
- Formal descriptive devices to make falsifiable predictions
- Cross-linguistic validity
- Simplicity



Desiderata for a grammar implementation

- Coverage of linguistic phenomena
- Accuracy of linguistic analyses
- Encoding of principles and generalizations
- Ambiguity tolerably constrained
- Reversibility (parsing and generation)



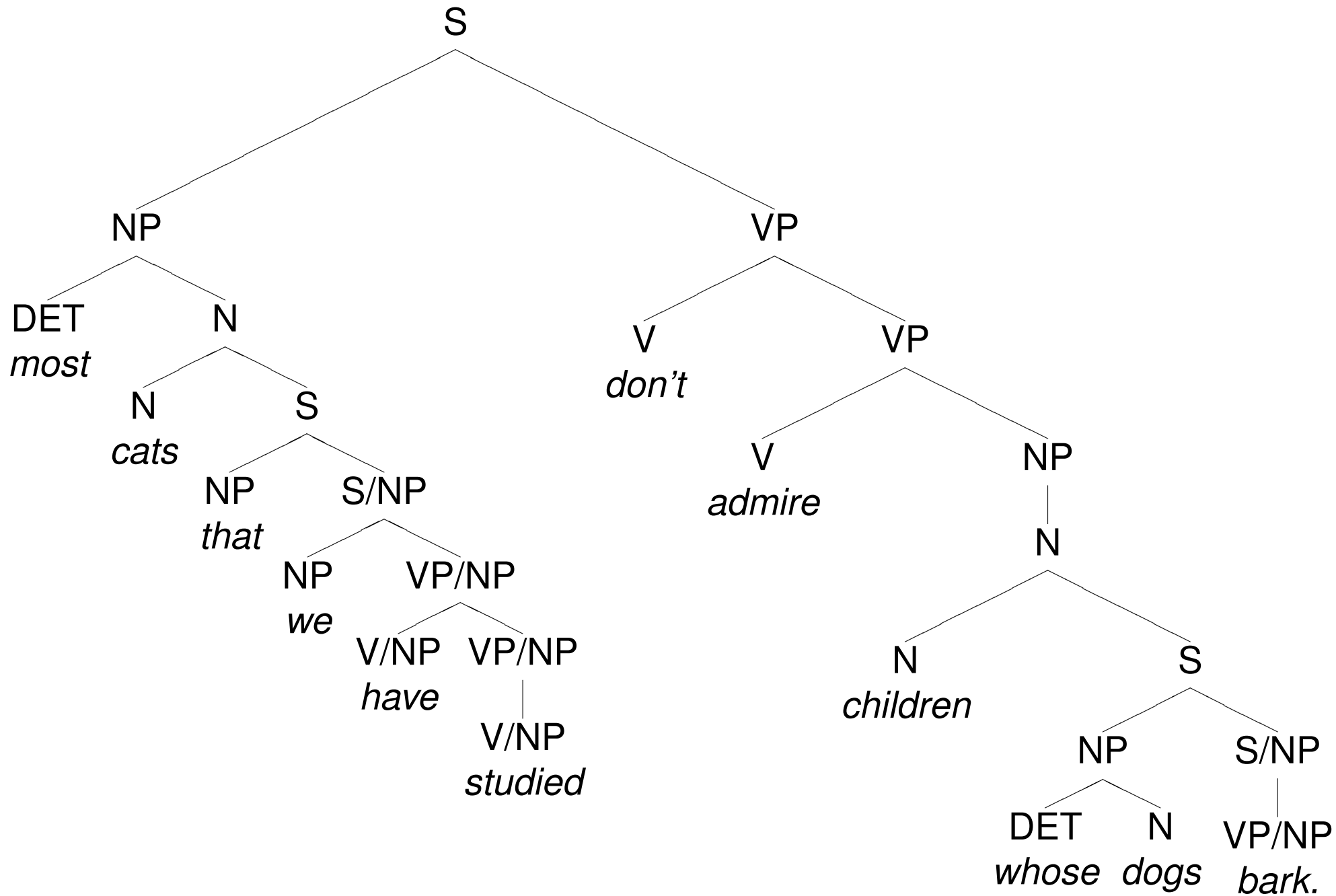
Desiderata for a grammar implementation

- Coverage of linguistic phenomena
- Accuracy of linguistic analyses
- Encoding of principles and generalizations
- Ambiguity tolerably constrained
- Reversibility (parsing and generation)

