21st Century Corpus Workbench!

Updating a query architecture for the new millennium

Stefan Evert (Osnabrück)

Andrew Hardie (Lancaster)

CL2011, Birmingham, July 2011

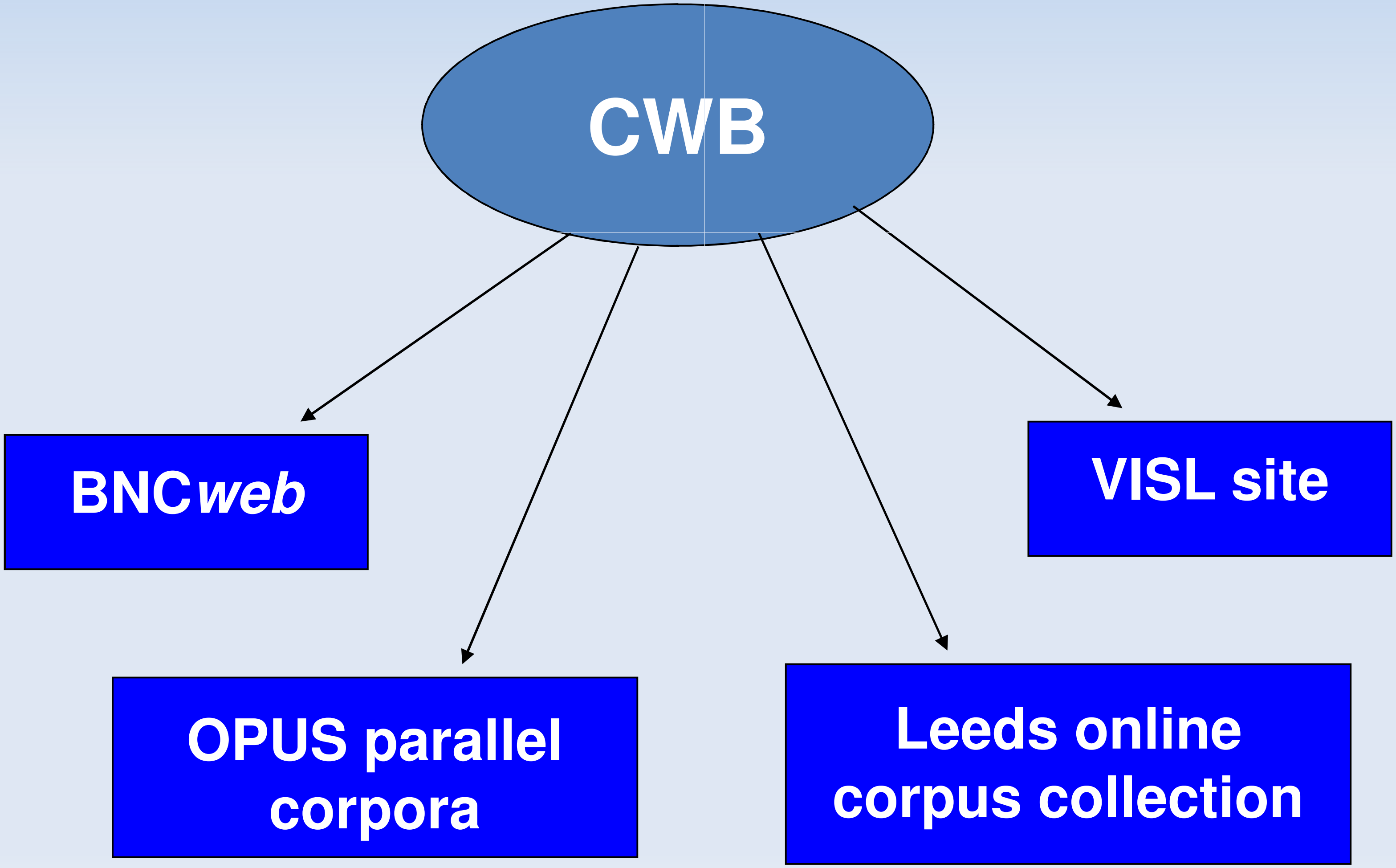
LANCASTER UNIVERSITY UCREL CWB UNIVERSITÄT OSNABRÜCK

