Manual for CLMET 3.1

1 Introduction

The *Corpus of Late Modern English Texts* (CLMET) is a corpus of roughly 35 million words of British English from 1710–1920, grouped into three 70-year periods (De Smet 2005; Diller et al. 2011). The history, versions and specifics of corpus composition can be followed up by referring to the CLMET3.0 website. CLMET3.0 is currently distributed in three formats: (i) plain text, (ii) plain text with one sentence per line, and (iii) a tagged version (one sentence per line).

Version CLMET3.1 is the result of making CLMET available in a CQP format for use in CWB and CQPweb-based corpus environments (Evert & Hardie 2011; Evert 2010a). While there is no change to the selection of texts, CLMET3.1 includes additions and changes in linguistic annotation. The changes in CLMET3.1 are of three general types: (a) retokenization and retagging, (b) fixing of some systematic issues that come with historical data, and (c) enhancing annotation by adding lemmas and simplified part-of-speech class tags (see §3.3). The types of improved and additional linguistic annotation are documented below.

2 Contact & citation

Questions about CLMET in general: Hendrik De Smet, Hendrik.DeSmet@arts.kuleuven.be
Questions about CLMET3.1/CQP (con)version: Susanne Flach, susanne.flach@fu-berlin.de

If you use CLMET3.1, please cite as follows:
De Smet, Hendrik, Susanne Flach, Jukka Tyrkkö & Hans-Jürgen Diller. 2015. *The Corpus of Late Modern English (CLMET), version 3.1: Improved tokenization and linguistic annotation*. KU Leuven, FU Berlin, U Tampere, RU Bochum. Available from https://perswww.kuleuven.be/~u0044428/clmet3_1.htm.

3 Annotation & clean-up

CLMET3.1 is based on the CLMET3.0 plain text files. The conversion of these files with available tools is documented in §3.1. The annotation levels for CLMET3.1 are described in §3.2–§3.4. Details on the CQP conversion can be found in §3.5.

3.1 Cleaning, tokenization & tagging

- **Pre-processing:** The following changes were made to the original files:
    - CLMET3_0_3_266: file is a 24,000 word duplicate of parts of CLMET3_0_3_268. Comments were added to the `<comment>`-tag in the header of both files.
    - CLMET3_0_2_104: file contains double quotes in `<notes>`-tag. These interfere with xml-styles and were changed to single quotes.
    - CLMET3_0_3_198: “a< far” changed to “as far” (interferes with tagging).
- **Cleaning OCR errors:** a number of characters that are possibly OCR and/or file format conversion errors between platforms (e.g. *donít* ‘don’t’) were corrected. This included removing all underscores, which presumably marked italics, but which is not consistent.

- **Insertion of paragraph tags:** text files in CLMET3.0 have empty lines that correlate with paragraph boundaries. Paragraph tags were written to empty lines (`<p>`). This is only an approximation and mirrors texts representation in OTA/Project Gutenberg, not necessarily the texts' typography in their original publication.
- **Page breaks:** a non-exhaustive heuristic (!) was employed to identify page breaks (*[Page: ixx]*) to avoid these items being recognized as actual text tokens (`<page>`).
- **Sentence recognition & tokenization:** text was tokenized with DocumentPreprocessor, part of the coreNLP (Manning et al. 2014). Its output is one sentence per line, which allowed for the inclusion of sentence boundary tags (`<s>`).
- **Tokenization errors clean up:** erroneous tokenization of historical forms was undone for *'d* (*cou 'd* > *cou'd*, *oblig 'd* > *oblig'd*), then re-tokenized for PRONOUN*'s* (*you'd*, *she'd*). Tokens were put on one line to be tagged with TreeTagger (Schmid 1994).
- **Tagging & lemmatization:** TreeTagger's built-in tokenizer was switched off for tagging. The tagger's lexicon file was extended to include about 20 historical variants of modals (e.g., *wou'd*, *canst*, or *mayest*), ~70 types of 2.SG forms (*saidst*, *dost*), ~2,900 participial or adjectival forms ending in *'d* (*wreck'd*, *restrain'd*), and ~500 *-th*-verb forms (*visiteth*, *blesseth*). In addition, frequent forms (*to-day*, *thy*) were added as were forms like *'tis*, *'twas*, or *'twould* (tags **`VBZ`**, **`VBD`**, **`MD`**; lemmas *it_be*, *it_would*, etc.).

