Implementing a theory at scale: the English Resource Grammar at 30

Dan Flickinger

HPSG 2024

Palacký University, Olomouc

Introduction to the ERG

- Origins: CSLI, Verbmobil (Wahlster 2000)
- Framework: HPSG (Pollard and Sag 1994)
- Contributors: Rob Malouf, Emily Bender, Stephan Oepen, Ivan Sag, Ann Copestake, Tom Wasow
- Platforms: DISCO, PAGE, PET (DFKI), LKB (Cambridge), ACE (Stanford), LKB-FOS (Sussex)
- Implementation formalism: TDL (Type Description Language)
  Rich type hierarchy, no defaults, few relational constraints
- Semantics: Minimal Recursion Semantics

Additions to core HPSG framework

- Lexicon (expanding on Flickinger 1987)
- Relative clauses (Sag 1997)
- Coordination (consistent with Chaves 2007)
- Punctuation (following Numberg 1990, Briscoe 2002 in H&P)
- Semantics: MRS Copestake et al. 2005
- Annotated corpora: Verbmobil, email, tourism, Wikipedia, WSJ

Syntactic Rule Schemata

- Head-Subject
- Head-Complement
- Head-Specifier
- Head-Adjunct
- Head-Filler
- Head-Marker

Syntactic Rule Schemata, expanded

- Head-Subject
- Head-Complement (re-ordering, optionality, extraction)
  - Head-Comp-2: *given to Kim by Pat* vs. *given by Pat to Kim*
  - Head-Optional-Comp (unary rule)
- Head-Specifier (semantic composition, optionality)
  - Semantic head: *every book* vs. *very tall*
- Head-Adjunct (ordering)
  - *expensive books* vs. *books more expensive than this*
- Head-Filler
  - WH vs non-WH in matrix clauses (inverted or not)
  - Relative clauses
- Head-Marker (only for coordination)

More constructions

- Constituent coordination
- Compounds: *hotel reservation*, *IBM report*, *tree-lined*, *dog-friendly*, *angry-looking*, *best-performing*, *long-eared*
- Apposition: *my neighbor, a famous doctor, ...*
- NP adverbials: *this month*, *the week before he arrived*
- Measure phrases: *fell three meters*, *five-meter tall tree*
- Partitives: *several of the books*, *the most beautiful of the flowers*
- Verbal gerunds: *we objected to your asking them to stay*
- Right-node raising: *They wrote about _ and admired _ that leader.*
- Sub-clausal modifiers *Angry at everyone, they stormed out.*
- Sentence fragments: *Just one, please. Not that one.*

300 syntactic constructions in the ERG

- 20 head-valence rules (subject, complement, specifier)
- 45 head-adjunct rules
- 15 filler-head rules
- 30 compounding rules
- 15 apposition rules
- 40 cagegory-changing rules (gerunds, NP-adverbials, subord)
- 70 coordination rules
- 35 fragment rules
- 30 other rules (RNR, parentheticals, run-ons, numbered items)