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…



CWB

BNC*web*

VISL site

OPUS parallel corpora

Leeds online corpus collection

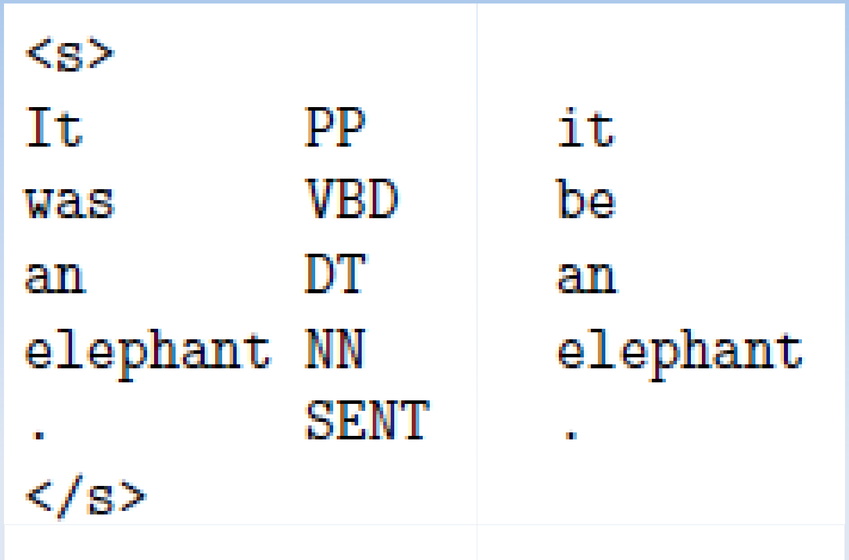
Corpus Workbench: an introduction

- 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



```
<s>
It        PP      it
was       VBD     be
an        DT      an
elephant  NN      elephant
.         SENT    .
</s>
```

- 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

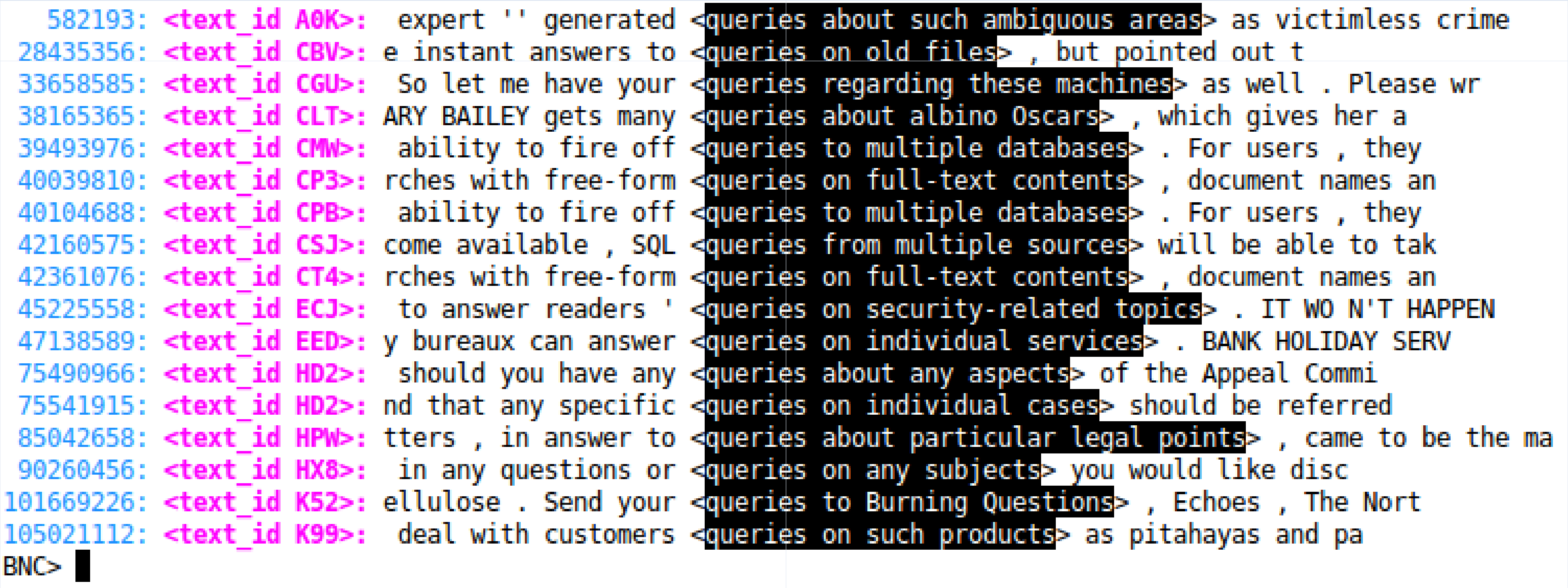
- Fast, efficient, two-level corpus searches

