



Corpus and software
resources
available at Lancaster

Andrew Hardie & Paul Rayson
UCREL CRS Introductory Talk
Michaelmas term, week 1

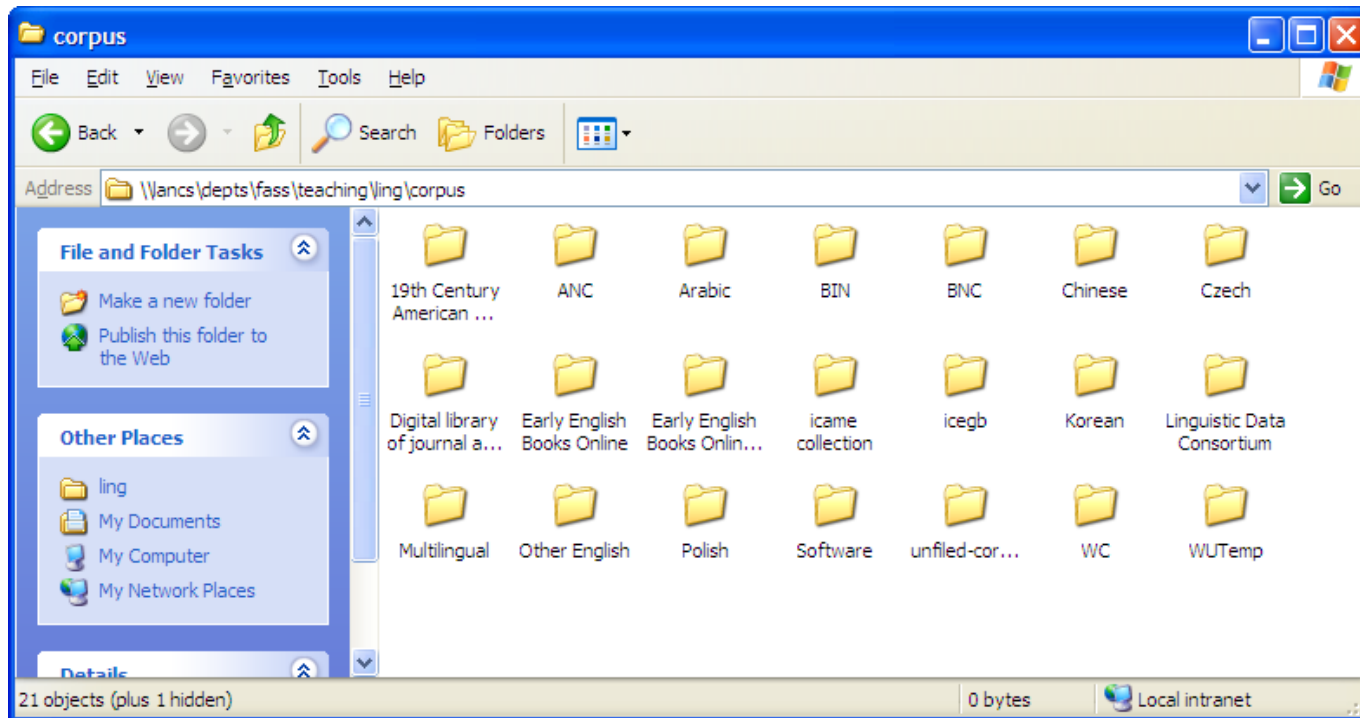
Today's outline

- A brief introduction to:
 - corpus resources
 - UCREL research centre
- Two software demonstrations:
 - CQPweb
 - Wmatrix

Corpus resources

 \\lancs\depts\fass\teaching\ling\corpus
 smb://username@depts.lancs.ac.uk/fass-teaching/ling/corpus

- <http://corpora.lancs.ac.uk/shareview>



Corpus resources (2)