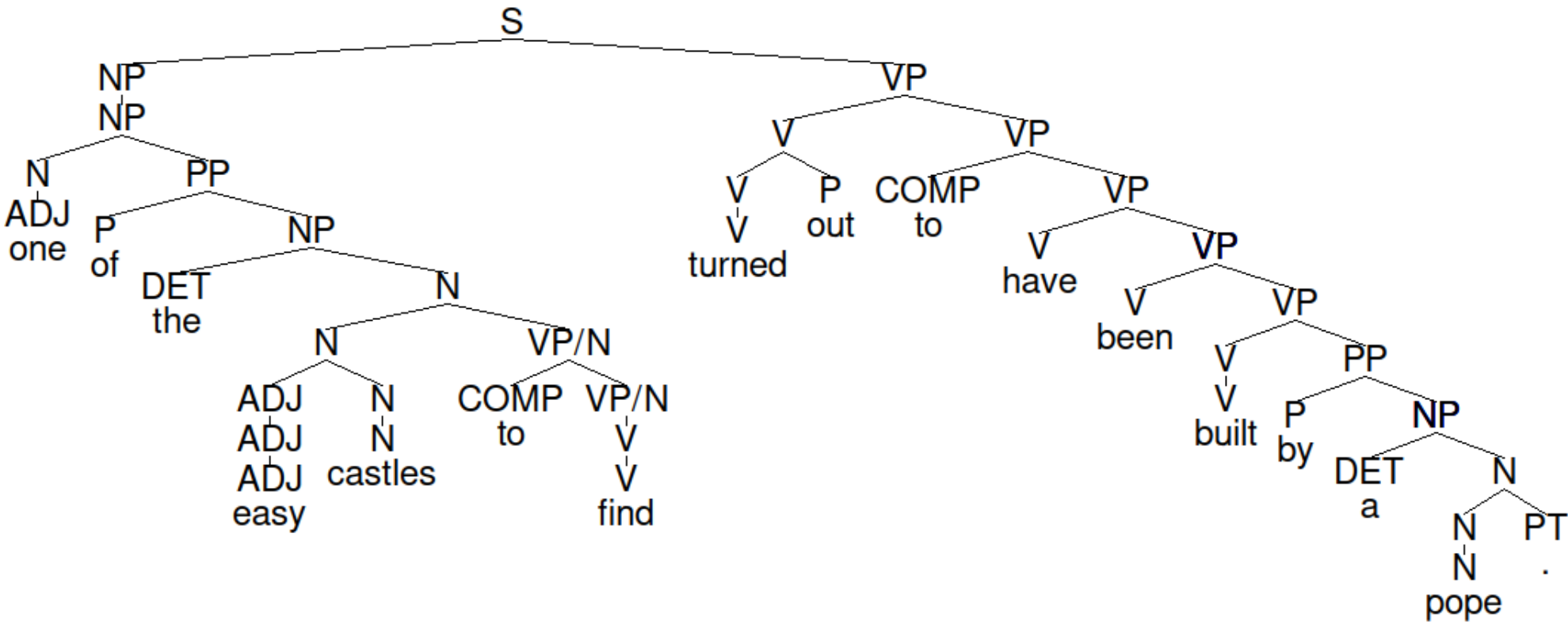
1400 Lexical Types in the ERG

Manually constructed lexicon of 45,000 entries

- 470 verb types (ditransitives, raising, verb-particle ...)
- 375 noun types (count/mass, relational, VP/CP complements ...)
- 145 adjective types (attrib/pred, comparative/superlative ...)
- 125 adverb types (position in VP, degree modification ...)
- 120 preposition types
- 75 conjunction types
- 50 determiner types
- 15 complementizer types
- 25 miscellaneous types (*a.m./p.m., BC/AD, how about, so*)

