Linking the ERG to the Cambridge Grammar of the English Language

Dan Flickinger

DELPH-IN Summit 2024

Olomouc

1 July 2024

Why the Cambridge Grammar

- 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 ERG in its theoretical assumptions

  Pullum was co-developer of GPSG, precursor to HPSG

Goals for CGEL data

- **Evaluation** of the ERG

  Coverage of linguistic phenomena found in CGEL

  Overgeneration

- **Planning** for further grammar development

  Which phenomena in CGEL remain unanalyzed in the ERG?

- **Linguistic advances in ERG**

  Which phenomena analyzed by ERG are missing in CGEL?

- **Documentation** of linguistic analyses

  Linking ERG rules and lexical types to CGEL descriptions

Some history

- 1987 Hewlett-Packard NLP test suite presented at CSLI
- 1994 ERG development began at Stanford
- 2002 CGEL was published
- 2017 Ned Letcher collaborated on ERG/CGEL links using Typediff
- 2023 Pullum gave all CGEL examples to Nathan Schneider
- 2024 ERG was evaluated on all 15,372 examples in CGEL

  Parsed with ACE using soon-to-be-released ERG version 2024

  Treebanked with FFTB

Using the CGEL example data

- GitHub repository provides full set of examples in several formats
- One example in CGEL may correspond to several sentences

  a. *They were eating/drinking/*devouring.* ⇒

  They were eating.

  They were drinking.

  **They were devouring.*

  b. *... got their results: all/both (of them) had passed.* ⇒

  ... all of them had passed..

  ... all had passed..

  ... both of them had passed..

  ... both had passed..
- Manual curation resulted in an ‘item’ file of 15,372 examples

  14,260 well-formed, 1,222 ill-formed

Evaluation of ERG coverage of CGEL examples

‘24-06-18/ace’ Coverage Profile

	total	positive	word	total	overall
Length	items	items	string	results	coverage
	$\sharp$	$\sharp$	ϕ	$\sharp$	$\%$
50 < 55	1	1	50.00	1	100.0
45 < 50	5	5	48.20	3	60.0
40 < 45	9	9	41.78	5	55.6
35 < 40	17	17	36.41	16	94.1
30 < 35	34	33	31.64	25	75.8
25 < 30	104	98	26.34	75	76.5
20 < 25	221	211	21.72	175	82.9
15 < 20	562	521	16.31	461	88.5
10 < 15	3265	2985	11.56	2704	90.6
5 < 10	8484	7770	6.70	7494	96.4
0 < 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

  The best thing to do is buy a new bicycle.

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

- Specifiers of specifiers and adverbs

  much more *important*

  **very more important*

  This problem was **more quickly** *solved than yours was.*

Linking ERG analyses to CGEL phenomena

- Annotate each item in profile with page number in CGEL
- Enrich derivation trees in profile with lexical type names
- Extract all rule names and lexical types from derivations
- For each rule and le-type, collect all pages using it in an example
- Gather frequencies in CGEL for each rule and lexical type

Rules and lexical types in CGEL derivations

- Used 932 of the 1423 lexical types in ERG 2024
- 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*