- Efficiency in processing
- Maintenance and extensibility



An example



An example: Semantic dependencies

Most cats that we have studied don't admire children whose dogs bark.

x4:_most_q[]

x9:pronoun_q[]

e13:_study_v_1[ARG1 x9:pron, ARG2 x4:_cat_n_1]

e16:neg[ARG1 e2:_admire_v_1]

e2:_admire_v_1[ARG1 x4:_cat_n_1, ARG2 x18:_child_n_1]

x18:undef_q[]

x26:def_explicit_q[]

e28:poss[ARG1 x26:_dog_n_1, ARG2 x18:_child_n_1]

e30:_bark_v_1[ARG1 x26:_dog_n_1]



An example: Semantic dependencies

Most cats that we have studied don't admire children whose dogs bark.

x4:_most_q[]

x9:pronoun_q[]

e13:_study_v_1[ARG1 x9:pron, ARG2 x4:_cat_n_1]

e16:neg[ARG1 e2:_admire_v_1]

e2:_admire_v_1[ARG1 x4:_cat_n_1, ARG2 x18:_child_n_1]

x18:undef_q[]

x26:def_explicit_q[]

e28:poss[ARG1 x26:_dog_n_1, ARG2 x18:_child_n_1]

e30:_bark_v_1[ARG1 x26:_dog_n_1]



An example: semantics with scope constraints

Most cats that we have studied don't admire children whose dogs bark.

jh1,e2,
h3:_most_q(x4, h5, h6),
h7:_cat_n_1(x4),
h8:pron(x9),
h10:pronoun_q(x9, h11, h12),
h7:_study_v_1(e13, x9, x4),
h14:neg(e16, h15),
h17:_admire_v_1(e2, x4, x18),
h19:udef_q(x18, h20, h21),
h22:_child_n_1(x18),
h23:def_explicit_q(x25, h26, h24),
h22:poss(e27, x25, x18),
h28:_dog_n_1(x25),
h22:_bark_v_1(e29, x25),
h5 qeq h7, h11 qeq h8, h15 qeq h17, h20 qeq h22, h26 qeq h28;



English Resource Grammar (ERG)

- 7000 types in multiple-inheritance monotonic hierarchy
- 975 leaf lexical types
- 39,000 manually constructed lexemes
- 225 syntactic rules
- 70 morphological rules (inflection and derivation)
- Statistical parse selection model trained on 1.5 million word corpus



English Resource Grammar (ERG)

- 7000 types in multiple-inheritance monotonic hierarchy
- 975 leaf lexical types
- 39,000 manually constructed lexemes
- 225 syntactic rules
- 70 morphological rules (inflection and derivation)
- Statistical parse selection model trained on 1.5 million word corpus
- Good analyses for 92% of newspaper text (WSJ used in Penn Treebank)



General-Purpose Syntactic Rules

- Head-Specifier
- Head-Complement
- Head-Modifier
- Head-Filler
- Coordination
- Head-Subject
- Head-Marker



Construction-Specific Rules

- Relative clauses

the book we tried to buy yesterday

- Compounds and appositives

Kim Smith, an ancient languages specialist

- Clause-modifying phrases headed by nouns or adjectives (depictives, absolutives, vocatives)

Their eyes as wide as saucers, the children gasped.

- Interleaving of complements and adjuncts

They said last week that we could stay.