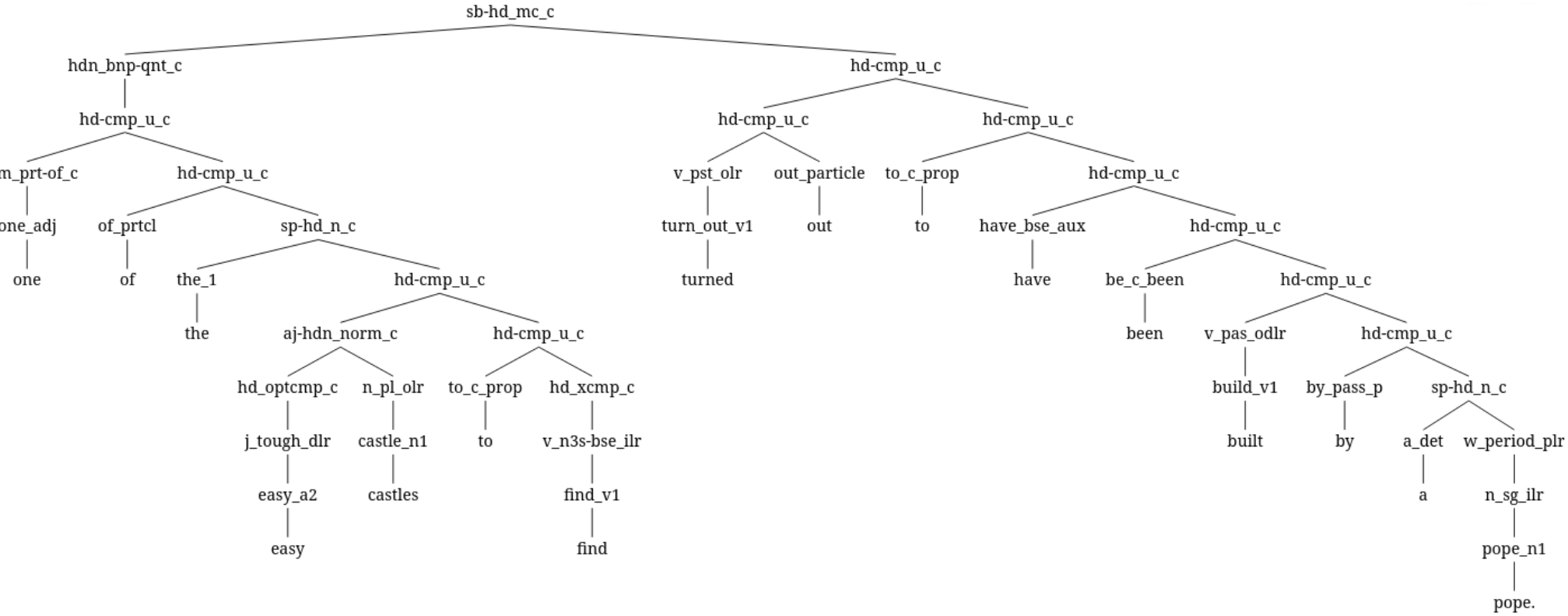
Sample parse

One of the easy castles to find turned out to have been built by a pope.



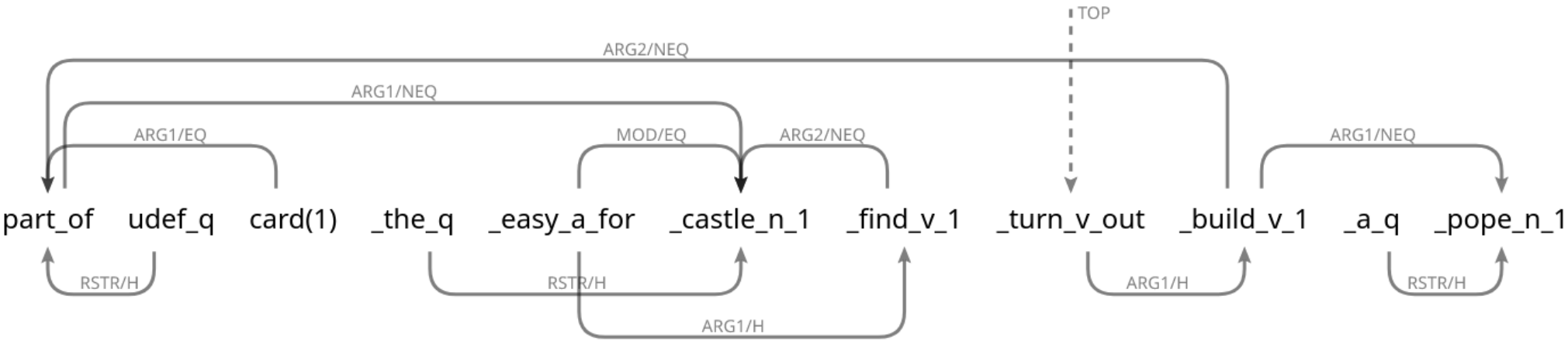
Sample derivation tree

One of the easy castles to find turned out to have been built by a pope.



