Know your corpus

Text typology for assessing evidence

Serge Sharoff

Centre for Translation Studies
University of Leeds

s.sharoff@leeds.ac.uk

Outline

A reminder: the history of corpus development

- Critical studies in China: Confucius index
- Concordances: studying the Bible in the middle Ages
- Early corpora: stenography (Käding), foreign language learning
- Computer corpora (1960s): **Brown Corpus** (1962), LOB
- Megacorpus era (1990s, 100+MW): Bank of English and **BNC**
- Internet corpora (2000s, 1+GW): **ukWac**, en10^{10}

Corpus design

The Brown Corpus (1MW)

1. 500 samples, 2000 words in each
2. written sources published in the USA in 1961
3. 15 genres (news, commentary, academic, 6 fiction genres: crime, scifi, humour . . .)

The British National Corpus (100 MW)

1. about 4,000 complete texts
2. 90% of written sources, mostly published in 1980s, 10% of spoken texts (context and demographic)
3. recording their domain, audience, genre: 70 categories (W.fict.prose, W.ac.polit.law, W.ac.socsci, W.med, S.meeting, W.non.ac.socsci, W.non.ac.tech, S.interview.oral.history, . . .)

Composition of Brown and BNC

	Brown	BNC	
80	J. Learned	501	W.misc
75	G. Belles-lettres	432	W.fict.prose
48	F. Popular lore	211	W.pop.lore
44	A. News, reportage	186	W.ac.polit.law.edu
36	E. Skill and hobbies	153	S.conv
31	N. Fiction, adventure	138	W.ac.socsci
30	H. Miscellaneous, government	132	S.meeting
29	K. Fiction, general	128	S.consult
29	P. Fiction, romance	127	W.non.ac.socsci
27	B. News, editorial	123	W.non.ac.techengin
24	L. Fiction, mystery	119	S.interview.oral
17	D. Religion	112	W.commerce
16	C. News, reviews	111	W.non.ac.humanities
9	R. Fiction, humor	100	W.biography
6	M. Fiction, science	95	W.newsp.brdsht.misc

Representativeness and sampling

- John Sinclair: 1987, 1991, 1996, 2003, 2005
  Chapter 1 in Developing Linguistic Corpora, 2005
  `http://users.ox.ac.uk/~martinw/dlc/chapter1.htm`
- What is a corpus for?
  "for studying linguistic constructions reliably"
  unlimitable language vs. limited corpus → sampling
- Constraints on sampling a language:
  1. Which language
  2. Which texts
  3. Which samples

Outline

Language types for sampling

- normative corpus – standard language ('prestigious')

  BC — Brown Corpus of Standard American English

  MICASE — Michigan Corpus of Academic Spoken English
- historical corpus – a sample of language at a time
- monitor corpus – the same kind of language at regular intervals
- demographic corpus
- learner corpora
- parallel and comparable corpora
  - parallel for exact translations
  - comparable for roughly similar texts
    Wikipedia articles in English and Chinese
    Internet corpora as snapshots

Google queries: Googleology

Page 2 of about 3,530,000,000 results (0.41 seconds)

www.healthline.com › health › weakness

Weakness: Causes, Symptoms, and Diagnosis - Healthline

Weakness may be temporary, but it's chronic or continuous in some cases. What **causes** asthenia? Common **causes** of ...

