Topic and Genre Classification of a Large English Web Corpus

Vít Suchomel, Jan Kraus
name.surname@sketchengine.eu

e-Lex 2023, Brno, Czechia

Understanding the Content of Web Corpora

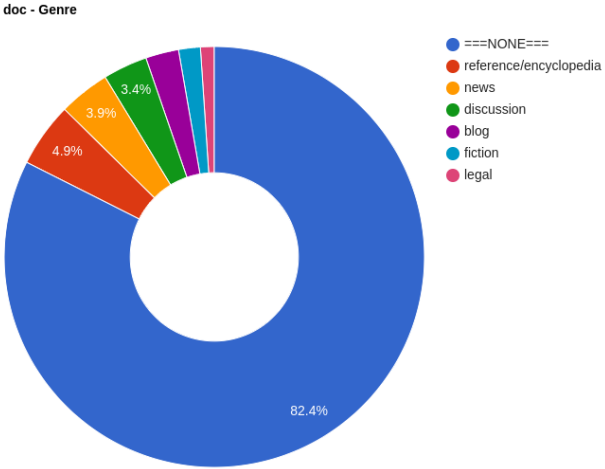
- Corpora from books, newspapers, magazines,...: rich metadata, controlled content
- Web corpora? Which texts are inside?

Understanding the Content of Web Corpora

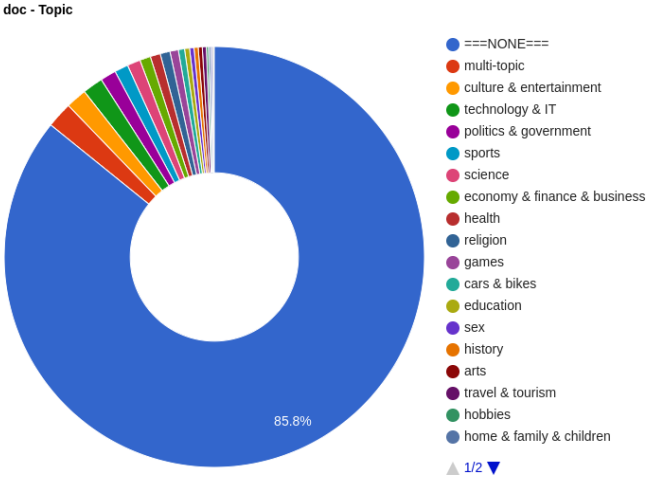
- Corpora from books, newspapers, magazines,...: rich metadata, controlled content
- Web corpora? Which texts are inside?

1. author
2. date of publishing
3. language variety
4. **text genre**
5. **text topic**