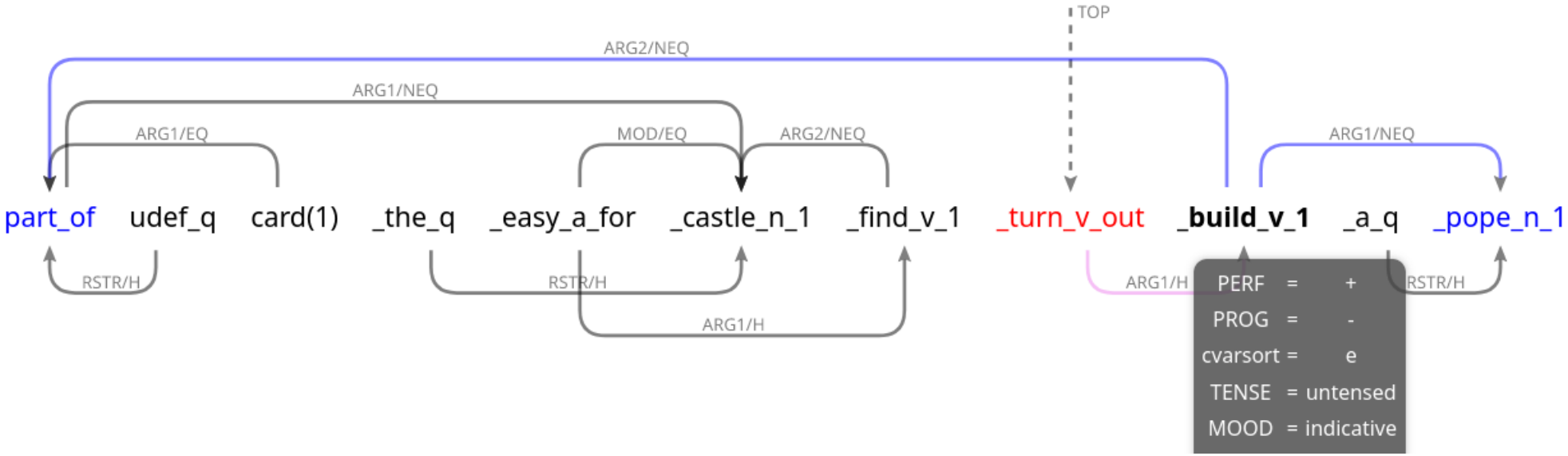
Dependency MRS

One of the easy castles to find turned out to have been built by a pope.



Dependency MRS

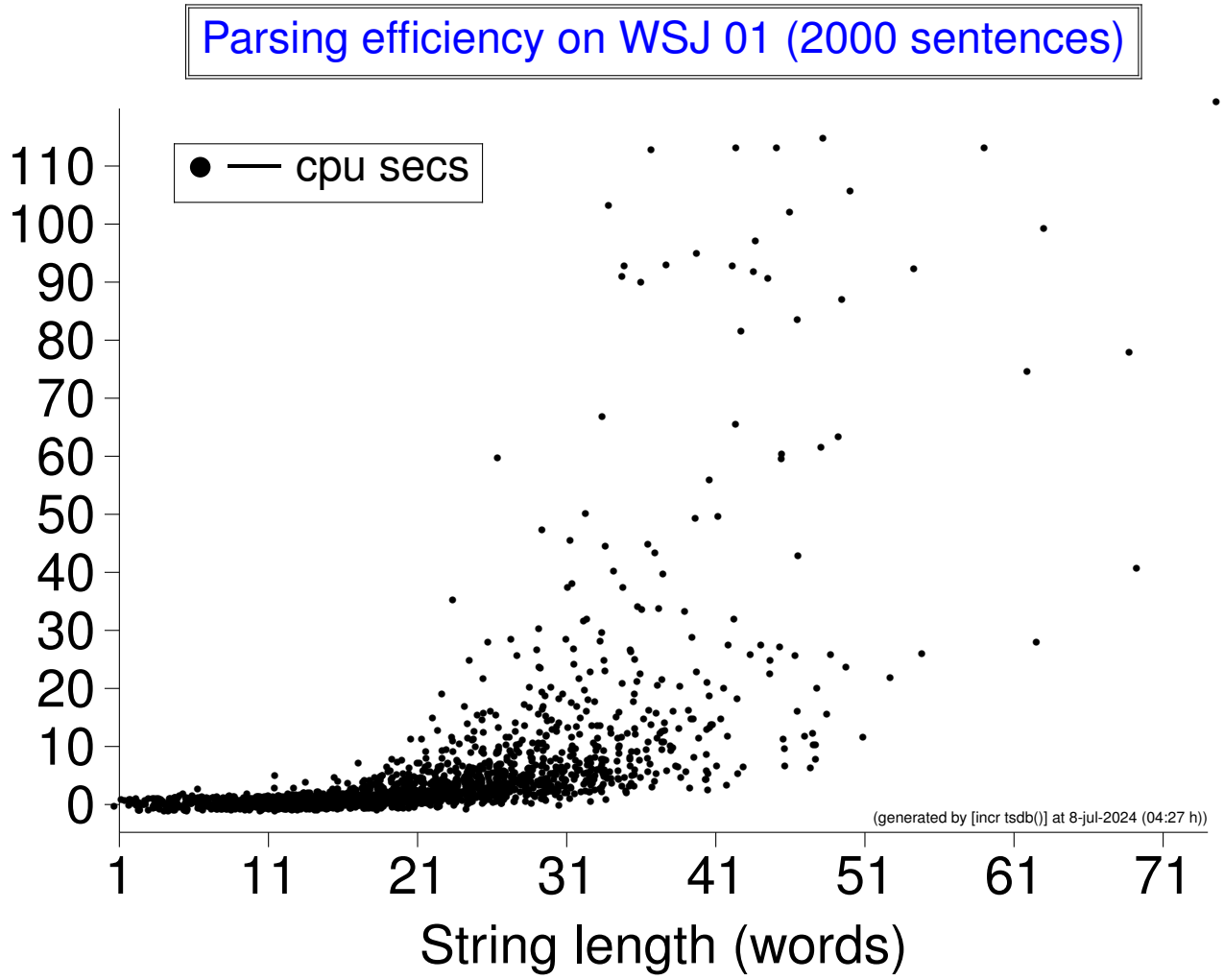
One of the easy castles to find turned out to have been built by a pope.