- 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

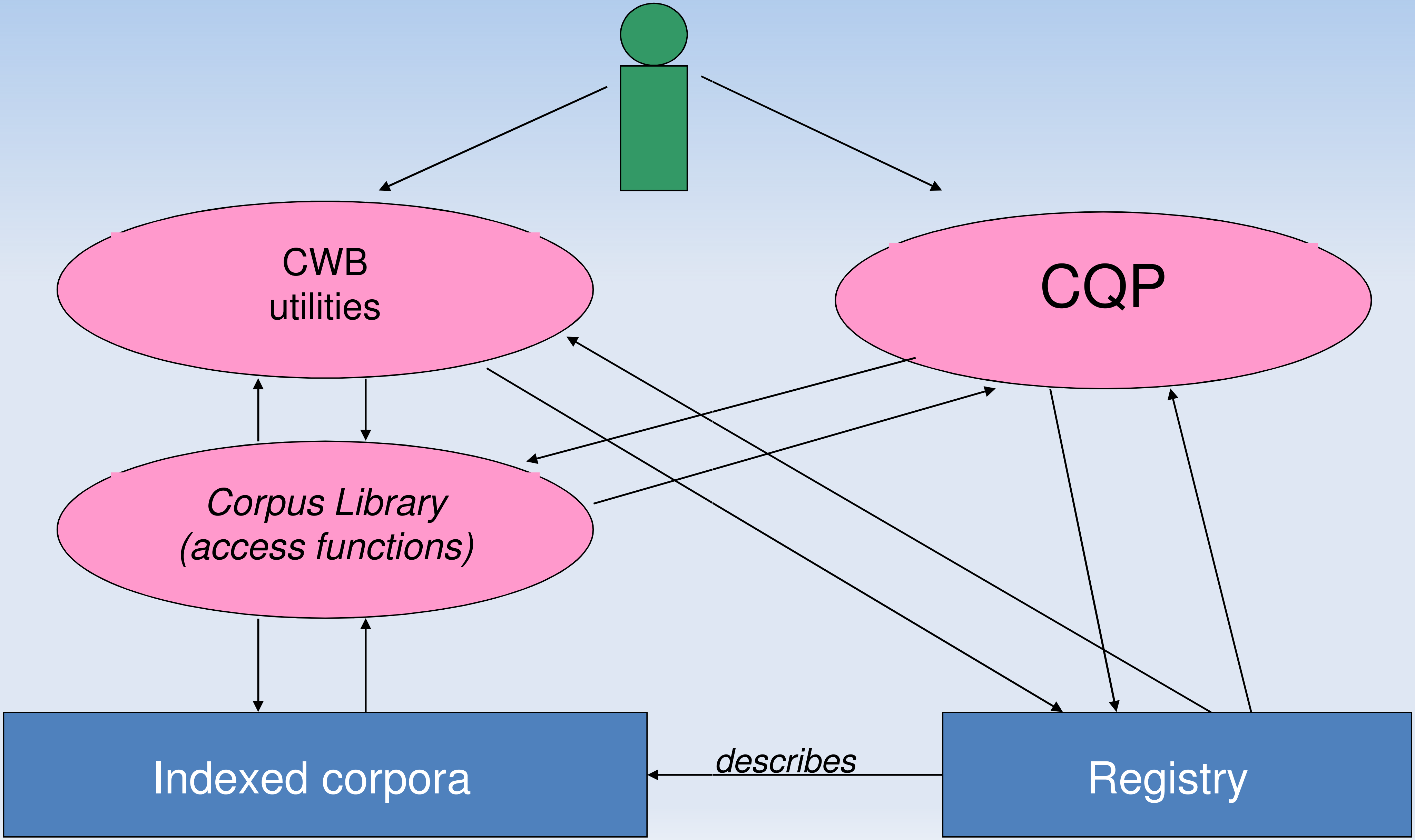
Corpus Query Processor

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



```
   582193: <text_id A0K>:  expert '' generated <queries about such ambiguous areas> as victimless crime
 28435356: <text_id CBV>: e instant answers to <queries on old files> , but pointed out t
 33658585: <text_id CGU>:  So let me have your <queries regarding these machines> as well . Please wr
 38165365: <text_id CLT>: ARY BAILEY gets many <queries about albino Oscars> , which gives her a
 39493976: <text_id CMW>:  ability to fire off <queries to multiple databases> . For users , they
 40039810: <text_id CP3>: rches with free-form <queries on full-text contents> , document names an
 40104688: <text_id CPB>:  ability to fire off <queries to multiple databases> . For users , they
 42160575: <text_id CSJ>: come available , SQL <queries from multiple sources> will be able to tak
 42361076: <text_id CT4>: rches with free-form <queries on full-text contents> , document names an
 45225558: <text_id ECJ>:  to answer readers ' <queries on security-related topics> . IT WO N'T HAPPEN
 47138589: <text_id EED>: y bureaux can answer <queries on individual services> . BANK HOLIDAY SERV
 75490966: <text_id HD2>:  should you have any <queries about any aspects> of the Appeal Commi
 75541915: <text_id HD2>: nd that any specific <queries on individual cases> should be referred
 85042658: <text_id HPW>: tters , in answer to <queries about particular legal points> , came to be the ma
 90260456: <text_id HX8>:  in any questions or <queries on any subjects> you would like disc
101669226: <text_id K52>: ellulose . Send your <queries to Burning Questions> , Echoes , The Nort
105021112: <text_id K99>:  deal with customers <queries on such products> as pitahayas and pa
BNC> 
```

How it works


CWB
utilities
CQP
Corpus Library
(access functions)
Indexed corpora
describes
Registry

21st Century Requirements

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

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
      - http://cwb.sourceforge.net/future.php
      - http://cwb.svn.sourceforge.net/viewvc/cwb/cwb/trunk/doc/unicode_roadmap.html

