Oxford Text Archive
Developing Linguistic Corpora

Oxford University Computing Services Summer Seminar

http:/ota.ahds.ac.uk/corpuslinguistics/	Friday 19 July 2002 9:30-17:00

Introduction

This is the webpage for the one-day seminar *Developing Lingusitic Corpora*. Here, the seminar participants can obtain ongoing access to the exercises, links and references which they used on the day. Other interested parties are also welcome to browse here and try the exercises. Please note that it is not possible to provide links from here to the corpora and tools used in the seminar. In order to get copies of these, please see the Oxford Text Archive, ICAME, COBUILD and Oxford University Press.

New Links

The text from the OHP from John's introductory talk, and the text of several chapters of Corpus, Concordance, Collocation.

There was some discussion of tests of statistical significance such as the T-score, and references were made to:

- a paper by Michael Stubbs,
- a contribution to the Corpora email discussion list by Jeremy Clear, and
- a book by Michael Oakes.

There were further references to:

- ICLE - International Corpus of Learner English
- EFL - English as Lingua Franca (if anyone has a better link to more information, please let me know).

Course Info

Seminar leaders:

- John Sinclair, Tuscan Word Centre
- Lou Burnard, Oxford University Computing Services
- Pernilla Danielsson, Centre for Corpus Linguistics, Birmingham University
- Ylva Berglund, Oxford Text Archive
- Martin Wynne, Oxford Text Archive

Programme

9:30	Arrival, registration, introduction
9:45	*Build your own corpus* John Sinclair text from OHP slide

10:30 Surgery session

11:00 coffee break

11:30 *How to retrieve linguistically relevant information from a corpus* John Sinclair

12:15 Practical exercises

13:00 Lunch

14:00 Practical exercises

14:45 Report back on exercises

15:00 tea break

15:30 *The British National Corpus and SARA* Lou Burnard

16:00 General round-up and discussion

17:00 End of seminar

Exercises

1. borrow
2. place - *sorry due to copyright restrictions, this is not available online*
3. diachronic comparisons
4. enormous
5. minitasks
6. visa

Exercises 1,2,5 and 6 above do not require the use of the online corpora or Wordsmith, as small example concordance files are provided in the exercises. You can however if you wish explore these exercises further by looking at the corpora listed below.

Resources

For the practical exercises, the following resources were available online (listed below with wordcounts). While we can't make these freely available here, you can have a look at the manuals, which are a very useful source of technical information.

BNC	Sampler corpus (spoken and written) extracted from the BNC (2m)
BROWN4	Brown Corpus of American English from the 1960s (1m)
BROWNTAG4	Tagged version of Brown
LOB4	Plain text file of LOB (1m)
LOBTAG4	Tagged version of the LOB corpus
FLOB4	Freiburg-LOB 1990s British English (1m)
FLOBTAG4	FLOB with part-of-speech tagging (1m)
FROWNTAG4	Plain text file of tagged Frown corpus (1m)
FROWN4	Freiburg-Brown 1990s American English (1m)
FROWNTAG4	Frown with part-of-speech tagging (1m)
LLC	London-Lund spoken English corpus (1m)
Polish	PELCRA ELAN corpus of Polish newspaper texts (2m)
Susanne	Parsed subcorpus of Brown
ST&WP	Corpus with Speech Presentation annotation (240k words)
Manuals	Manuals for these and other corpora
Wordlist	wordlists for use with the Wordsmith *keywords* utility

Wordsmith Tools

Further exercises

Take a look at the different formats of the various corpora on the CD using Wordpad, or some other text viewer. What are the advantages and disadvantages of the different formats? What sorts of information will it be possible or difficult to extract from them? Take a look at some of your own data if you have any with you. What decisions do you have to make about the format and design of your corpus?

Using Wordsmith Tools, try to extract some information from the corpora listed above. This way you can really find out the advantages and disadvantages of the different types of corpus.