Parsing coverage of WSJ Section 01

Length	items $\sharp$	words ϕ	results $\sharp$	coverage %
$66 \leq 70$	2	68.50	1	50.0
$61 \leq 65$	3	62.33	2	66.7
$56 \leq 60$	3	56.33	1	33.3
$51 \leq 55$	5	50.60	3	60.0
$46 \leq 50$	26	46.81	20	76.9
$41 \leq 45$	41	41.54	31	75.6
$36 \leq 40$	80	36.53	71	88.8
$31 \leq 35$	191	31.68	185	96.9
$26 \leq 30$	272	26.86	262	96.3
$21 \leq 25$	397	21.87	384	96.7
$16 \leq 20$	406	17.17	397	97.8
$11 \leq 15$	313	12.17	304	97.1
$6 \leq 10$	195	7.21	192	98.5
$1 \leq 5$	58	3.07	58	100.0
Total	**1993**	**20.60**	**1911**	**95.9**

(generated by [incr tsdb()] at 8-jul-2024 (04:06 h))


Parsing efficiency on WSJ 01 (2000 sentences)
cpu secs
110
100
90
80
70
60
50
40
30
20
10
0
(generated by [incr tsdb()] at 8-jul-2024 (04:27 h))
1
11
21
31
41
51
61
71
String length (words)

Redwoods Treebank

Resource	# Sentences
Verbmobil spoken dialogues	12393
E-commerce emails	5793
Norwegian tourism	11596
Eric Raymond essay	769
Brown corpus	5420
Wikipedia	11556
Wall Street Journal	46162
Tanaka English learner data	3000
Open Multilingual WordNet	1545
Sherlock Holmes story	599
Cambridge Grammmar examples	15372
Total	114205

Constructions used in Redwoods

114,000 sentences, 1.5M words

Head-Complement	557,081
Bare-NP	348,887
Head-Adjunct	251,164
Head-Opt-Comp	202,832
Head-Specifier	199,267
Head-Subject	140,784
Coordination	51,451
Head-Filler	35,849
...	...