plato.stanford.edu › entries › aristotle-causality

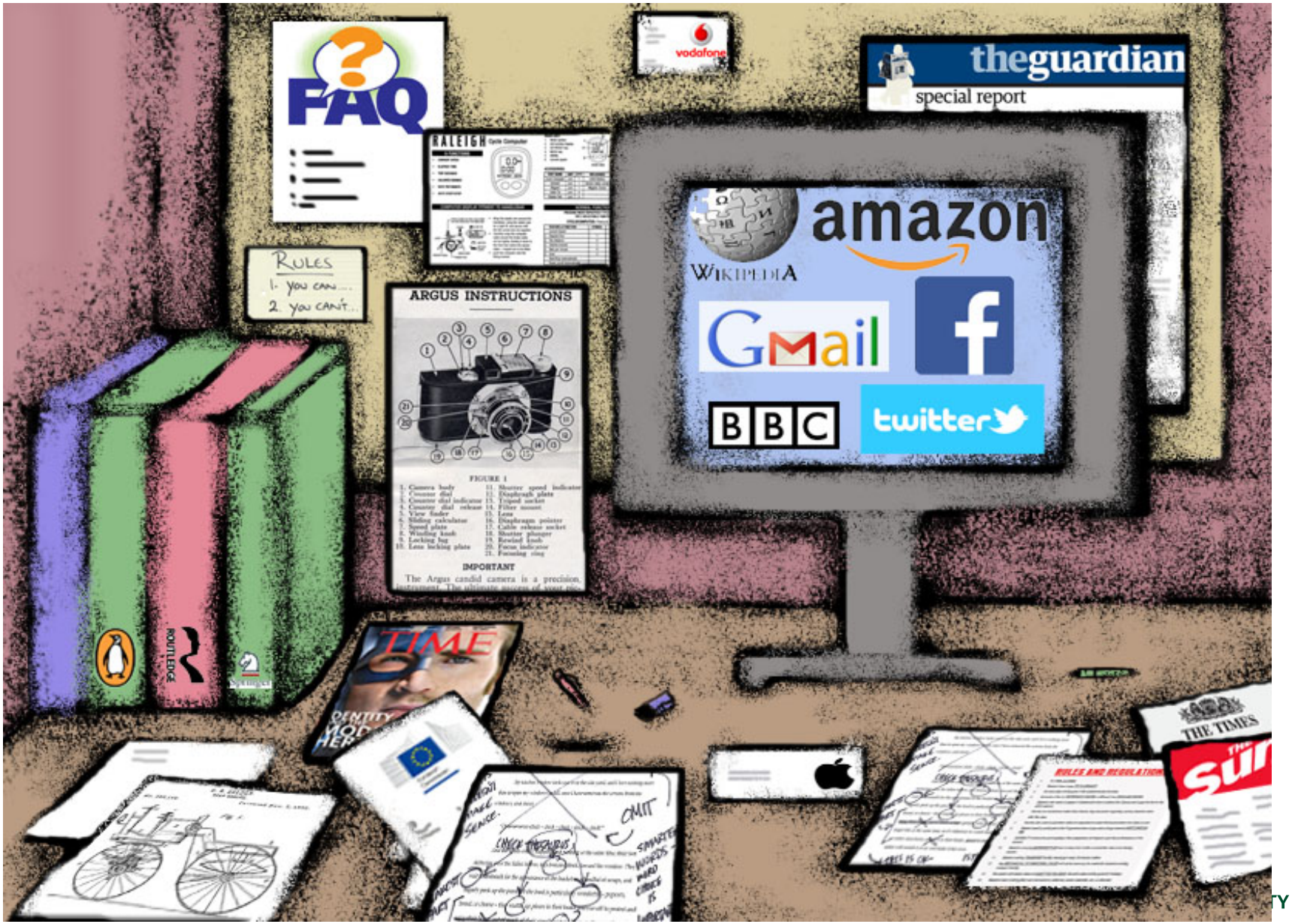
Aristotle on Causality (Stanford Encyclopedia of Philosophy)

11 Jan 2006 - For Aristotle, a firm grasp of what a **cause** is, and how many kinds of **causes** there are, is essential for a successful investigation of the world ...
by A Falcon - 2006 - Cited by 297 - Related articles

- Advantages of corpora:
  - Provenance of texts
  - Stable frequency counts, no repeated texts
  - Stable sort order
  - KWIC vs web snippets
  - Possibility of queries for lemmas and POS tags