NOTE: The additions to TreeTagger were manually created in an iterative fashion by looking at what the TreeTagger could not identify, then adding the most frequent such forms to the tagger's lexicon, then retagging, and re-identifying unknown lemmas, etc. Once 'complete', all lemmas which the TreeTagger could not identify were replaced by lower case versions of their corpus strings. Also note that the TreeTagger performs rather poorly on capitalization, which TreeTagger often identifies as proper names (**`NP`**).

3.2 Text-level information (*structural attributes*)

CWB can deal with extensive text-structural annotation (*structural attributes*). CLMET3.1 contains `<p>` and `<s>` (see above), as well as meta information extracted from the 333 file headers. The latter are enclosed in `<text>` tags (e.g., *file id*, *author*, *year*, *title*) and technical information on the files in `<file>` tags (e.g., *source URL*, *comments*).

3.3 Token-level annotation (*positional attributes*)

In contrast to structural attributes, *positional attributes* refer to individual tokens. CLMET3.1 has `pos`, `lemma`, and `class` as positional attributes (in addition to `word`, the corpus form).

3.3.1 POS-tagging & lemmatization

Dealing with historical data, tagging remains problematic in all areas, and should be treated with caution (especially with noun recognition) and/or combined with more coarse-grained `class` queries (see below). Also bear in mind that the lemmas for unknown items are in lower case, while proper names that the tagger *did* recognize are not necessarily all lower case. In addition, lemmatization may not be consistent, e.g. in the area of *-ize*/*ise* spellings; these were not homogenized to preserve as much of the original orthography as possible. The use of regular expressions or CQP's case-insensitivity operators (`%c` or `%cd`) is always highly recommended.

The tagset used in CLMET3.1 is the Penn Tagset, with the following major changes and additions (see detailed list in the Appendix and in the CQP `.info` file):

i. Negation particles *not* and *n't*: the tag for these strings was changed from **RB** to **XX0** to set it off from general adverbs, which are also tagged **RB**. **XX0** is the negation tag used in the CLAWS tagset (e.g. BNC). In addition, the lemma for both variants is *not*, as TreeTagger lemmatizes the two forms separately as *not* and *n't*, respectively.
ii. Non-sentence-final punctuation (, ; :) was assigned **PUN**; sentence-final punctuation has **SENT** (the PennSet default).
iii. Left and right brackets received **LBR**- and **RBR**-tags, respectively; quotation mark tags were assigned **LQUO**/**RQUO**. Both tag pairs must be used with caution: opening and closing tags are not always pairwise and/or may have been missed or simply guessed by TreeTagger. (Using their **class**-value in these cases might be better.)
iv. Currency symbols were assigned **CURR**.

3.3.2 CLASS tags

A **class** scheme was added to CLMET3.1 based on the XML version of the BNC corpus. This includes 11 so-called 'Oxford simplified wordclass tags' (Burnard 2007) and subsumes groups of **pos** tags under their more general word classes, i.e. all verb tags (**VVI**, **VBN**, **VDD**, **VM0**, **VHI** etc.) are assigned the simplified class tag **VERB**, or all noun tags **SUBST**. This strategy was also applied to the Penn set (see Appendix), though CLMET3.1 uses a slightly modified set (**ADJ**, **ADV**, **ART**, **CONJ**, **INTJ**, **PREP**, **PRON**, **PUNC**, **SUBST**, **SYM**, **UNC**, **VERB**). The wordclass tags allow for coarse-grained searches, and are, especially with respect to the inaccurate tagging of historical data, a major strategy for workarounds in terms of precision and recall.

3.4 Additional time period attributes

The original meta information on year and period were retained; in addition, the attributes **decade** and **quartcent** were added, according to the following criteria:

i. DECADE:

   a. If a text can unambiguously be attributed to a year, it was placed in the respective decade; this also includes text that has a range within a single decade, i.e. text with a year value of *1844* is assigned the decade value *1840s*, as is text with a year value of *1866-9*.

   b. If the year value spans several decades, the range is given in full, preceded by X, i.e. the year value *1859-60* is given as *X1859-60*. This leaves it up to the analyst whether to subsume such a text as 1850s or 1860s.

ii. QUARTER CENTURY:

   a. 25-year-spans: Text with unambiguous dating is assigned its quarter century, beginning with 1700–1724.

   b. If dating is unclear or extends a range, the range value is given (*X1868-1894*)

ATTRIBUTE	VALUE	DESCRIPTION
Decade	1710s	1710–1719
text_decade	1720s	1720–1729
	…	…
	X1859–60	Text spanning decades, preceded by X

	?1770s	Unsure dating, if year value is ?1770
Quarter century `text_quartcent`	1700–1724 1725–1749 …	25 year intervals for clear and unambiguous quarter centuries
	X1746–1771	Text spanning several quarter centuries

Note: CQPweb does not allow ? in metadata, thus a conjectured time value has a preceding u *(for unknown), i.e. u1775 instead of ?1775.*

3.5 CQP conversion

The tokenized, tagged, lemmatized, and classed vertical file (in UTF-8) was indexed and compressed with CWB-3.0 (Evert 2010b). If you want to index CLMET3.1 yourself in CWB-3.4.x or for CQPweb, use the accompanying **`clmet.vrt`** file (desired metadata for CQPweb can be selected from the accompanying metadata spreadsheet [sheet CQPweb]).