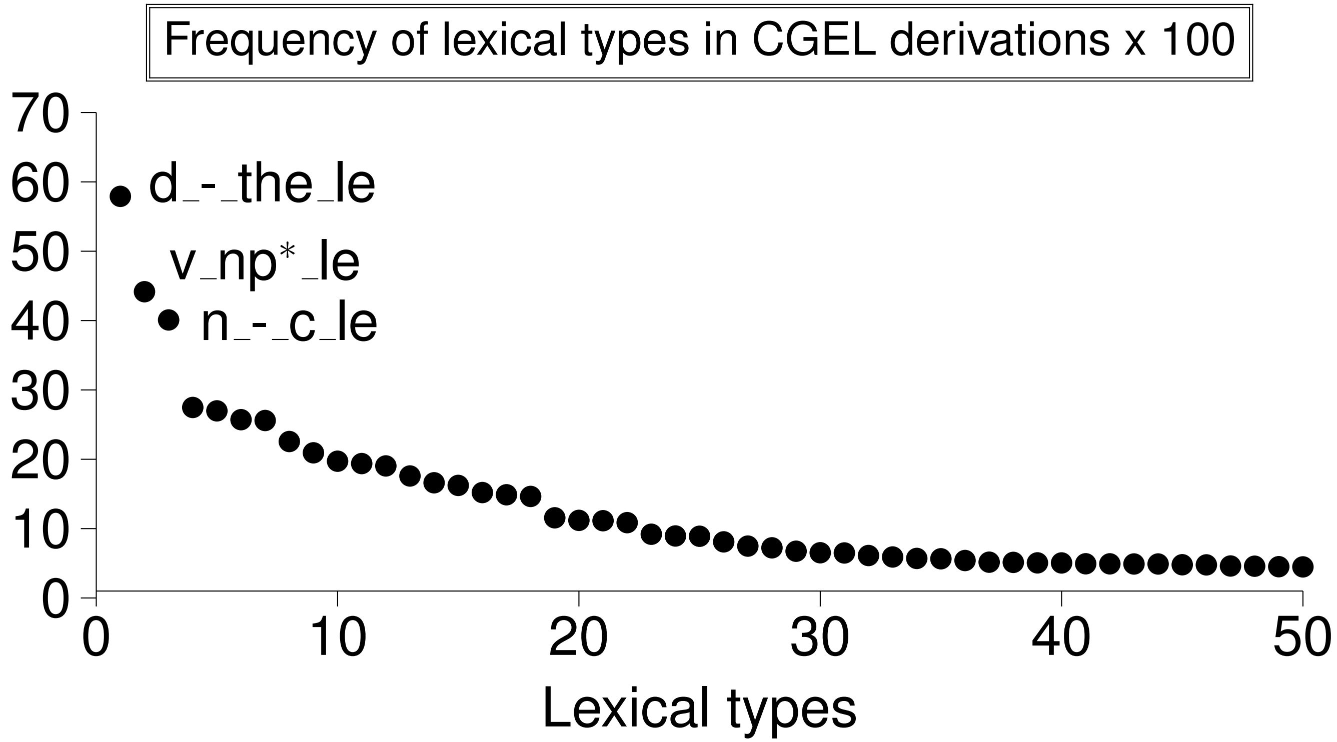
  flr-hd_nwh-inv-nmc_c: *He claimed that* **only yesterday did they finally arrive***.*

  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

Using the ERG-CGEL mapping

- Parse a sentence exhibiting some construction of interest, 1-best
- Extract rules and le-types from the derivation tree
- Sort by CGEL frequency, and report CGEL pages for rarest sign
- Ideally (but not yet), for each rule/type, identify the canonical pages in CGEL discussing the associated phenomenon

Demo of CGEL-ERG indexing search

```
Everyone admires and respects that professor.
NOTE: 1 readings, added 2265 / 398 edges to chart (153 fully instantiat
NOTE: parsed 1 / 1 sentences, avg 5861k, time 0.02915s
hd-hd_rnr_c 500 800 813 1001 1044 1286 1320 1323 1343 1344 1424 1548

What we should really do is make an effort to present a really complica
NOTE: 1 readings, added 7621 / 3046 edges to chart (884 fully instantia
NOTE: parsed 1 / 1 sentences, avg 37652k, time 0.36174s
v_vp_do-is_le 1422

That was too easy a problem for her.
NOTE: 1 readings, added 4005 / 1511 edges to chart (607 fully instantia
NOTE: parsed 1 / 1 sentences, avg 14991k, time 0.11706s
d_-_sg-caj_le 61 62 350 433 435 443 529 540 551 634 910 920 923 967 108
```

Next steps

- For each rule/type, manually identify the canonical page(s)
- For each page, report the section header (phenomenon) in CGEL

  Documentation of ERG rule/type names
- Persuade someone to set up a web server running this process