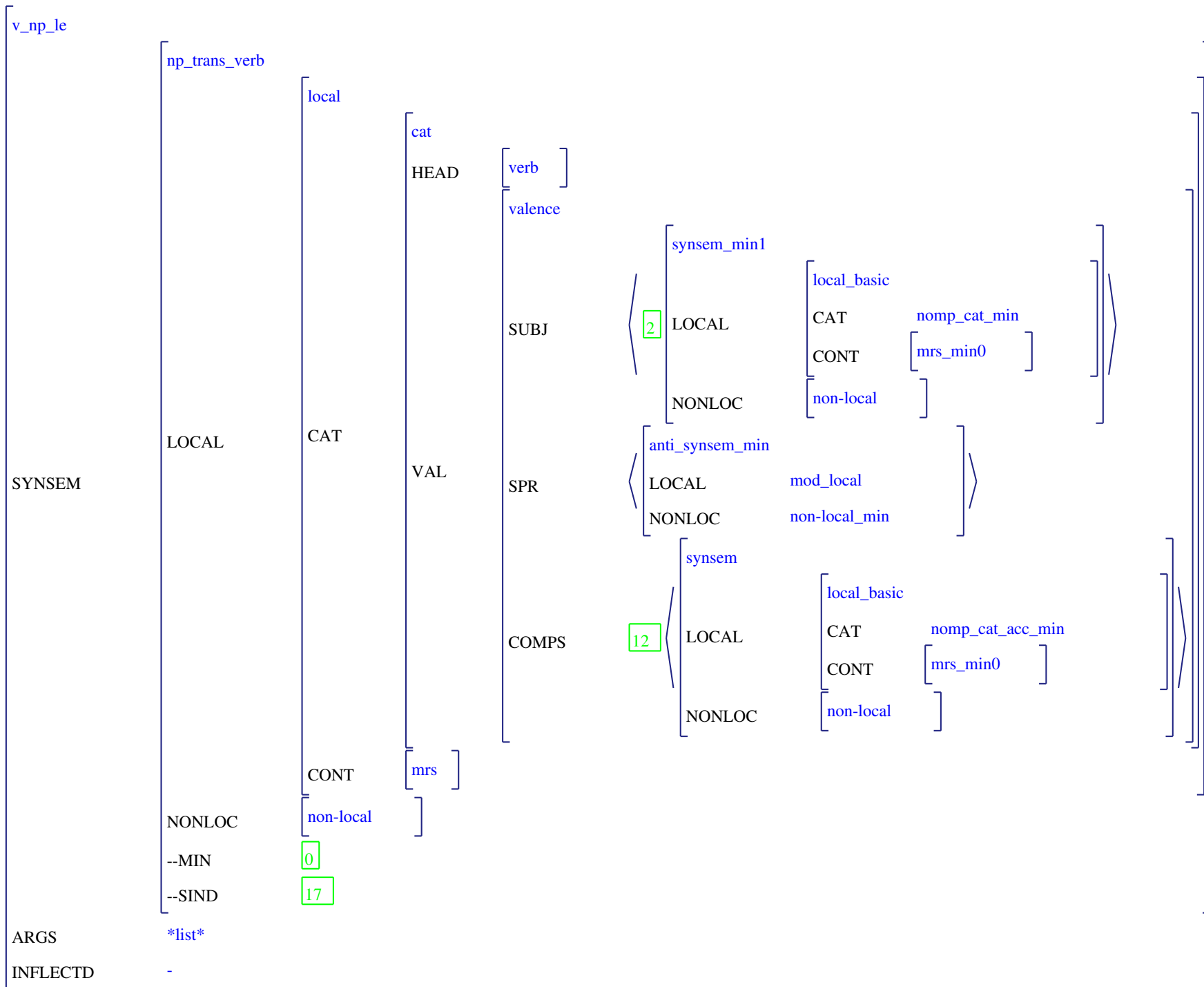
We'll see dolphins and if we're lucky some whales.

