Extensibility as a focus for corpus analysis software

The CQPweb plugin framework

Andrew Hardie

ESRC Centre for Corpus Approaches to Social Science

Lancaster University UK

@HardieResearch

The background

- Corpus Workbench (CWB)
  - Open project: http://cwb.sf.net

- CQPweb : web-based UI to CQP
  - Interface design follows BNCweb
  - 2008: Internal at Lancaster https://cqpweb.lancs.ac.uk
  - 2010: open to other users
  - As part of CWB: open source
    - Many other servers across the world

CASS
Corpus Approaches to Social Science

Why EXTENSIBILITY?

Features in demand

CASS
Corpus Approaches to Social Science

- Users to be able to upload their own corpora
- Interface to corpus construction systems
- More options to customise download formats
- Wider range of analyses
  - Of concordances, of corpora
  - Traditional, experimental
  - Statistics, visualisations

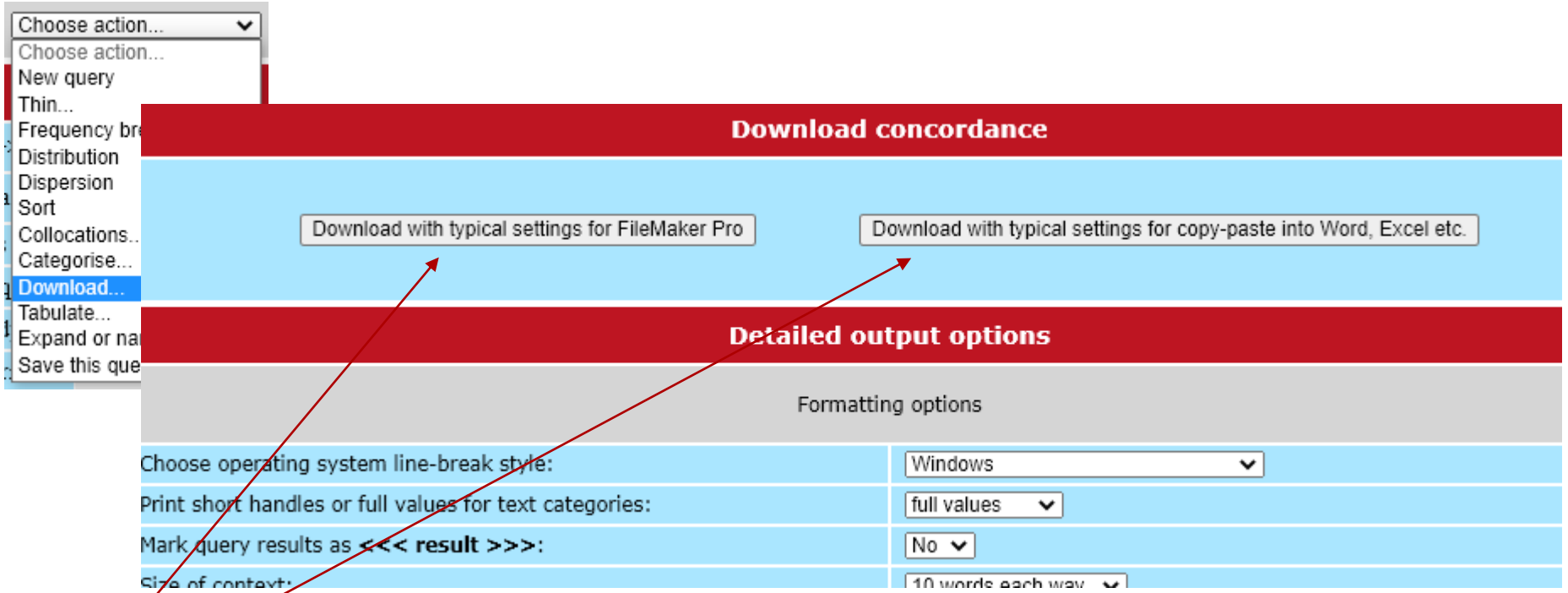
Plugins

CASS
Corpus Approaches to Social Science

- What is a **plugin**?