- 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*

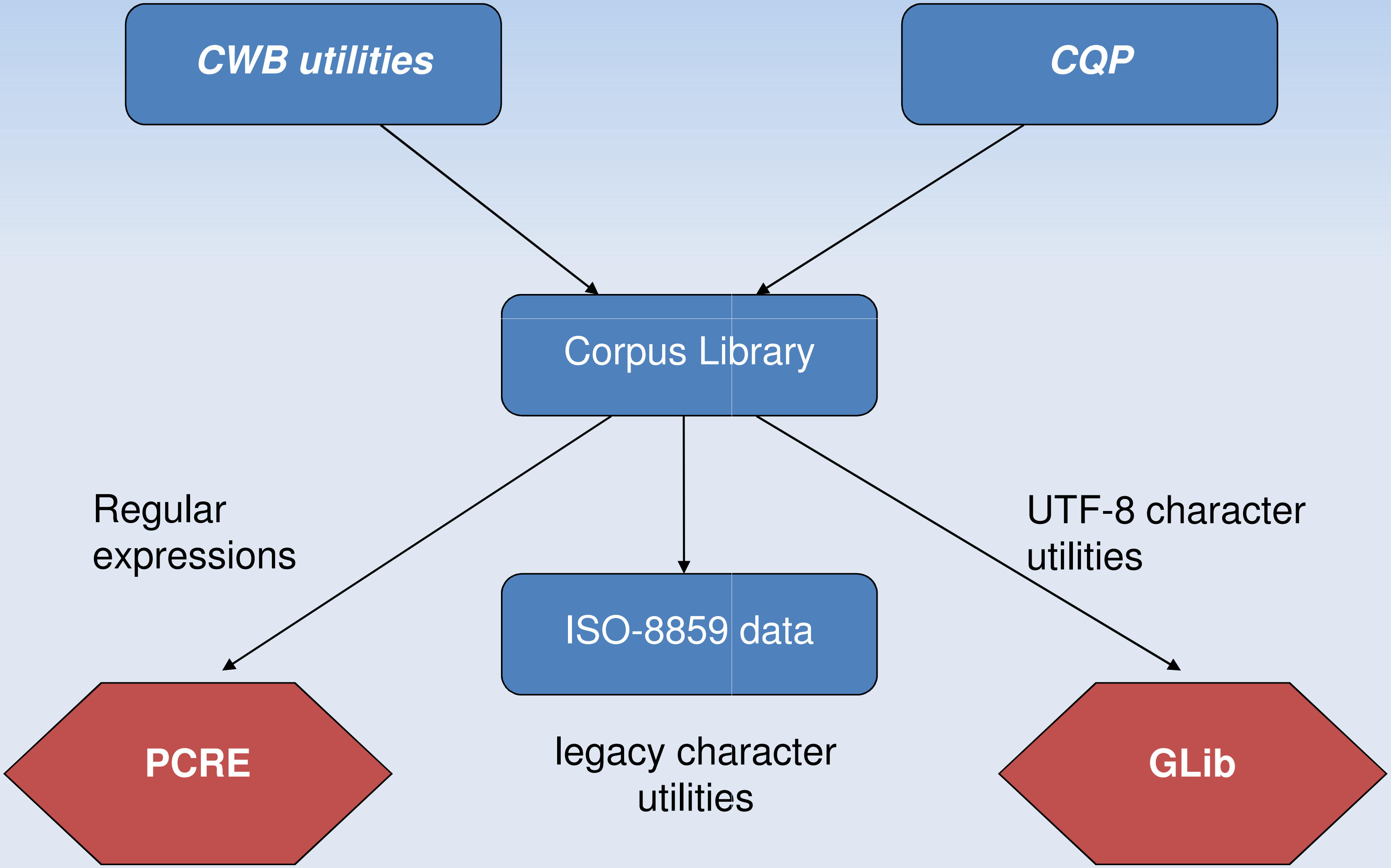
Character set support

- 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

Unicode

- 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

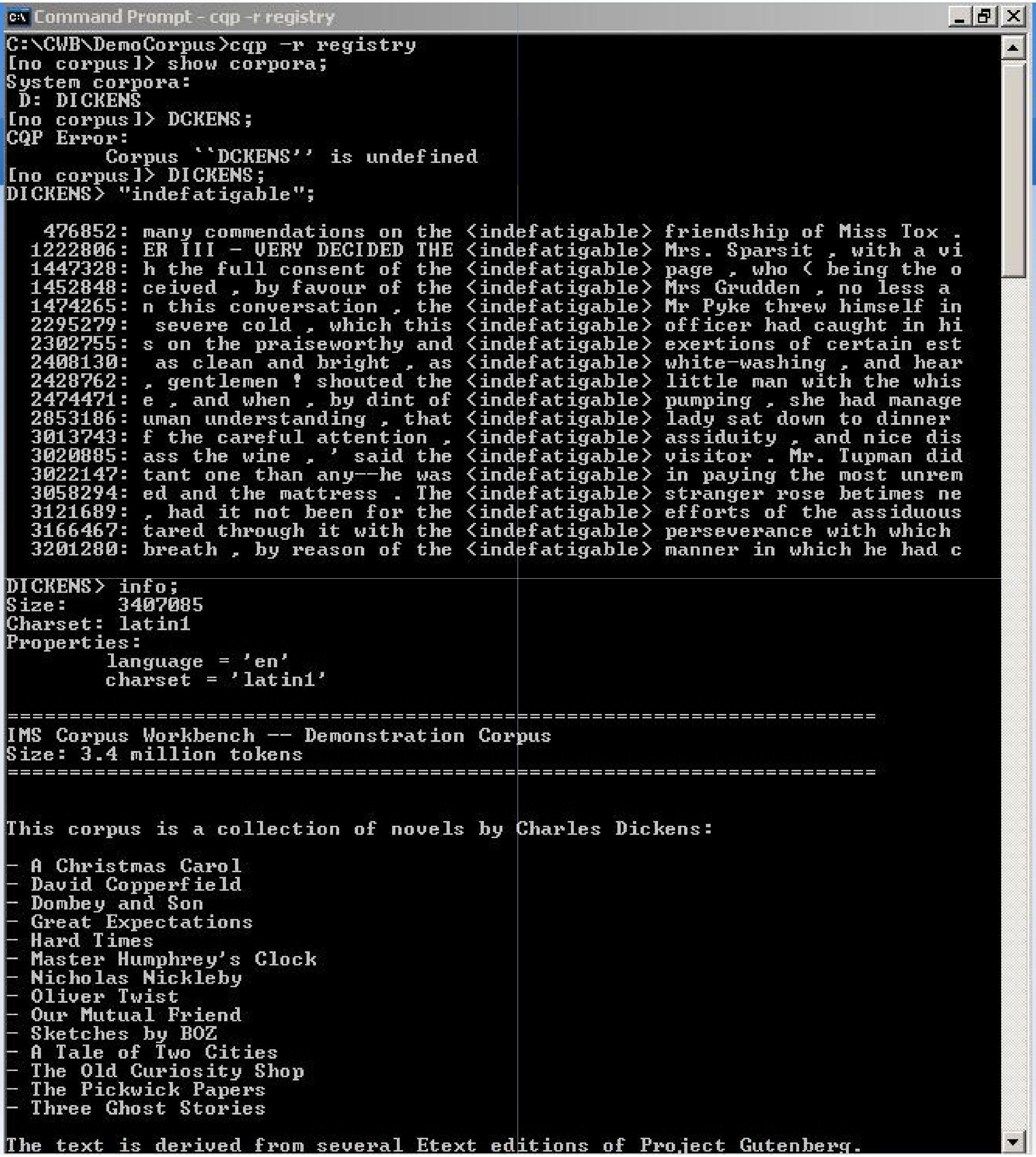


Interface user-friendliness

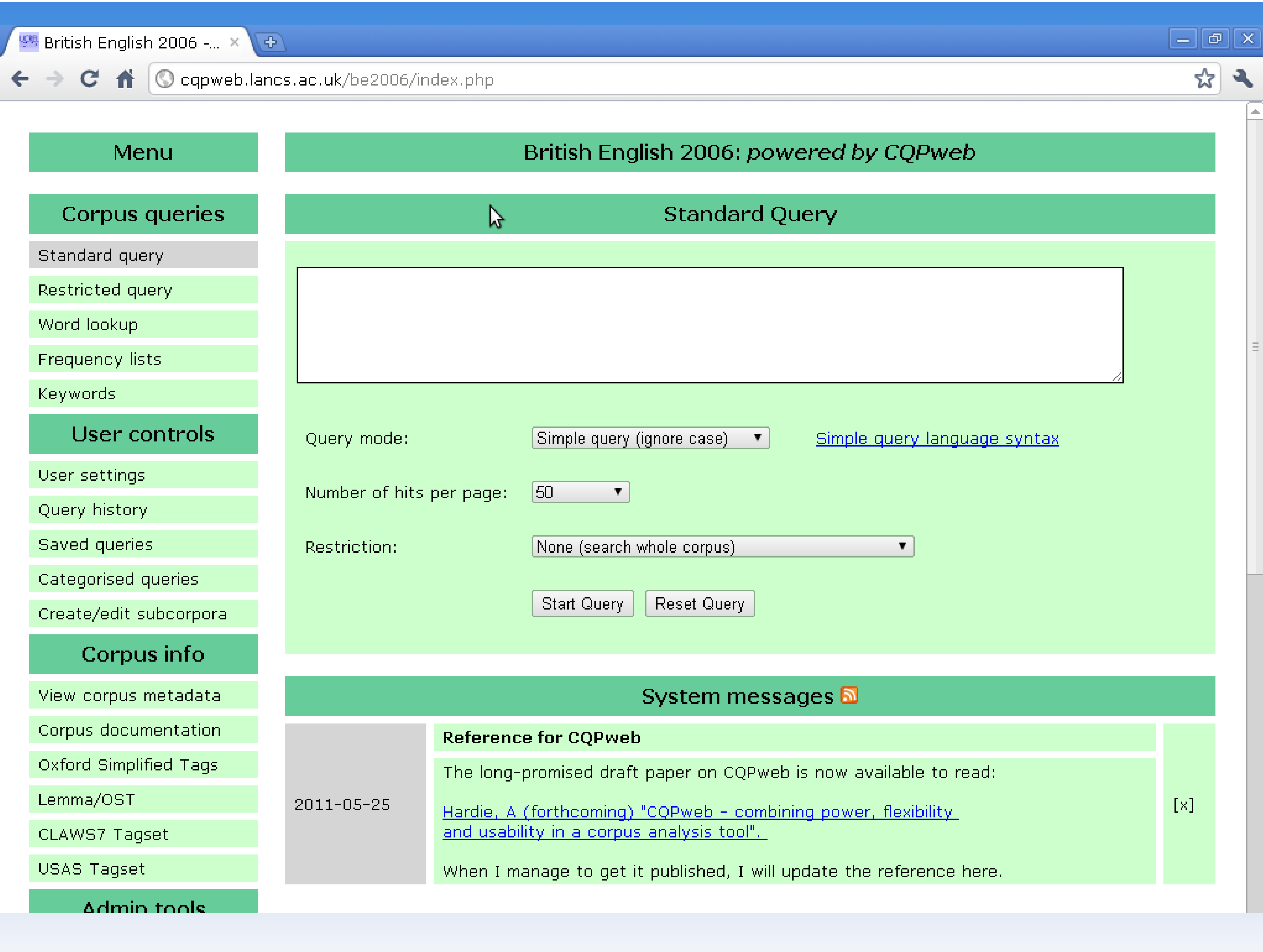
- CWB/CQP’s great limitation
- The CWB-Perl interface
  - v3.0+: for the creation of bespoke web or command-line interfaces
- Common Elementary Query Language (CEQL)
  - used in BNCweb (see Hoffmann et al. 2008)

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)



```
C:\CWB\DemoCorpus>cqp -r registry
[no corpus]> show corpora;
System corpora:
 D: DICKENS
[no corpus]> DCKENS;
