21st Century Corpus Workbench!

Updating a query architecture for the new millennium

Stefan Evert (Osnabrück)
Andrew Hardie (Lancaster)

CL2011, Birmingham, July 2011
A brief introduction

- What CWB is
- What CWB isn’t
- Where you may know it from
- What it’s good for
Where you may know it from…

- CWB as the server backend for several widely used interfaces…

  - BNCweb
  - OPUS parallel corpora
  - Leeds online corpus collection
  - VISL site
Corpus Workbench: an introduction

• What kind of tool is CWB?
  • System for indexing and searching large corpora via a powerful data model and query language
  • Targeted especially at annotated corpora
  • Can handle medium-large to very-large datasets

• What is CWB not?
  • Not a concordancer meant for ad-hoc analysis of smaller datasets (cf. WordSmith, AntConc, …)
  • Not really targeted at the beginner
What’s good about Corpus Workbench

- Staying power!
  - A tool of long standing
  - Original implementation: see Christ (1994)
  - De facto standard input format and query language
  - Has influenced other systems, e.g. Manatee backend used by SketchEngine (Kilgarriff et al 2004)
- Works both as a server backend for centralised systems and as an install-it-yourself tool for the individual user (command line or GUI)

- So how does it all work?
CWB corpora: the input format

- **P-attributes** (word-level annotations): represented by tab-delimited values
- **S-attributes** (regions): represented by XML tags
  - these don’t count as extra tokens
- One token or XML tag per line
- **A-attributes** for sentence alignment (→ OPUS etc.)
Corpus Query Processor

• Fast, efficient, two-level corpus searches

• Patterns specified at the token level…
  • = regular expressions on words-tags/…

• … and the token-sequence level
  • = regular expressions across patterns of tokens

• Flexible, de facto standard query language
Corpus Query Processor

[word = "queries" %c] [class="PREP"]
[class="ADJ"]+ [pos="NN.*" & word=".*s"];
How it works

- CWB utilities
- Corpus Library (access functions)
- Indexed corpora
- CQP

Indexed corpora describes Registry
# 21st Century Requirements

<table>
<thead>
<tr>
<th>What was reasonable in the 1990s ...</th>
<th>What’s required for a 21st century system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running only on commercial version of Unix</td>
<td>Linux, Mac OS X, and Windows</td>
</tr>
<tr>
<td>Designed for the 32-bit world (limited RAM and disk size)</td>
<td>Full 64-bit support</td>
</tr>
<tr>
<td>Western European character support only (Latin 1)</td>
<td>Full multilingual support via Unicode</td>
</tr>
<tr>
<td>Command-line interface</td>
<td>User-friendly graphical interface</td>
</tr>
<tr>
<td>System run by and available from a single institution (IMS)</td>
<td>Public, open source project</td>
</tr>
</tbody>
</table>
Open CWB

- Licence is open-source (GPL)
- Code / builds openly available to download
- The development of CWB is conducted in public
  - Our public bug / feature-request database:
    - http://sourceforge.net/tracker/?group_id=131809
  - Discussions about the development of CWB are public
    - Mailing list
    - Online roadmaps
- Find all information on **cwb.sf.net**
64-bit support

- CWB’s limits in the 32-bit version:
  - Maximum corpus size 200-500 million words (less if heavily annotated)
  - Reasonable limitation by 1990s standards

- Under 64-bit operating systems:
  - Maximum corpus size rises to 2.1 billion words (due to internal data format / API)
Win32 version

- Mac OS X / Linux since early 2000s: Windows version a longstanding demand

- Win32 compatibility integrated from Textometrie project’s fork of CWB

- MinGW cross compiler used to build on Linux, deploy to Windows
  - http://www.mingw.org/

- Some debugging still needed
Character set support

- v3.0 and below: ISO-8859-1 (Latin 1)
- v3.2+: Also Unicode (UTF-8)

- Character set validation

- Unicode support contributed by ANR Textométrie, ENS de Lyon
Advantages:
- Covers all the world’s writing systems
- Plenty of open-source support available
  - GLib, PCRE, …

Disadvantages:
- Can take up more disk space (even UTF-8)
- Would require existing users to upgrade their corpora

Solution: implement legacy ISO-8859 encodings for Greek, Hebrew, Arabic etc. as well
Unicode support

