Summer School in Corpus Linguistics, Centre for Corpus Research, University of Birmingham, 20 – 24 June 2016

DESIGNING AND INTERPRETING SPECIALISED CORPORA
Viola Wiegand

Activity 1. Defining a specialised corpus

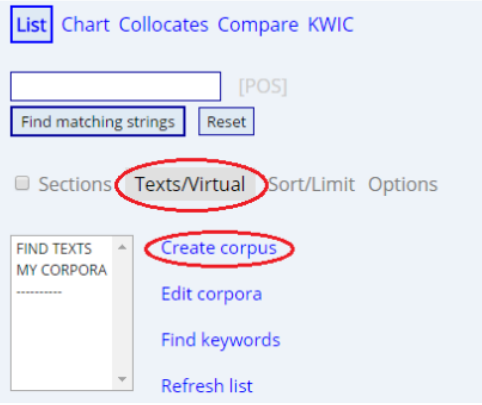
a. What is a specialised corpus?
b. What research questions call for a specialised corpus?
c. What are examples of criteria according to which a specialised corpus can be designed?
d. How would you design a corpus to meet these criteria? Which tools would you use to compile the corpus?

Activity 2. Accessing the CORE corpus via the BYU corpus interface

a. To use the BYU corpus interface, you need to go to http://corpus.byu.edu/ and log in.
b. Under "English" look for the "CORE Corpus" and check its specifications in the table.
c. Click on the corpus name.
d. Read the corpus description.
e. Click on the link to "many different registers" (also see appendix or the direct link http://corpus.byu.edu/help/texts/asp). What range of registers is included? Which register is most frequent in the CORE corpus?

Activity 3. Creating a BYU 'virtual corpus' (example)

a. Click on "Texts/Virtual".
b. Read the description of "Virtual corpora" on the right.
c. Click on "Create Corpus".
d. In the "Genre" box, select the option "Recipe".
e. Click "Submit". (The website may now take a few seconds to load.)
f. On the next screen, examine the list of files. How many files belong to the "Recipe" genre? Check some of the



documents in their original context (open the website link in a new tab).

g. Label the corpus by entering "Recipes" into the "Save as" box on the top left and clicking "Submit".

Activity 4. Extracting keywords from the 'virtual corpus'

a. Go back to the "Search" tab of http://corpus.byu.edu/core .
b. Click on "Texts/Virtual".
c. Select our virtual corpus, "Recipes", from the list.
d. Click on "Find keywords" and check the resulting virtual corpus overview. How big is the corpus?
e. First, select the options "FREQ" and "NOUN", then click on Submit. Examine the resulting keyword list.
f. Now, select the options "SPECIFIC" and "NOUN"; what is the difference between the lists? (Note: The help menu provides some information on the keyword function at http://corpus.byu.edu/help/keyword.asp). What happens if you click on the "+" under "SPECIFIC"?

Activity 5. Creating and analysing your own 'virtual corpus'

a. Create another virtual corpus (or several) according to your own criteria. Try to make use of a search word (instead of or in addition to the register selection). Write down all the choices you make.
b. Check the documents in the resulting corpus.
c. Create a keyword list and examine concordance lines of several keywords.

Activity 6. Reflecting

a. Does the BYU interface allow you to operationalise all your criteria for designing a specialised corpus?
b. Can you think of applications for your own research?
c. What are other ways of designing and interpreting specialised corpora? How do they compare to the BYU interface?

References

1. Specialised corpora/ corpus + discourse/ sampling issues

Barnbrook, G., & Sinclair, J. (2001). Specialised corpus, local and functional grammars. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT: Theory and Practice* (pp. 237–276). Amsterdam: John Benjamins.

Gabrielatos, C. (2007). Selecting query terms to build a specialised corpus from a restricted-access database. *ICAME Journal, 31*, 5–43.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.

Teubert, W. (2007). Natural and human rights, work and property in the discourse of Catholic social doctrine. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, Discourse and Corpora: Theory and Analysis* (pp. 89–126). London: Continuum.

2. Web corpora

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In Proceedings of LREC 2004. Retrieved from http://www.cs.utah.edu/nlp/readinglist/BaroniB04.pdf

Biber, D., Egbert, J., & Davies, M. (2015). Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora*, *10*(1), 11–45. http://doi.org/10.3366/cor.2015.0065

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, *66*(9), 1817–1831. http://doi.org/10.1002/asi.23308

Gatto, M. (2014). *The Web as Corpus: Theory and Practice*. London: Bloomsbury.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, *29*(3), 333–347. http://doi.org/10.1162/089120103322711569

3. Specific tools/ tutorials

- Video tutorials on creating virtual corpora on the BYU corpus interface: http://corpus.byu.edu/wikipedia.asp#tutorials
- BootCaT download: http://bootcat.sslmit.unibo.it
- BootCaT front-end tutorial: http://docs.sslmit.unibo.it/doku.php?id=bootcat:tutorials:basic_1
- WebBootCaT access via https://the.sketchengine.co.uk

Appendix: Registers in the CORE corpus

(retrieved from http://corpus.byu.edu/help/texts/asp)

	Register	# texts	# words
1	Lyrical	636	251,703
2	TV/Movie Script	17	32,502
3	Interview	299	451,593
4	Formal Speech	24	80,109
5	Spoken	103	224,703
6	Interactive Discussion	3,156	2,772,040
7	Recipe	126	80,254
8	How-to/Instructional	1,392	1,400,469
9	Opinion Blog	6,104	9,705,811
10	Advice	1,146	1,286,170
11	Religious Blog/Sermon	721	1,245,410
12	Review	1,925	1,854,423
13	Opinion	87	83,292
14	Narr+Infor Desc+Opinion Hybrid	1,455	1,379,396
15	Narrative + Opinion Hybrid	98	125,541
16	Description with Intent to Sell	1,452	965,447
17	Informational Persuasion	38	70,885
18	Short Story	272	633,734
19	Personal Blog	2,957	2,896,851
20	Travel Blog	371	330,918
21	Sports Report	2,938	2,350,526
22	News Report/Blog	10,399	8,491,496
23	Historical Article	497	828,577
24	Narrative	258	424,614
25	Informational Description + Opinion Hybrid	722	596,308
26	Narrative + Informational Description Hybrid	756	686,424
27	FAQ about Information	415	595,324
28	Description of a Person	759	791,992
29	Informational Blog	1,699	1,902,945
30	Encyclopedia Article	556	1,305,343
31	Research Article	924	1,696,959
32	Informational Description	3,924	3,803,838
33	Other	2,343	3,587,946
	TOTAL	48,569	52,933,543