- Verbal gerunds

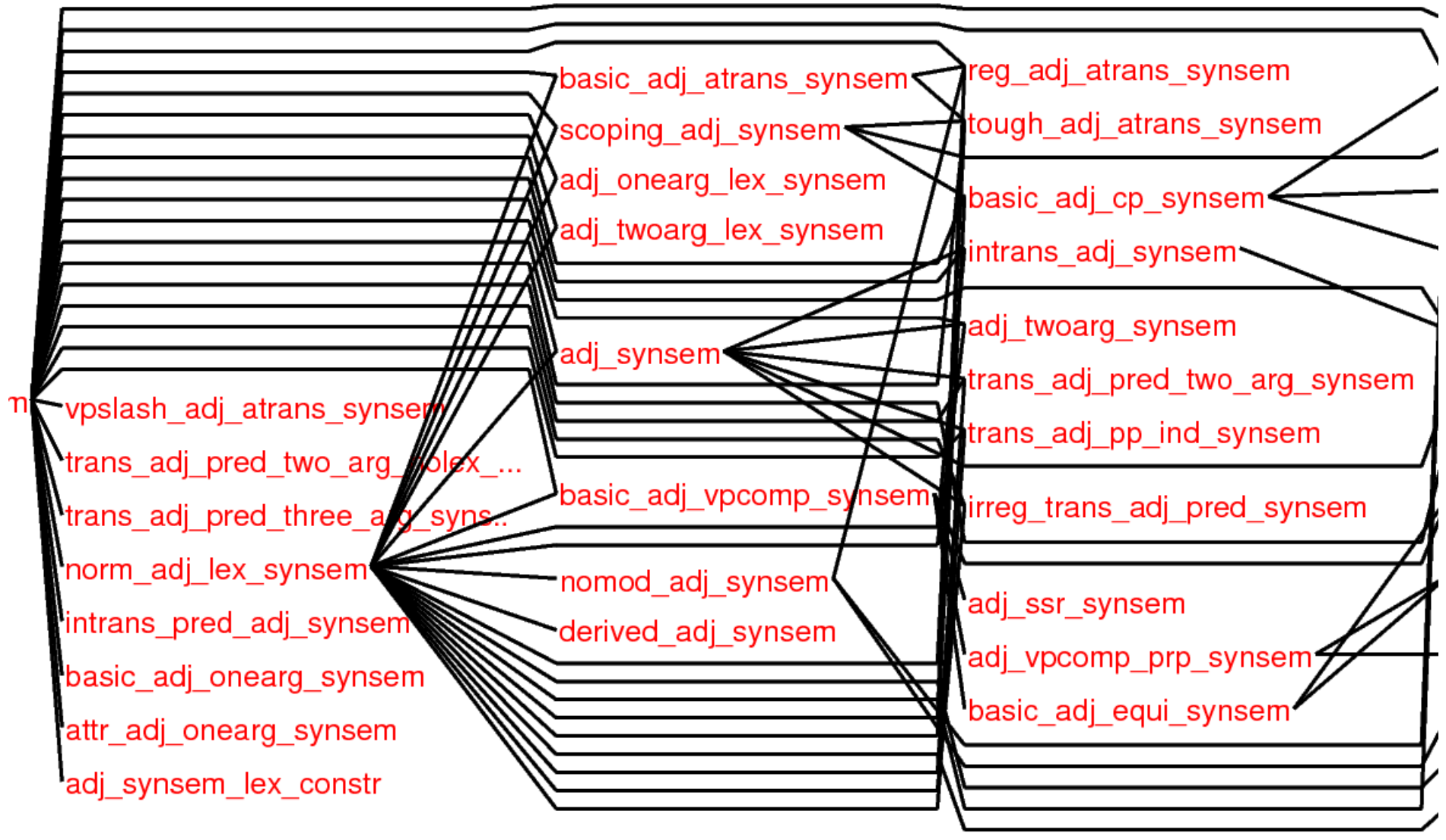
His habit of noisily slamming the door is well known