Tokens by Genre in enTenTen21

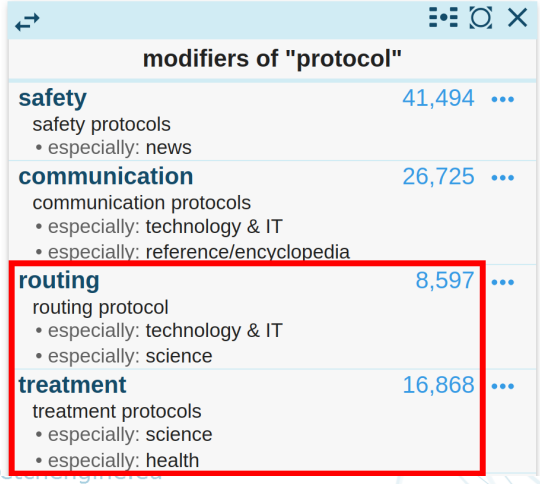


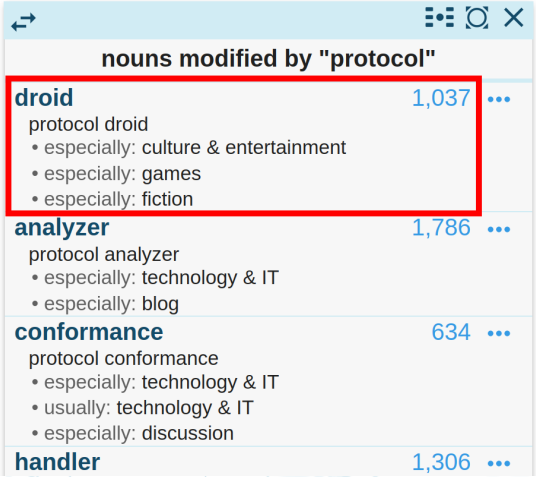
Tokens by Topic in enTenTen21



Word Sketches – Topic-specific and Genre-specific Collocations

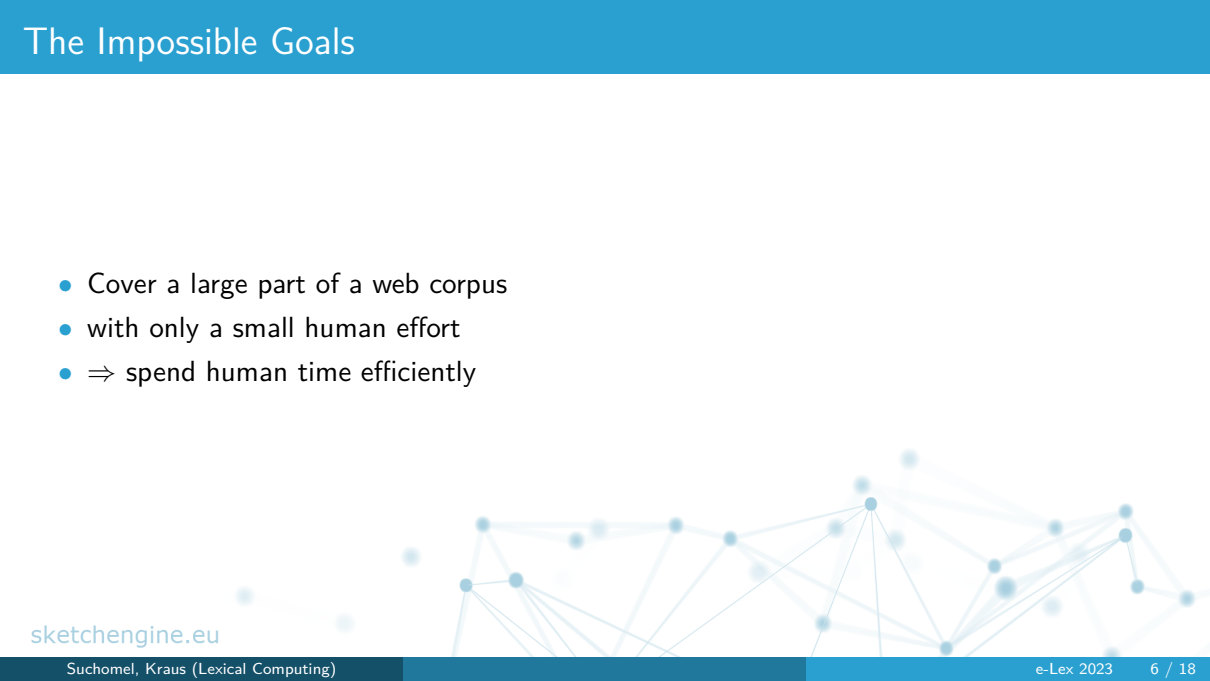
protocol as noun 1,861,633×





The Impossible Goals

- Cover a large part of a web corpus
- with only a small human effort
- $\Rightarrow$ spend human time efficiently



Website Homogeneity Assumption

- Assign the same topic/genre label for all pages from a website
- Holds for topics in 92 % of cases [Papčo, 2022]
  - based on a 12-topic scheme in enTenTen20

Topic and Genre Annotation of Whole Websites

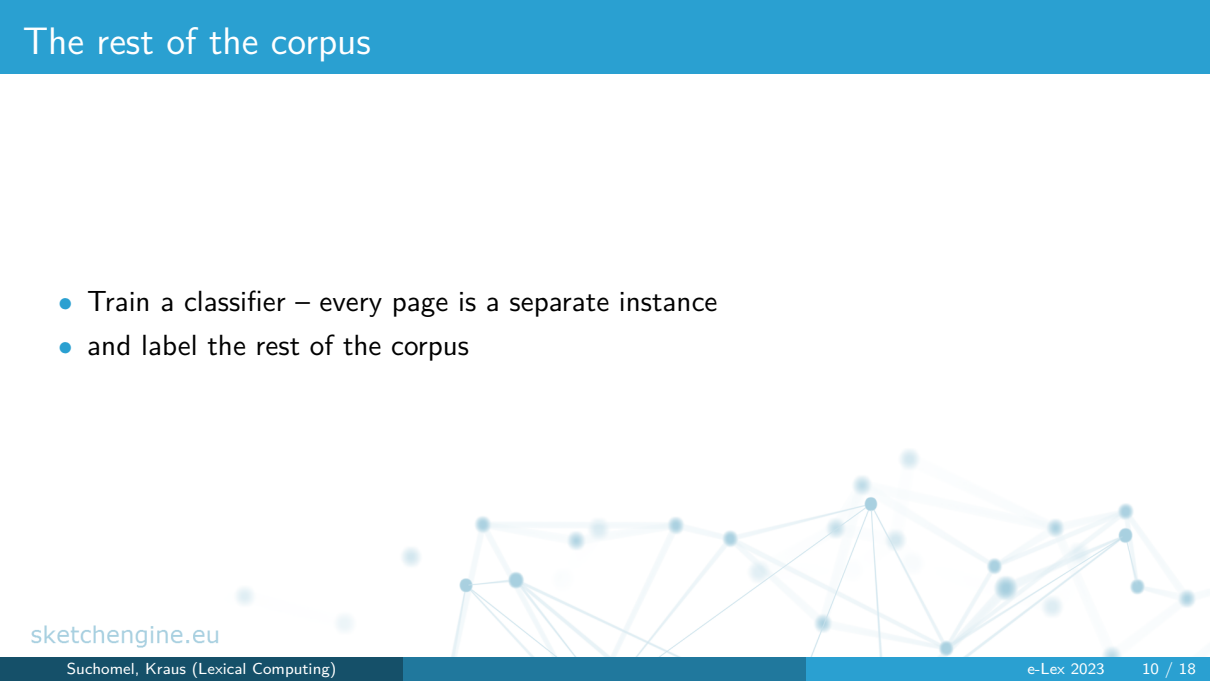
1. Rank websites by token count in the corpus
2. Select top N sites
   - English: 3,000 $\Rightarrow$ 40 % of corpus tokens
   - other languages: 300 – 1,500 $\Rightarrow \geq$ 60 % and even up to 90 % of corpus tokens
3. Split site documents by frequent path prefixes, e.g. `/sports/`, `/culture/`
4. Spend time checking the website content proportionally to its rank
5. Generate table (a website per row) to record annotations
   - hostname (e.g. `bbc.com`)
   - link to the site landing page
   - link to concordance of random sentences from the site in Sketch Engine
6. Check site quality, topic, genre – all at the same time