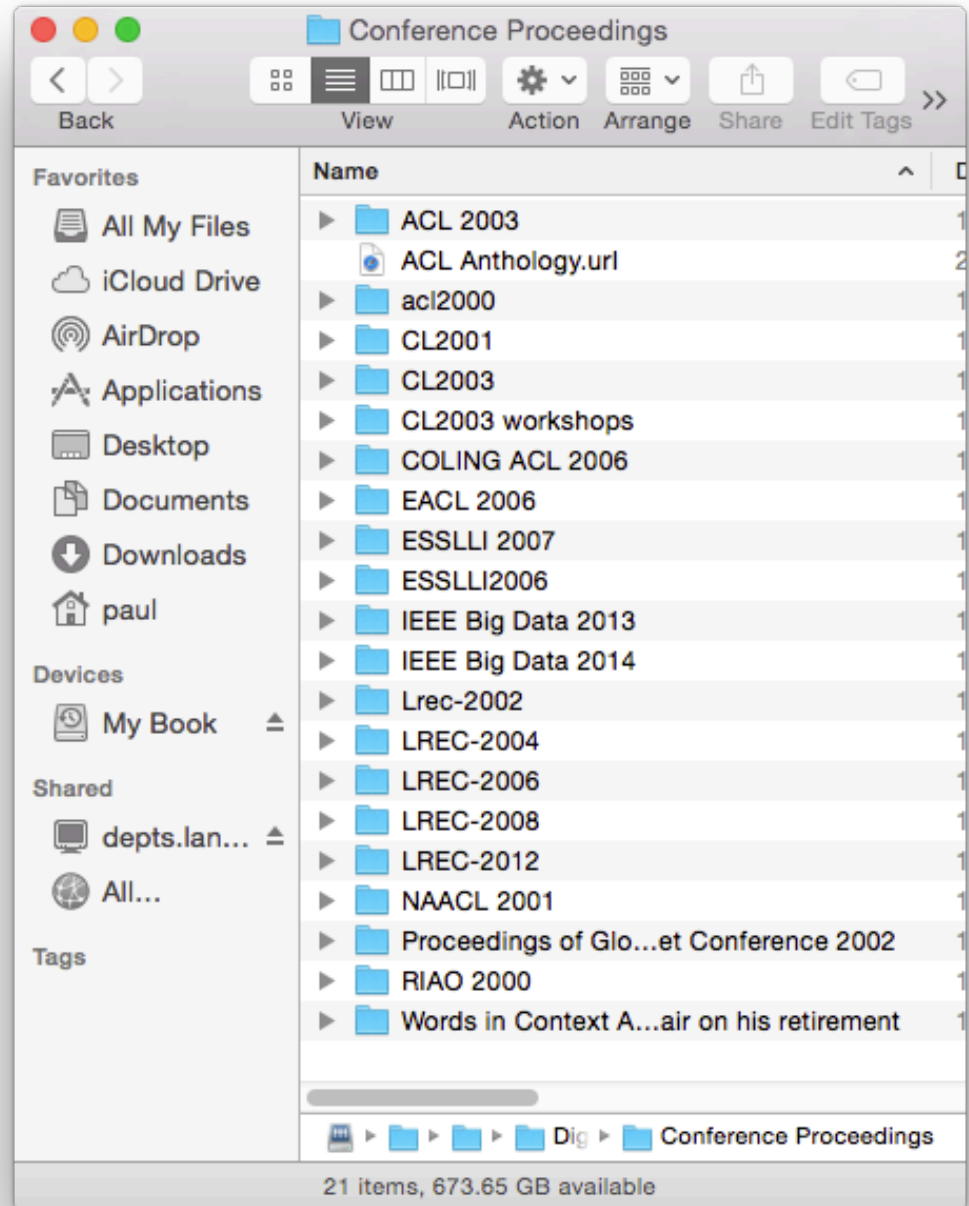
- Linguistic Data Consortium
 - <http://www ldc.upenn.edu/>
 - Membership years: 2001-4, 2007, 2008, 2016
- ICAME collection (2nd edition)
 - <http://icame.uib.no/cd/>
- Bank of English (contact Paul Thompson)
 - <http://www.cqpweb.bham.ac.uk/>
- Archer corpus (contact Paul Rayson)
 - multi-genre corpus of British and American English covering the period 1650-1999
 - also on CQPweb