- CQPweb modular elements
  - Download
  - Install corpus
  - Annotate text data
  - Process concordance to another form
- … places where the normal process could be replaced *as a unit* by another unit that connects up the right way

Downloader Plugin



Now plugins!

Downloader plugins activated in Admin interface

Existing Plugins					
ID	**Type**	**Description**	**Class**	**Extra setup**	**Acti**
7	QueryDownloader	Filemaker download button	TypicalForFilemaker		[Activate]
8	QueryDownloader	Word/Excel download button	TypicalWordExcel		[Activate]

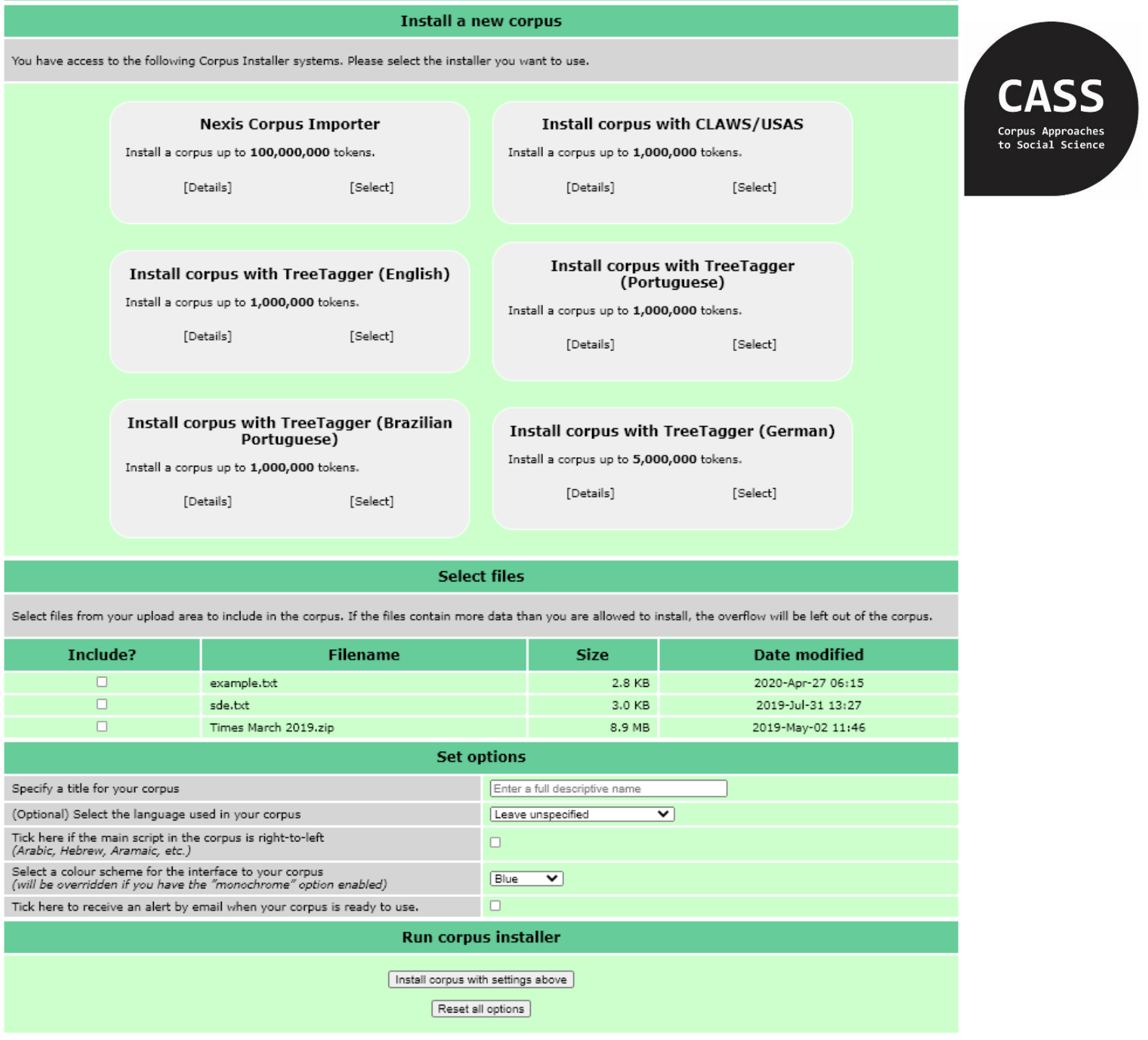
Annotator and Corpus Installer

Existing Plugins				
ID	**Type**	**Description**	**Class**	**Extra setup**
1	CorpusInstaller	Nexis Corpus Importer	NexisCorpusCreate	**semtag-resources** => "/opt/ucrel/usas/share/semtag" **annotation_template_id** => "4" **xml_template_id** => "11" **nexis_script_location** => "/home/hardiea/code-checkouts/build-tools/nexis-to-corpus.php"
2	CorpusInstaller	Install corpus with CLAWS/USAS	StandardToolInstaller	**tool** => "UCREL" **semtag-resources** => "/opt/ucrel/usas/share/semtag" **annotation_template_id** => "4" **xml_template_id** => "2"
3	CorpusInstaller	Install corpus with TreeTagger (English)	StandardToolInstaller	**tool** => "TreeTagger" **language** => "english" **annotation_template_id** => "2" **xml_template_id** => "2"
4	CorpusInstaller	Install corpus with TreeTagger (Portuguese)	StandardToolInstaller	**tool** => "TreeTagger" **language** => "portuguese" **annotation_template_id** => "2" **xml_template_id** => "2"
5	CorpusInstaller	Install corpus with TreeTagger (Brazilian Portuguese)	StandardToolInstaller	**tool** => "TreeTagger" **language** => "portuguese-br" **annotation_template_id** => "2" **xml_template_id** => "2"