Outline

Genres of everyday life



Annotation of ukWac

- ukWac: 2.5 mln .uk texts (Baroni, et al, 2009)

  http://news.ulster.ac.uk/releases/2001/319.html

  http://otis.scotcit.ac.uk/onlinebook/otisT203.htm

  http://site.commedia.org.uk/article/view/694/1/1/

  http://www.afterdinnerspeaker.co.uk/testimonial.html

  http://www.aldermanpeel.norfolk.sch.uk/private/News_easter_04/newslt.htm

  http://www.arun.gov.uk/cgi-bin/buildpage.pl?mysql=2120

  http://www.bettertogether.ac.uk/news.cfm?id=42

  http://www.bias.org.uk/news_story.php?id=153_0

  http://www.bjhc.co.uk/news/1/2004/n40929.htm

  http://www.brookgallery.co.uk/artist.php?arid=77

  http://www.buyagift.co.uk/products/5573.htm

  http://www.cclrc.ac.uk/Activity/DL;WEBNAME=CCLRC;

  http://www.chelmsford.gov.uk/index.cfm?articleid=10020

ukWac URL domains

# Docs	URL domain
23875	cam.ac.uk
16653	ox.ac.uk
15870	ed.ac.uk
15780	demon.co.uk
14555	classic-literature.co.uk
11253	guardian.co.uk
9911	leeds.ac.uk
9051	bham.ac.uk
8423	gla.ac.uk
7617	ucl.ac.uk
6136	open.ac.uk
6126	soton.ac.uk
5665	freeserve.co.uk
5423	independent.co.uk
5135	man.ac.uk
4568	ex.ac.uk

ukWac examples in ac.uk

http://aspirations.english.cam.ac.uk/converse/alevel/growing.

Patrick Leigh Fermor. A Time of Gifts
Patrick Leigh Fermor is a British author, scholar and soldier, who was born in 1915. After being thrown out of school, he decided to walk across Europe from Holland to Turkey, and set out in 1933, just as Hitler began his rise to power.
Germany!... I could hardly believe I was there. For someone born in the second year of World War 1, those three syllables were heavily charged. Even as I trudged across it, early subconscious notions, when one first confused Germans with germs and knew that both were bad, still sent up fumes; fumes, moreover, which the ensuing years had expanded into clouds as dark and baleful as the Ruhr smoke along the horizon and still potent enough to unloose over the landscape a mood of - what?

ukWac examples in ac.uk

http://caialumni.admn.cai.cam.ac.uk/alumni/famous/c17.php

Preacher, best known for inventing the 'Popish Plot', framing a number of Catholics at a time of great religious tension. Not seen as one of the College's great successes, and (I would like to make quite clear) he was only at Caius for two years before leaving for St. John's, so the college influence can (hopefully) be said to be small. **Jeremy Taylor** (1613-1667) Theologian, Bishop of Down and Connor and Vice-Chancellor of Trinity College Dublin. He owed his entire education to the foundation of Dr Perse. Taylor first attended the Perse school, before becoming a Perse Scholar then a Perse Fellow of Caius.

A text typology according to Sinclair

External parameters

- E1 — origin
- E2 — mode
- E3 — aims

E3.1 — audience

E3.2 — intended outcome

Internal parameters

- I1 — domain
- I2 — style

E1: the origin of a text

- the year of text creation
- the authorship (*single*|*multiple*|*corporate*|*unknown*)
- the author's age (*child*|*teen*|*young*|*mid*|*senior*)
- the author's sex (*male*|*female*)
- the place of author's origin (native/non-native, British/American, Kent/Yorkshire . . .)
- originality (*original*|*compiled*|*translated*)

E2: the mode and appearance of the text

- the mode: *written*|*spoken*|*to – be – spoken*|*electronic*
- for written texts:

  printed (books, newspapers, magazines, ephemera)

  typed (all sorts of reports and documentation)

  correspondence (official, personal)

E3.1: the audience of the text

- the size of the audience
  private: 2 / 3 / 5 / 6-20 / 21-50
  public: small, medium, large, very large
- the age of the audience
- the constituency of the audience:
  general|*informed*|*specialist*

E3.2: Generalised aim of text production

- **instruction** – how-tos, FAQs, tutorials, handbooks
- **propaganda** – adverts, political pamphlets
- **recreation** – fiction, biographies and popular lore
- **regulations** – laws, admin, small print
- **reporting** – newswires, police reports
- **discussion** – all texts expressing positions and discussing a state of affairs

Internal parameters: domains (BNC)

- natsci: mathematics, biology, physics, chemistry, geo, ...
- appsci: medicine, engineering, computing, military, ...
- socsci: law, history, philosophy, language, education, ...
- politics: home affairs, world affairs
- commerce: finance, industry, agriculture, ...
- life: fiction
- arts: drawing, literature, architecture, performing, ...
- leisure: sports, travels, fashion, entertainment ...