CQP Error:
        Corpus ``DCKENS'' is undefined
[no corpus]> DICKENS;
DICKENS> "indefatigable";

   476852: many commendations on the <indefatigable> friendship of Miss Tox .
  1222806: ER III - VERY DECIDED THE <indefatigable> Mrs. Sparsit , with a vi
  1447328: h the full consent of the <indefatigable> page , who ( being the o
  1452848: ceived , by favour of the <indefatigable> Mrs Grudden , no less a
  1474265: n this conversation , the <indefatigable> Mr Pyke threw himself in
  2295279:  severe cold , which this <indefatigable> officer had caught in hi
  2302755: s on the praiseworthy and <indefatigable> exertions of certain est
  2408130:  as clean and bright , as <indefatigable> white-washing , and hear
  2428762: , gentlemen ! shouted the <indefatigable> little man with the whis
  2474471: e , and when , by dint of <indefatigable> pumping , she had manage
  2853186: uman understanding , that <indefatigable> lady sat down to dinner
  3013743: f the careful attention , <indefatigable> assiduity , and nice dis
  3020885: ass the wine , ' said the <indefatigable> visitor . Mr. Tupman did
  3022147: tant one than any--he was <indefatigable> in paying the most unrem
  3058294: ed and the mattress . The <indefatigable> stranger rose betimes ne
  3121689: , had it not been for the <indefatigable> efforts of the assiduous
  3166467: tared through it with the <indefatigable> perseverance with which
  3201280: breath , by reason of the <indefatigable> manner in which he had c

DICKENS> info;
Size:     3407085
Charset: latin1
Properties:
        language = 'en'
        charset = 'latin1'

======================================================================
IMS Corpus Workbench -- Demonstration Corpus
Size: 3.4 million tokens
======================================================================


This corpus is a collection of novels by Charles Dickens:

- A Christmas Carol
- David Copperfield
- Dombey and Son
- Great Expectations
- Hard Times
- Master Humphrey's Clock
- Nicholas Nickleby
- Oliver Twist
- Our Mutual Friend
- Sketches by BOZ
- A Tale of Two Cities
- The Old Curiosity Shop
- The Pickwick Papers
- Three Ghost Stories

The text is derived from several Etext editions of Project Gutenberg.
```