6	CorpusInstaller	Install corpus with TreeTagger (German)	StandardToolInstaller	**tool** => "TreeTagger" **language** => "german" **annotation_template_id** => "2" **xml_template_id** => "2"

Annotator configuration

Corpus setup configuration

Interface


Install a new corpus
You have access to the following Corpus Installer systems. Please select the installer you want to use.
Nexis Corpus Importer
Install a corpus up to 100,000,000 tokens.
[Details] [Select]
Install corpus with CLAWS/USAS
Install a corpus up to 1,000,000 tokens.
[Details] [Select]
Install corpus with TreeTagger (English)
Install a corpus up to 1,000,000 tokens.
[Details] [Select]
Install corpus with TreeTagger (Portuguese)
Install a corpus up to 1,000,000 tokens.
[Details] [Select]
Install corpus with TreeTagger (Brazilian Portuguese)
Install a corpus up to 1,000,000 tokens.
[Details] [Select]
Install corpus with TreeTagger (German)
Install a corpus up to 5,000,000 tokens.
[Details] [Select]
Select files
Select files from your upload area to include in the corpus. If the files contain more data than you are allowed to install, the overflow will be left out of the corpus.
Include?
Filename
Size
Date modified
example.txt
2.8 KB
2020-Apr-27 06:15
sde.txt
3.0 KB
2019-Jul-31 13:27
Times March 2019.zip
8.9 MB
2019-May-02 11:46
Set options
Specify a title for your corpus
Enter a full descriptive name
(Optional) Select the language used in your corpus
Leave unspecified
Tick here if the main script in the corpus is right-to-left
(Arabic, Hebrew, Aramaic, etc.)
Select a colour scheme for the interface to your corpus
(will be overridden if you have the "monochrome" option enabled)
Blue
Tick here to receive an alert by email when your corpus is ready to use.
Run corpus installer
Install corpus with settings above
Reset all options