Outline

Reliability of annotation

Brown cats A) News, reportage, B) News, editorial, C) News, Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Misc, J) Learned, K) Fiction, general, L) Fiction, mystery and crime ...

Reportage or Editorial?

The most positive element to emerge from the Oslo meeting of North Atlantic Treaty Organization Foreign Ministers has been the freer, franker, and wider discussions, animated by much better mutual understanding than in past meetings. This has been a working session of an organization that, by its very nature, can only proceed along its route step by step and without dramatic changes...

("NATO Welds Unity" The Christian Science Monitor, 1961)

Functional Text Dimensions

A1: argum To what extent does the text argue to persuade the reader? (For example, an editorial or an argumentative blog entry)

A8: news To what extent is the text an informative report of recent events? (For example, a newswire)

A17: eval To what extent does the text evaluate something? (For example, a product review)

A11: Personal To what extent does the text report from a first-person point of view? (For example, a personal diary)

Rating Levels:

0	none or hardly at all;	(Sharoff, 2018)
.5	slightly;	
1	somewhat or partly;	
2	strongly or very much so.	

Genres as syndromes

	A1	A3	**A4**	A5	A6	**A7**	**A8**	**A9**	**A11**	**A12**	**A13**	**A14**	A15	**A17**	A19	**A20**
TeleHTC						2.0							0.5			
TelsGoog					0.5	2.0				1.0						
UnitHR								2.0			2.0					
OpacTeam	2.0						1.0				2.0		0.5			
TediJordan	2.0	2.0		0.5	0.5				2.0		0.5					
NewsGueye		1.0		0.5	0.5		2.0									
Bib1Amos		1.0	1.0					0.5	1.0		2.0		1.0		1.0	
FictTolstoy		1.0	2.0	1.0											2.0	

(Halliday, 1992)

a register is a syndrome of lexicogrammatical probabilities

Automatic genre classification

	Brown
163	fiction
111	argument
78	info
29	news
14	personal
13	review
12	promotion
11	academic
11	instruction
7	legal
6	academ/info
2	argum/info
2	info/fiction
2	news/argum
1	argum/news
1	argum/pers
1	fiction/info
1	fiction/pers

	BNC
786	argument
616	fiction
429	personal
390	info
336	news
271	promotion
132	legal
101	instruction
91	academic
83	review
10	pers/argum
7	academ/info
6	argum/news
6	argum/pers
5	fiction/pers
5	news/argum
5	pers/fiction
4	argum/info

	ukWac
500371	promotion
305376	info
266763	argument
236857	news
220880	instruction
154900	review
150201	personal
74291	legal
71147	academic
39150	fiction
5335	news/argum
4913	argum/news
3671	academ/info
3125	info/academ
2968	promo/news
2793	news/promo
2334	instruct/promo
2288	instruct/info

Specialised domains

BNC `arts,medical,natsci,socsci,techeng`

?? Domains are not well represented:

24 texts, 1.4 mln words for medicine

15 texts, 0.6 mln words for linguistics

4 texts, 0.1 mln words for chemistry

Trafilatura for web scraping

Starting with a list of URLs:

```
topic_list=['International law','Human rights','ius gentium
top_url='https://en.wikipedia.org/wiki/'
!pip install trafilatura
import trafilatura
for (i,topic) in enumerate(topic_list):
  url=top_url+topic
  downloaded = trafilatura.fetch_url(url)
  plain_text = trafilatura.extract(downloaded)
  out_file = open(str(i)+".txt",mode="w")
  print(f'{url} to {out_file}')
  print(url, file=out_file)
  print(plain_text, file=out_file)
  out_file.close()
```

Basic points

- Corpora vs. text collections
- Basic types of corpora
- Representativeness and sampling
- Assessing texts using text typologies
- Reliability of annotation:
  Assessing for functional dimensions
- Web corpus collection