Website Checks – Quality and Text Types Together

1. Hostname (e.g. `bbc.com`)
   - quality check: long phrases, language code, generic/foreign TLD are suspicious
2. live site checks in a browser
   - non-text, low quality text
   - hijacked/unrelated content
   - selectors with too many language mutations (high chance of MT)
   - MT scripts in the source code
   - a dead site (a high quality content does not get shut down often)
3. link to 100 random triples of consecutive sentences in context in Sketch Engine
   - 3 to 10 sentence triples are inspected
   - the rest is briefly seen and consulted more in the case of a doubtful content or suspicious site
   - each chunk of text can be tracked to the original web page
4. topic and genre
   - lexical or syntactic features typical for a recognized text type
   - unsure or multiple classes – don't label

The rest of the corpus

- Train a classifier – every page is a separate instance
- and label the rest of the corpus



sketchengine.eu

Topic and Genre Selection

Classes expected by users vs. data driven labels

Inspiration

- [Sharoff, 2018]: 18 Functional text dimensions
- [Koppel and Kallas, 2022]: 5 genres and 24 topics
- `curlie.org` (web directory): 15 classes

Our approach

- enough corpus evidence
- inter-annotator agreement
- comprehensible label
- precision over recall

Topics recognized in enTenTen21 (1/2)

Topic	Websites	Tokens
arts	12	169 655 242
beauty & fashion	6	45 899 006
cars & bikes	49	268 201 168
construction & real estate	1	4 610 212
culture & entertainment	123	695 609 769
economy, finance & business	62	387 271 125
education	15	79 155 574
food & drinks	2	9 774 572
gambling & casinos	1	7 839 308
games	52	324 004 431
health	59	426 786 724
history	24	176 510 675
hobbies	18	111 828 110

Topics recognized in enTenTen21 (2/2)

Topic	Websites	Tokens
home, family & children	7	47 126 547
lifestyle	0	0
nature & environment	6	64 495 602
pets & animals	9	33 432 198
politics & government	27	243 239 797
religion	71	424 919 420
science	51	594 461 579
sex	10	209 398 259
sports	103	647 268 352
technology & IT	138	887 566 212
travel & tourism	31	162 020 069
Total	**877**	**6 021 073 951**

Genres recognized in enTenTen21

Topic	Websites	Tokens
blog	99	748 208 188
discussion	194	1 327 118 539
fiction	55	1 009 319 746
legal	37	507 984 084
news	226	1 284 058 175
reference/encyclopedias	10	4 210 237 110
Total	**621**	**9 086 925 842**

Annotators' Agreement at the Website Level

- Four students of applied linguistics at Swansea uni coordinated by Giovanna Donzelli
- 1056, 2600, 1300, 1066 websites analyzed in ~20 hours
- agreement of at least 3/4 students: 498 topic labels, 454 genre labels
- agreement of students' value with our expert: 89 % topic labels, 86 % genre labels

Management of Annotation Resources

- Time spent with each website is proportional to the contribution of the site to the corpus
- several minutes to as less as 20 seconds with each item to inspect
- not assigning any labels is encouraged to reduce noise
- no expert linguistic or computer skills required
- no expert language skill required – live site and MT (GT/DeepL) of sentences is enough

sketchengine.eu

Conclusions

- Semi-manual cost-efficient approach to topic & genre annotation of web corpora
- Enables identification of topic-specific and genre-specific collocations
- enTenTen21: 40 % tokens in annotated documents (3,000 sites)
  - corpus coverage: 14 % tokens by 20 topics & 18 % tokens by 6 genres
- cooperation with native speakers on Estonian, Italian, Spanish, Ukrainian web corpora

sketchengine.eu

References I

Koppel, K. and Kallas, J. (2022).
Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu.
Eesti Rakenduslingvistika Ühingu aastaraamat, 18:207–228.

Papčo, R. (2022).
Topic classification for web corpora: Method comparison and crosslingual transfer.
Master's thesis, Masaryk University.
Supervisor: V. Suchomel.

Sharoff, S. (2018).
Functional text dimensions for the annotation of web corpora.
Corpora, 13(1):65–95.