Once indexed…

CASS
Corpus Approaches to Social Science

Menu
Your account
Overview
Settings
Macros
Corpus permissions
Your files and corpora
View your corpora
Install a new corpus
View install queue
Manage your files

CQPweb User Page

Your installed corpora

Corpus	Indexing date	Size			Disk space	Actions
		Tokens	Types	Texts		
dsaaedx (_00137)	2019-Jun-10 06:12	4,032,258	87,021	5,758	138 MB	[Delete corpus]
A test! 9 (_0016e)	2020-Jun-25 07:20	494	273	1	380 KB	[Delete corpus]

Your corpus data storage allowance

Amount of disk space your installed corpora currently take up:	139 MB
Your disk space allocation for installed corpora:	10,240 MB

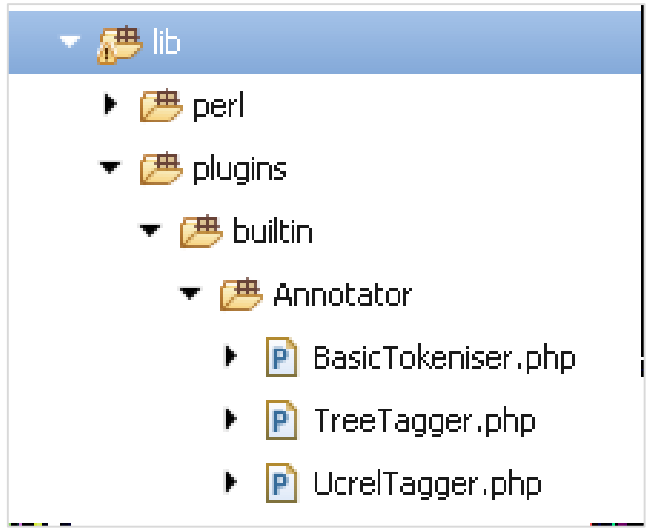
How do plugins work?

- Each type has certain rules about what it *receives* and what it *produces*
  - Downloader
    - Receives a concordance from command-line CQP
    - Produces the file that the user gets as download
  - Annotator
    - Receives input text (file uploaded by user)
    - Produces vertical format text for import
  - Installer
    - Receives list of input files
    - Produces settings for corpus installation plus annotated versions of the files.

How do plugins work?

- Write a PHP file containing a Class for the plugin
- Put the file in the right place in CQPweb's code

```php

/**
 * Annotator plugin for the UCREL toolchain with CLAWS/USAS.
 *
 * Note this assumes the UCREL toolchain is installed as it is
 * on our systems at Lancaster.
 *
 * Elsewhere, it probably won't work without tweaks.
 */
class UcrelTagger extends AnnotatorBase implements Annotator
{
    private $script = 'ucrel-tagger-toolchain';
    private $script_opts = [];


    // TODO in future there will be 4 modes, depending on the Claws resources (maybe also USAS resoruces? )

    const MODE_CLASSIC = 1; /* classic lexicon used for all those years in Wmatrix etc., based on Written BNC1994 */
    const MODE_WRITTEN = 2; /* including new Writ BNC2014 vocab (added to classic lexicon) */
    const MODE_SPOKEN  = 3; /* including new Spok BNC2014 vocab (added to Old BNC Spoken lexicon) */
    const MODE_EMODENG = 4; /* including new Shakespearean vocab (added to classic lexicon) */
    // all depends on the toolchain knowing about them,  of course.


    private $mode = self::MODE_CLASSIC;

    public function __construct(array $extra_config = [])
    {
        /* the only extra params we accept are (a) options for ucrel-tagger-toolchain, or (b) a few specials. */
        foreach($extra_config as $param => $val)
        {
            switch($param)
            {
            case '_special_use_this_script':
                $this->script = realpath($val);
                break;
```



```php
/**
 * Interface for CorpusInstaller Plugins.
 *
 * A CorpusInstaller Plugin is a driver for corpus setup by users.
 *
 * It carries out all steps in the procedure. This may include running an Annotator,
 * checking file formats, doing cleanup, etc. Files may be generated anew,
 * or created from operations on the user's existing files.
 *
 * The plugin does not actually run cwb-encode and friends.
 * But it does supply options for setup based on what it's done.
 */
interface CorpusInstaller extends CQPwebPlugin
{
    /**
     * Make the plugin restrict the amount of data by tokens.
     * If 0 or a negative limit is set, no restriction at all applies.
     * @param int $max
     */
    public function set_max_input_tokens(int $max) : void;


    /**
     * Tell the plugin what the CQPweb handle ofr the corpus will be.
     * (Necessary for it to generate the correct SQL statements.)
     * @param string $name    A lowercase CQP-corpus name.
     */
    public function set_corpus_name(string $name) : void;


    /**
     * Ascertain whether this CorpusInstaller needs input files
     * (some plugins generate data externally). This controls
     * whether the file selector is enabled/disabled, and whether
     * any input files are passed in.
     *
     * @return bool  True if input files are needed.
     */
    public function needs_input_files() : bool;

    /**
     * Add a given path to the list of files to be used as input to the installation.
     * @param string|array $path    A path, or an array of paths.
     */
    public function add_input_file($path) : void;

    /**
     * Run setup - that is,  anything that needs to be done to get files
     * ready to be encoded. This might include tagging, or even building a corpus.
     *
     * @return bool     True if setup worked OK; false if not.
     */
    public function do_setup() : bool;