Sample feature structure



Sample type hierarchy: Adjectives

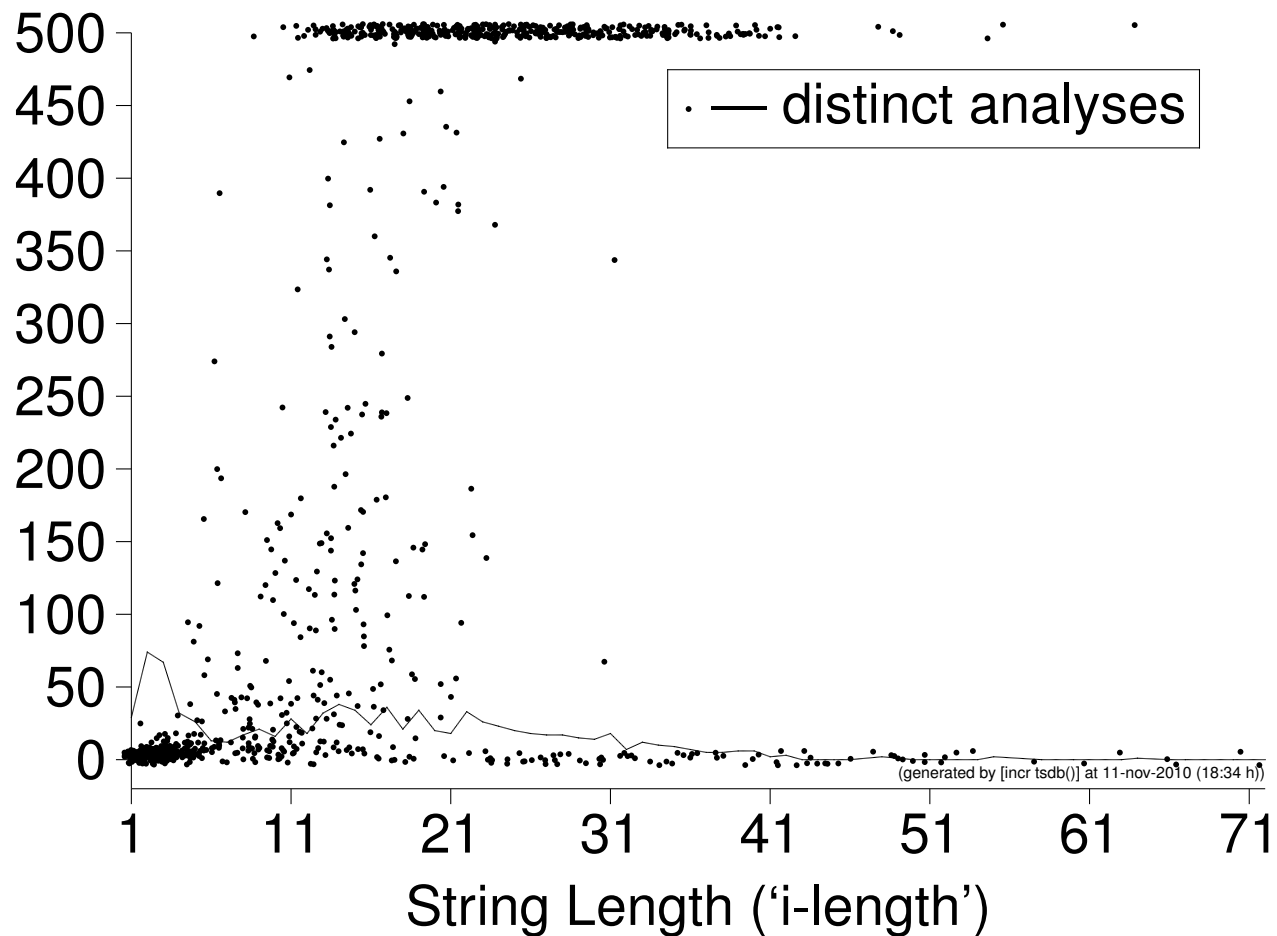




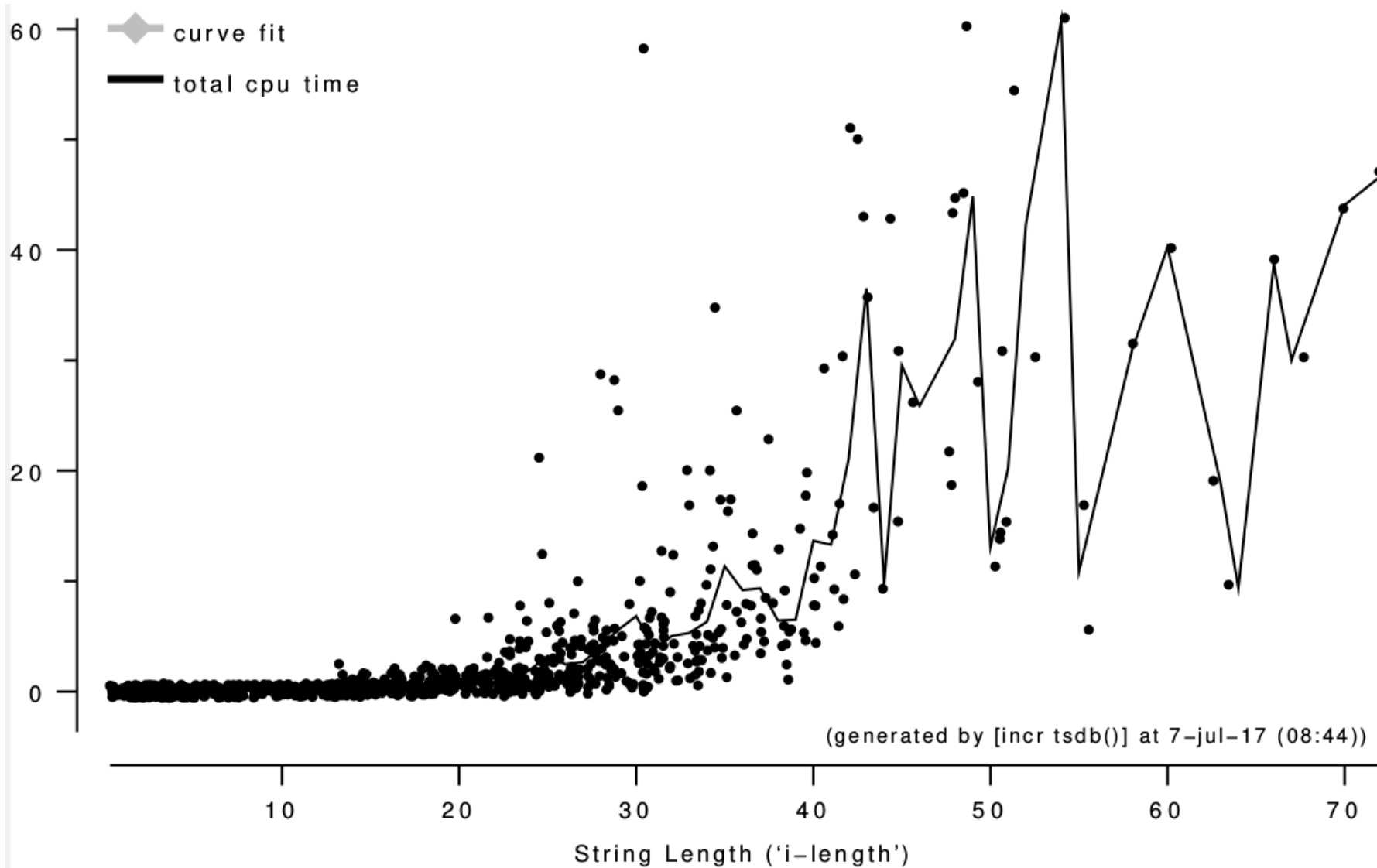
Wikipedia parsing with ERG/ACE: Coverage