- CWB utilities
- CQP
- Corpus Library
  - ISO-8859 data
  - Regular expressions
  - Legacy character utilities
- PCRE
- GLib
  - UTF-8 character utilities
Interface user-friendliness

- CWB/CQP’s great limitation
- The CWB-Perl interface
  - v3.0+: for the creation of bespoke web or command-line interfaces
- Common Elementary Query Language (CEQL)
  - used in BNCweb (see Hoffmann et al. 2008)
CQPweb

- Extension of BNCweb-like interface to any corpus
- Adopted as GUI front-end of the CWB project
- Multilingual support (UTF-8 throughout)
- Additional functions via MySQL databases
  - Collocation, query distribution, corpus metadata
- See Hardie (forthcoming)
476652: many commendations on the \texttt{indefatigable} friendship of Miss Tox. Mrs. Sparsit, with a very
1222806: h the full consent of the \texttt{indefatigable} page, who \texttt{being} the 0
1447328: ceived, by favour of the \texttt{indefatigable} Mrs Grudden, no less a
1452848: n this conversation, the \texttt{indefatigable} Mr Pyke threw himself in
2295229: severe cold, which this \texttt{indefatigable} officer had caught in hi
2302755: s on the praiseworthy and \texttt{indefatigable} exertions of certain est
2488130: as clean and bright, as \texttt{indefatigable} white-vashing, and hear
2492876: gentlemen! shouted the \texttt{indefatigable} little man with the whis
2474471: e, and when, by dint of \texttt{indefatigable} pumping, she had manage
2853186: man understanding, that \texttt{indefatigable} lady sat down to dinner
3013743: f the careful attention, \texttt{indefatigable} assiduity, and nice dis
3028885: ass the wine, said the \texttt{indefatigable} visitor. Mr. Tupman did
3022147: tant one than any—he was \texttt{indefatigable} in paying the most unrem
3058284: ed and the mattress. The \texttt{indefatigable} stranger rose betimes ne
3121689: had it not been for the \texttt{indefatigable} perseverance with which
3166467: tared through it with the \texttt{indefatigable} manner in which he had c
3201280: breath, by reason of the \texttt{indefatigable}
British English 2006: powered by CQPweb

**Standard Query**

Query mode: Simple query (ignore case)  
Number of hits per page: 50  
Restriction: None (search whole corpus)  

Start Query  Reset Query

**System messages**

Reference for CQPweb

The long-promised draft paper on CQPweb is now available to read:

Hardie, A (forthcoming) "CQPweb - combining power, flexibility and usability in a corpus analysis tool".

When I manage to get it published, I will update the reference here.
Onwards: What the future holds

- Bigger corpora yet: many billions of tokens
- Interchangeable concordance output format (XML)
- Multiple target positions (for complex frequency data)
- Better XML indexing & queries
  - recursive nesting of elements, start tag attributes
- Query optimisation
  - Changes in the QL?
  - Different, specialised QLs?
- New query features
  - Google-style IR searches (“MU queries”)
  - Queries on dependency parse graphs
  - Fuzzy search & phrase queries
- What do you need?
Thank you!

BNC:Spoken[908]> MU(meet "thank"%c "you"%c 1 1) cut 10;
42848715: <text_id D90>: eeting. So Brenda. <Thank> you. Well some of
42853032: <text_id D90>: the blackout. Well <thank> you very much .. I
42853683: <text_id D90>: rather reluctantly. <Thank> you very much [uncl
42853691: <text_id D90>: really interesting. <Thank> you. and it 's [un
42853707: <text_id D90>: le in the audience. <Thank> you very much indee
42854621: <text_id D91>: I see Okay. Okay. <Thank> you. [unclear] . C
42854746: <text_id D91>: f people 's comments <thank> you. Thank you. O
42854749: <text_id D91>: comments thank you. <Thank> you. Okay. Can we
42854872: <text_id D91>: rd maintained. Well <thank> you for that that ' t to the evening .. <Thank> you. Jan [gap:name
Ask us about CWB!

- http://devel.sslmit.unibo.it/mailman/listinfo/cwb
- severt@uos.de | purl.org/stefan.evert
- a.hardie@lancaster.ac.uk
References


Hardie, A. Forthcoming. "CQPweb - combining power, flexibility and usability in a corpus analysis tool". http://www.lancs.ac.uk/staff/hardiea/cqpweb-paper.pdf (draft)