```

INTERFACES

Define functions that a plugin Class must provide

Only PHP?

CASS
Corpus Approaches to Social Science

- Right now, yes
- RFace
- PyFace

RFace in action: Log Ratio

CASS
Corpus Approaches to Social Science

```
1276 /**
1277  * Calculates a Z-unit to be used as the offset for the LR confidence interval.
1278  *
1279  * @param  float $alpha   The alpha to use. Caller should adjust for familywise if necessary. Defaults to 0.05.
1280  * @param  RFace $r       An R Face object to use. Defaults to null, in which case the function starts an R slave itself.
1281  * @return float          The Z unit (to use, for instance, embedded into an SQL query).
1282  */
1283 function calculate_Z_for_LR_confinterval(float $alpha = 0.05 , ?RFace $r = null) : float
1284 {
1285     switch ($alpha)
1286     {
1287     /* optimise for some frequently-used cases
1288      * (see calculate_LL_threshold() for more notes on this!)
1289      *
1290      * alpha here changes to the CI width (0.05 = 2.5% each way...)
1291      */
1292     case (float)'0.05':         $z_unit =  1.959964;    break; /* 95% CI; in R: qnorm(0.025, lower.tail=FALSE) evals to 1.959964 */
1293     case (float)'0.01':         $z_unit =  2.575829;    break; /* 99% CI; etc. */
1294     case (float)'0.001':        $z_unit =  3.290527;    break;
1295     case (float)'0.0001':       $z_unit =  3.890592;    break;
1296     case (float)'0.00001':      $z_unit =  4.417173;    break;
1297     case (float)'0.000001':     $z_unit =  4.891638;    break;
1298     case (float)'0.0000001':    $z_unit =  5.326724;    break;
1299
1300     default:
1301         global $Config;
1302         $internal_r = is_null($r);
1303         if ($internal_r)
1304             $r = new RFace($Config->path_to_r);
1305
1306         list($z_unit) = $r->read_execute(sprintf("qnorm(%E, lower.tail=FALSE)", $alpha/2.0));
1307
1308         if ($internal_r)
1309             unset($r);
1310
1311         break;
```

PyFace… coming soon honest

CASS
Corpus Approaches to Social Science

- Plugin writing in R / Python
  - Write a very small PHP class which fulfils the Plugin Interface
  - Make the PHP class pass commands to an RFace or PyFace
  - Write the main program in [language of choice]
  - Retrieve the results from the RFace / PyFace
  - Make the PHP class return the result to CQPweb

Thank you for your attention