Constructions used in Redwoods

114,000 sentences, 1.5M words

Head-Complement	557,081	
Bare-NP	348,887	
Head-Adjunct	251,164	
Head-Opt-Comp	202,832	
Head-Specifier	199,267	
Head-Subject	140,784	
Coordination	51,451	
Head-Filler	35,849	
...	...	
Parenthetical-S	1632	*I read a book (you'd like it) today.*
Right-Node-Raise	992	*They wrote about _ and admired _ that leader.*
Vocative-NP	244	*People in back, we can't hear you.*
VP-Gerund-Gap	24	*It was too big to imagine taking home.*
Year-Name	2	*It began on January eleventh, two thousand three.*

All 300 constructions used at least once
30 constructions used fewer than 10 times

Lexical types used in Redwoods

114,000 sentences, 1.5M words

d_-_the_le	99,385
p_np_i_le	94,290
v_np_le	93,661
n_-_c_le	78,919
n_-_mc_le	63,728
n_-_pn_le	53,951
aj_-_i_le	44,827
...	...

Lexical types used in Redwoods

114,000 sentences, 1.5M words

d_-_the_le	99,385	
p_np_i_le	94,290	
v_np_le	93,661	
n_-_c_le	78,919	
n_-_mc_le	63,728	
n_-_pn_le	53,951	
aj_-_i_le	44,827	
...	...	
v_np-rc_be_le	304	*It* **s** *you who should apologize!*
d_-_sg-caj_le	78	*That was too strong* **a** *drink to enjoy.*
v_vp_do-be_le	55	*What we had them do* **was** *make coffee.*
v_np-np-vpslnp_tgh_le	16	*This paper* **took** *me hours to read.*

1158 of 1400 lexical types used at least once
277 lexical types used fewer than 10 times

Cambridge Grammar of the English Language

- Most comprehensive text on English grammar to date
  - Published in 2002, by Rodney Huddleston and Geoffrey Pullum
  - Morphology, syntax, punctuation
  - 1800 pages, 20 chapters
- Rich in examples, both positive and negative
  - 15,000+, averaging about 10 per page
- Compatible with HPSG/ERG in its theoretical assumptions

Rules and lexical types in CGEL derivations

- Used 932 of the 1400 lexical types in ERG
- Example of unused lexical type:

  v_p-cp_it-s_le: *It matters a lot to Kim that the cat disappeared.*

- Used 296 of the 402 rules (syntactic and lexical)
- Examples of unused syntactic rules:

  j-v_j-cpd_c: *an angry-looking cat*

  flr-hd_nwh-inv-nmc_c:

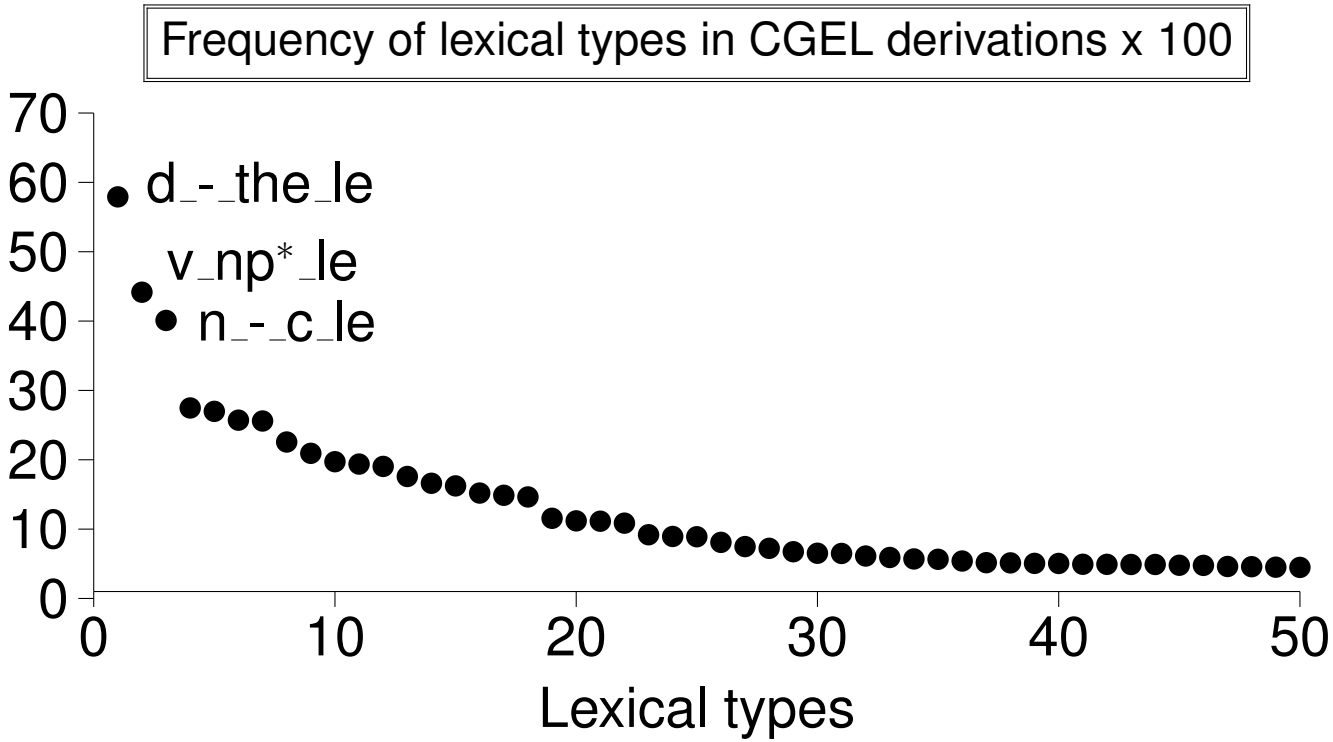
  He claimed that only yesterday did they finally arrive.