Corpus resources (3)

- Early English Books Online (EEBO-TCP) v3
 - 1.2 billion words 1473-1700
- UK Hansard
 - 2 billion words, 7 million speeches, 1803-2003
- ~16K Annual Financial Reports, press releases & media articles, conference calls
- Text reuse corpora
 - English-Urdu news, Urdu PA & newspapers
- Twitter dataset(s)
 - See FireAnt software

Digital library

- Conference proceedings
- Corpora
Journal



University Centre for Computer Corpus Research on Language

- <http://ucrel.lancs.ac.uk/>
 - Members
 - Projects
 - Bookshelf
 - Publications list
 - Corpora
- Mailing list
 - <http://scc-lists.lancs.ac.uk/cgi-bin/mailman/listinfo/ucrel>
 - (also: link from UCREL homepage)



Software – web-based tools

- <http://ucrel.lancs.ac.uk/tools.html>
- BNCweb (web based software tied to BNC)
- CQPweb (web based software – multiple corpora)
- BNC Web-Index
- Significance and Effect Size calculator (LL, LR, etc)
 - <http://ucrel.lancs.ac.uk/llwizard.html>
- Wmatrix (web based corpus analysis and comparison)
- <http://corpora.lancs.ac.uk>
 - Significance test system
 - Clustertool
 - DICER variant analysis
 - TreeTagger
 - New General Service List
 - #LancsBox homepage

Software – processing/annotation

- CLAWS part of speech tagger (English)
- USAS semantic tagger
 - Originally English only
 - Now beta versions for Chinese, Dutch, Italian, Portuguese, Spanish, French, Swedish, Welsh, Urdu
- Historical Thesaurus Semantic Tagger
 - <http://phlox.lancs.ac.uk/ucrel/semtagger/english>
- CFIE-FRSE tool
 - PDF to text and structure extraction from annual financial reports
 - Metrics, readability and word list counting
 - <http://ucrel.lancs.ac.uk/cfie/>
- VARD (Variant spelling detector)
 - EmodE historical corpora
 - SMS, Twitter & other online social media
 - <http://ucrel.lancs.ac.uk/vard/about/>

Software – analysis tools

- #LancsBox (incl. GraphColl)
- LWAC (Longitudinal Web As Corpus)
- Geoparser and SHPPS
- Measuring Text Reuse
- Collocation Network Explorer (CONE)
- Fast and memory efficient n-gram tool (Lgram)

Software – from beyond Lancaster

- Netapps (\\lancs\depts\fass\teaching\ling\netapps)
 - AntConc (Free Concordancer by Laurence Anthony)
 - WordSmith (Mike Scott)
 - ICECUP (For ICE corpora)
- SketchEngine via Lancaster University licence
 - <http://sketchengine.co.uk>
 - Using “Log in” > “Authenticate using your institution account (Single Sign On)” > Pick Lancaster Univ.

Linux Virtual Servers

- stig.lancs.ac.uk
 - Hosts Wmatrix (and the UCREL website)
 - (managed by Paul)
- leech.lancs.ac.uk
 - <http://bncweb.lancs.ac.uk>
 - <http://cqpweb.lancs.ac.uk>
 - <http://corpora.lancs.ac.uk>
 - (managed by Andrew)
- Perl, PHP, MySQL; CWB/CQP; UCREL tools
- Research cluster for Hadoop and VMs
 - (managed by Paul, Alistair, Andrew and Matt)
- GitLab (internal/private projects): <https://delta.lancs.ac.uk/>
- GitHub (external/public projects): <https://github.com/UCREL>