'ws13' Coverage Profile						
Length in tokens	total items	word string	lexical items	distinct analyses	total results	overall coverage
65 – 70	2	66.50	755.00	0.00	0	0.0
60 – 65	3	62.33	538.67	500.00	1	33.3
55 – 60	3	56.00	343.33	500.00	2	66.7
50 – 55	8	51.37	563.50	0.00	0	0.0
45 – 50	11	47.27	542.55	500.00	3	27.3
40 – 45	21	41.38	422.90	500.00	11	52.4
35 – 40	42	36.69	433.64	500.00	32	76.2
30 – 35	77	31.78	321.78	490.26	61	79.2
25 – 30	101	26.89	329.47	499.68	87	86.1
20 – 25	130	22.17	220.02	465.31	120	92.3
15 – 20	158	16.98	177.37	374.31	149	94.3
10 – 15	137	12.32	126.58	197.77	132	96.4
5 – 10	97	6.86	67.01	42.46	89	91.8
0 – 5	209	2.47	11.02	3.42	202	96.7
Total	1001	17.55	184.78	270.04	889	88.8

Wikipedia parsing with ERG/ACE: Ambiguity



Wikipedia parsing with ERG/ACE: Efficiency



Wikipedia parsing with ERG/ACE: Accuracy

Corpus type	Number of items	Av. item length	Observed coverage	Verified coverage
Meeting scheduling	11660	7.5	96.8%	93.8%
E-commerce	5392	8.0	96.1%	93.0%
Norwegian tourism	10834	15.0	94.3%	88.5%
SemCor (partial)	2501	15.0	94.3%	88.5%
Newspaper (WSJ)	31441	20.4	93.4%	84.9%
Wikipedia (CmpLng)	11558	19.5	92.9%	81.7%
Online user forum	578	12.5	85.5%	77.5%
Dictionary defs.	10000	6.0	81.2%	75.5%
Essay	769	21.6	83.2%	69.4%
Chemistry papers	637	27.0	87.8%	65.3%
Technical manuals	4000	12.5	86.8%	61.9%



Frequency of Linguistic Phenomena in Wikipedia 100

Phenomenon	#Items	%Corpus
Measure NPs	44	0.3
Appositives	1048	9.1
NP Fragments	2126	18.4
NP Coordination	1960	17.0
Multi-NP Coord	558	4.8
VP Coordination	491	4.2
S Coordination	381	3.3
Relative Clauses	2239	19.4
Long-Dist. Deps	2273	19.7
Yes-No Questions	11	0.1
WH Questions	10	0.1
Imperatives	222	1.9
Free relatives	107	0.9
Passives	3534	30.6