  n-j_crd-m_c: *the marble and wooden stairs*

- Examples of unused lexical rules

  j_tough-compar_dlr: *Kim is tougher to admire than Pat.*

  v_pas-p-t_odlr: *Our bill has been added to.*


Frequency of lexical types in CGEL derivations x 100
70
60
50
40
30
20
10
0
d_-_the_le
v_np*_le
n_-_c_le
0
10
20
30
40
50
Lexical types

Evaluation of ERG coverage of CGEL examples

Length	total items ♯	positive items ♯	word string ϕ	total results ♯	overall coverage %
$51 \leq 55$	1	1	50.00	1	100.0
$46 \leq 50$	5	5	48.20	3	60.0
$41 \leq 45$	9	9	41.78	5	55.6
$36 \leq 40$	17	17	36.41	16	94.1
$31 \leq 35$	34	33	31.64	25	75.8
$26 \leq 30$	104	98	26.34	75	76.5
$21 \leq 25$	221	211	21.72	175	82.9
$16 \leq 20$	562	521	16.31	461	88.5
$11 \leq 15$	3265	2985	11.56	2704	90.6
$6 \leq 10$	8484	7770	6.70	7494	96.4
$1 \leq 5$	2670	2500	3.53	2446	97.8
Total	**15372**	**14150**	**8.01**	**13405**	**94.7**

(generated by [incr tsdb()] at 29-jun-2024 (21:04 h))

Examples of CGEL phenomena missing in ERG

- Correlative comparatives

  The harder the task, the more she relished it.

- Gapping

  I gave $10 to Kim and $5 to Pat.

  Kim wasn't at work on Monday or Pat on Tuesday.

- Imperatives with subjects

  Nobody move.

  Somebody get me a screwdriver.

- Asymmetric coordination

  He'll reject it because it's too long or for some other reason.

- Topic + sentence

  The other one, they don't think she'll survive.

  Garlic, I eat it and pretty soon my stomach's upset.

Examples of ERG phenomena missing in CGEL

- *do-be* construction (1 sentence per 3000 in COCA)

  The best thing to do is buy a new bicycle.

  All we can do this year is hope for a better candidate.

  What we think he should not do is get a new car.

- Specifiers of specifiers and adverbs

  much more important

  **very more important*

  not very much more important

  This problem was somewhat more quickly solved than yours.

Next steps

- Attending to missing coverage for CGEL sentences
  Gapping, correlative comparatives, imperatives with subjects
- Aligning ERG lexicon with Open English WordNet (150K entries)
  Coarse-grained semantic senses, more robust generation
- Expanding inventory of idioms
- Reparsing all of English Wikipedia (5 billion words)
- Improving treatment of dialect variation (UK/OZ, Singlish)
- Extending mal-grammar for robustness, language learners