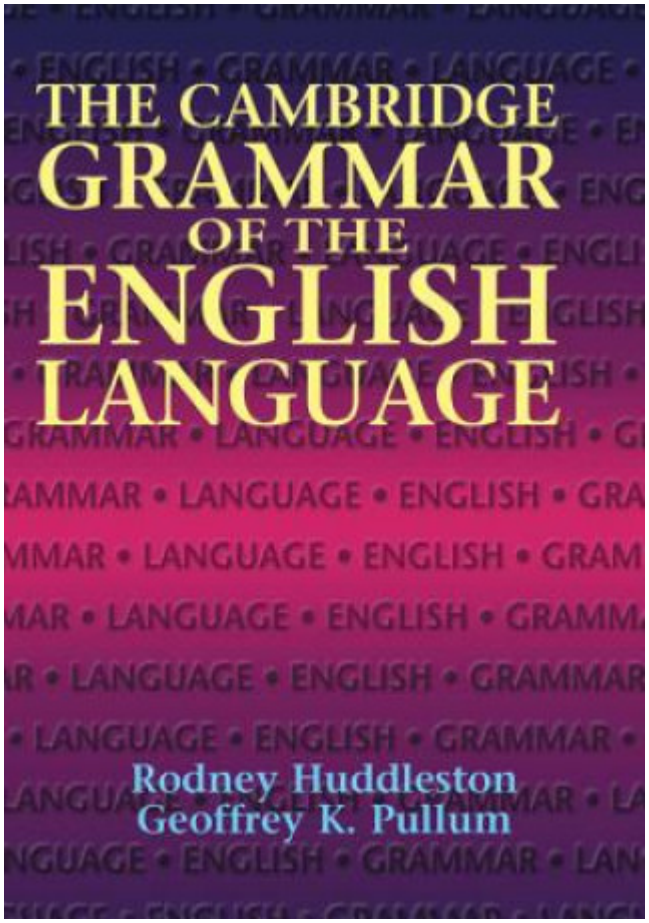
Linking the ERG to the Cambridge Grammar of the English Language

Dan Flickinger

DELPH-IN Summit 2025

Amsterdam

8 July 2025


THE CAMBRIDGE
GRAMMAR
OF THE
ENGLISH
LANGUAGE
Rodney Huddleston
Geoffrey K. Pullum

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 ERG in its theoretical assumptions
  - Pullum was co-developer of GPSG, precursor to HPSG

Benefits of linking to CGEL

- Documentation of linguistic analyses in ERG

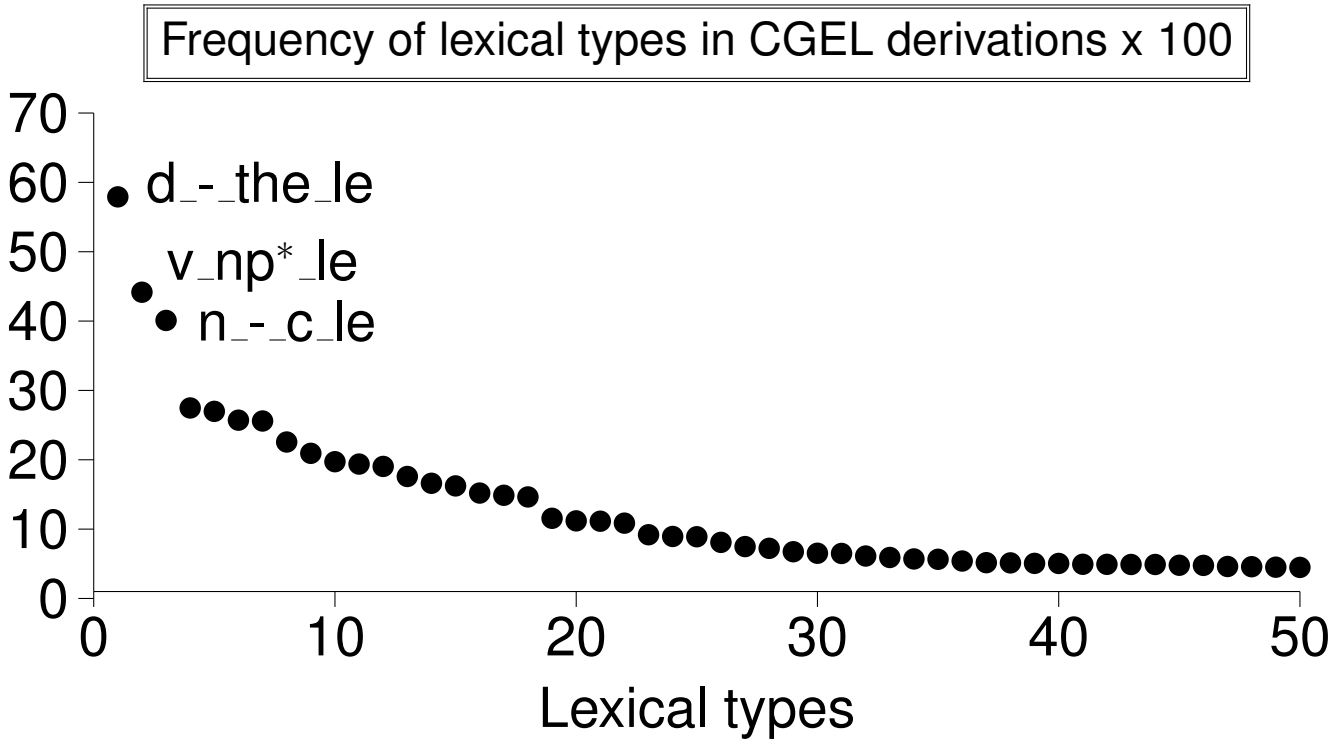
  Linking ERG rules and lexical types to CGEL descriptions

- Finding relevant sections in CGEL by parsing sentence with ERG

  ERG lexical types and rules are anchored to CGEL pages

Using the CGEL example data

- GitHub repository provides full set of examples in several formats
- Manual curation resulted in an 'item' file of 15,372 examples
  14,260 well-formed, 1,222 ill-formed
- Annotated each item in profile with page number in CGEL
- Extracted all rule names and lexical types from derivations
- For each rule/le-type, collected all pages using it in an example
- Gathered frequencies in CGEL for each rule and lexical type


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.
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
complicated sentence.
v_vp_do-is_le 1422

That was too easy a problem for her.
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 in LTDB
- Set up a web server running this process