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)

System messages

2011-05-25

Reference for CQPweb

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

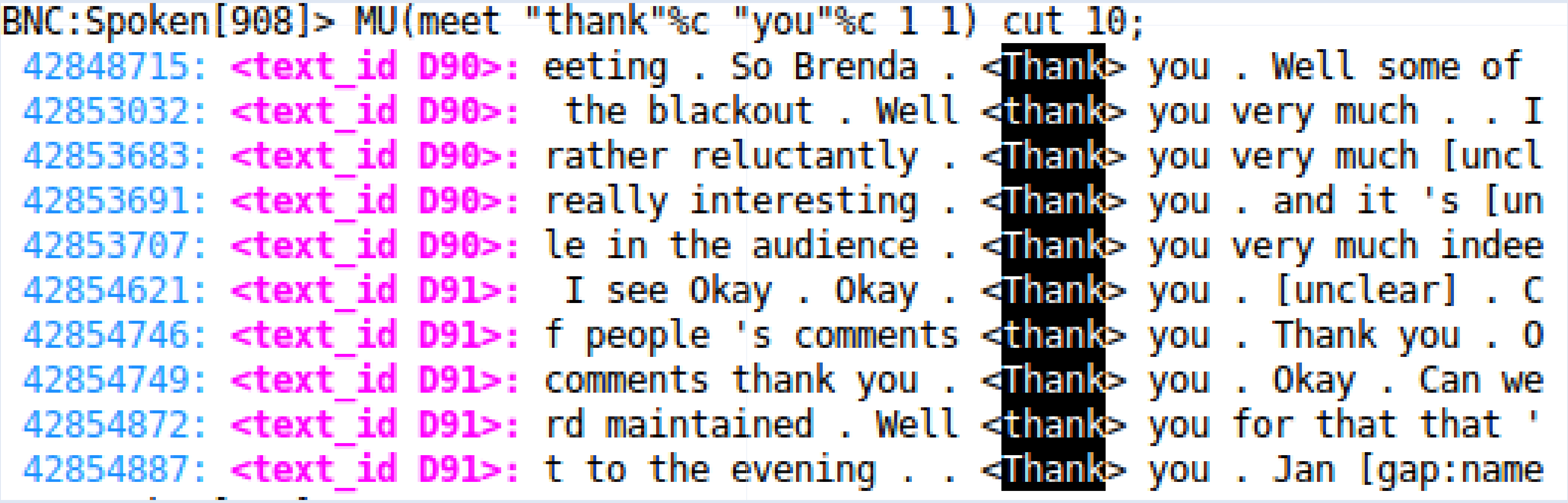
Hardie, A (forthcoming) "CQPweb - combining power, flexibility and usability in a corpus analysis tool".

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 '
 42854887: <text_id D91>: t to the evening . . <Thank> you . Jan [gap:name
```

Ask us about CWB!

- http://cwb.sourceforge.net
- http://devel.sslmit.unibo.it/mailman/listinfo/cwb

- severt@uos.de | purl.org/stefan.evert
- a.hardie@lancaster.ac.uk

References

Christ, O. 1994. "A modular and flexible architecture for an integrated corpus query System", in Proceedings of COMPLEX '94, pp. 23–32. Budapest.

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)

Hoffmann, S., Evert, S., Smith, N., Lee, D. and Berglund Prytz, Y. 2008. Corpus Linguistics with BNCweb – a Practical Guide. Frankfurt am Main: Peter Lang.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. "The Sketch Engine", in G. Williams and S. Vessier (eds) Proceedings of EURALEX 2004, pp. 105–116. Bretagne, France: Université de Bretagne-Sud.