Bibliography

- Guy Aston, and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA.* Edinburgh University Press, Edinburgh.
- Sue Atkins, Jeremy Clear and Nicholas Ostler. 1992. \91Corpus Design Criteria\92 in Literary and Linguistic Computing 7(1): 1-16.
- Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press, Cambridge, UK.
- Geoff Barnbrook. 1996. *Language and Computers.* Edinburgh University Press, Edinburgh.
- Francis Condron, Michael Fraser and Stuart Sutherland. 2000. *Guide to Digital Resources in the Humanities*. CTI Textual Studies, Oxford.
- Nelson Francis and Henry Kučera. 1964. *Manual of Information to accompany the a standard corpus of present-day edited American English, for use with digital computers.* Department of Linguistics, Brown University, Providence, Rhode Island.
- Roger Garside, Geoffrey Leech and Anthony McEnery (eds.). 1997. *Corpus Annotation.* London: Longman.
- Mohsen Ghadessy, Alex Henry and Robert L. Roseberry (eds.). 2001. *Small Corpus Studies and ELT: Theory and practice.* John Benjamin, Amsterdam.
- Stig Johannson and Geoffrey Leech. 1986. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers.* Department of English, University of Oslo.
- Graeme Kennedy. 1998. *An Introduction to Corpus Linguistics.* Longman, London.
- Anthony McEnery and Andrew Wilson. 2001. *Corpus Linguistics* (2nd edition). Edinburgh University Press, Edinburgh.
- Alan Morrison, Michael Popham and Karen Wikander, *Creating and Documenting Electronic Texts*, Oxbow Books, Oxford and freely available online at http://ota.ahds.ac.uk/documents/creating/.
- Oakes, M. P. 1998. *Statistics for corpus linguistics.* Edinburgh, Edinburgh University Press.
- John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford. Currently out of print but PDFs available here: Introduction Chapter 1

Chapter 2 Chapter 6

- John Sinclair (ed.). 1987. *Looking Up*. HarperCollins, London.
- C. M. Sperberg-McQueen and Lou Burnard, (eds.). 1999. *Guidelines for Electronic Text Encoding and Interchange. TEI P3 Text Encoding Initiative*. Revised reprint: Oxford May 1999 (http://www.hcu.ox.ac.uk/TEI/Guidelines/index.htm).
- Stubbs, M. 1995. "Collocations and semantic profiles: On the cause of the trouble with quantitative studies." *Functions of Language* 2(1): 23-55. Available free online.
- Svartvik, J. (1992). *Directions in corpus linguistics : proceedings of Nobel Symposium 82, Stokholm, 4-8 August 1991*. Mouton de Gruyter, Berlin.
- Martin Wynne, Mick Short and Elena Semino. 1998. \91A corpus-based investigation of speech, thought and writing presentation in English narrative texts\92 in Antoinette Renouf (ed), *Explorations in Corpus Linguistics*. Rodopi, Amsterdam.
- Antonio Zampolli and Nicholas Ostler (eds.). 1993. \91Special Section on Corpora\92, Literary and Linguistic Computing 8(4).

Online Resources

Corpora

British National Corpus http://sara.natcorp.ox.ac.uk/
(Oxford University users, see http://www.bodley.ox.ac.uk/oxlip/bnc.htm)
Bank of English (demo) http://www.cobuild.collins.co.uk/
Corpus del Espagnol http://www.corpusdelespanol.org/
COSMAS German corpus http://corpora.ids-mannheim.de/cosmas/
Croatian National Corpus http://www.hnk.ffzg.hr/30m.htm
Czech National Corpus http://ucnk.ff.cuni.cz/

Tools

Wordsmith Tools license from OUP plus information and extra resources from Mike Scott's homepage. Thanks to Oxford University Press for giving special permission to make copies the program for this seminar.

The CONCAPP program was mentioned a couple of times. This is free and easy to use, and you can do some of the clever things with collocations with it that John was demonstrating with the Bank of English software.

Archives

Oxford Text Archive http://ota.ahds.ac.uk/
OLAC Search engine http://saussure.linguistlist.org/cfdocs/new-website/LL-WorkingDirs/olac/olac-search1.cfm
Linguistic Data Consortium http://www.ldc.upenn.edu/
ELDA - The European Language resources Distribution Agency http://www.elda.fr/
TRACTOR http://www.tractor.de/
Electronic Text Center http://etext.lib.virginia.edu/
Project Gutenberg http://promo.net/pg/

Standards

Text Encoding Initiative http://www.hcu.ox.ac.uk/TEI/Guidelines/index.htm
EAGLES http://www.ilc.pi.cnr.it/EAGLES/home.html

Corpus Encoding Standard http://www.cs.vassar.edu/CES/
Open Language Archives Community (OLAC) http://www.language-archives.org/

More information and links\85

W3-Corpora at the University of Essex http://clwww.essex.ac.uk/w3c/
Corpus Linguistics website by Michael Barlow
http://www.ruf.rice.edu/~barlow/corpus.html