4 CLMET3.1: Formats

CLMET3.1 is available in two major formats: (i) a CQP-version, and (ii) individual .txt files; the latter were converted from **`clmet.vrt`** and contain pos-tagged, class-tagged, and plain versions in 'one-sentence-per-line' formats. Thus, CLMET3.1 comes in most formats that are known from CLMET3.0, but direct comparison to the tagged and plain versions of CLMET3.0 should be avoided primarily due to different tagging/tokenizing strategies.

4.1 CQP version

The CQP folder contains all relevant data for instant use in CQP, i.e. the compressed data, a **`clmet`**-registry file and an **`.info`** file. The doc folder also includes **`clmet.vrt`**. While the **`clmet.vrt`** file is not necessary to run CLMET3.1 in CQP, it may have to be used in the setup for CQPweb or for conversion to CWB-3.4. and above.

You can download CWB/CQP for all major platforms from the CWB website (http://cwb.sourceforge.net). The website offers several demo corpora that were uploaded for use in the CQP query language tutorial, so simply follow the instructions in the tutorial on how to add CLMET to a CQP environment. Remember to update the file path to your data folder in the **`clmet`**-registry file to enable access to the corpus (Windows probably requires the file path in quotes).

4.2 Tagged version

The CWB file **`clmet.vrt`** was converted into a 'one-sentence-per-line' format for all 333 files (removing **`<s>`** tags, but preserving **`<p>`**). The **`pos`** tags are separated by an underscore, i.e. **`I_PP hope_VBP I_PP may_MD be_VB forgiven_VBN ,_PUN`** (as in CMET3.0). The xml-style **`<text>`** and **`<file>`** tags were converted back to the format familiar from CLMET3.0 (a **`<file>`** tag containing the file name was added, as were **`<decade>`** and **`<quartcent>`** with their respective values). The **`<page>`** tags introduced for CQP were retained in the horizontal versions, but were changed to self-closing xml tags: **`<page type="[PAGE ELEMENT AS IT APPEARED IN TEXT]"/>`**.

4.3 Classed version

The 'class' format includes the simplified wordclass tags separated from their tokens by an underscore, i.e. `I_PRON hope_VERB I_PRON may_VERB be_VERB forgiven_VERB ,_PUNC`. Otherwise this version is identical to the tagged version.

4.4 Plain version

A plain version was created by stripping all token-level annotation. Note, however, that this version is of course still tokenized (i.e., CLMET3.1's tokenization), i.e. clitics and punctuation are treated as separate tokens and set off by whitespace, i.e. `Where I may dose out what I 've left of life ,`. Otherwise, this version is identical to the tagged and class version.

5 Comparison

Cleaning and tokenization changed the number of tokens of CLMET. The final version contains 34,342,857 tokens, excluding punctuation (40,340,760 with punctuation). While the actual numerical difference is not that large to CLMET3.0, the qualitative differences on all levels of annotation are considerable. An overview of the token count can be found in the Appendix (and in the CQP info file). Two comparison frequency lists of string and verb tokens are provided in a spreadsheet document. Some examples are pointed out below.

While tokenizing and tagging should of course always only be used with great caution, the increased usability applies most notably to the verbal area. For instance, manual post-editing ensured that *wou'd* is recognized as *would* (rather than `wou_NN 'd_MD` in CLMET3.0). The manual retokenization of (VERB)*'d* decreased *'d* tokens in the corpus from 30,814 in CLMET3.0 to 3,635 in CLMET3.1 (–90%). In the former version, only 10.2% of *'d* tokens were preceded by pronouns (and most of them erroneously tagged), indicating that *'d* was not correctly tokenized; conversely, most preceding verb stems tagged as `NN`. CLMET3.1 now has ~99% pronouns or nouns preceding the token *'d*.

Similar improvements are observable in the non-modal verbal area, where lemmatization now allows for disambiguation of the very general verb tags (`VB.*`) to find only forms of, e.g. *be*, *do*, or *have* by combining attributes in CQP's powerful query language (e.g., finding their base form with `[pos="VB" & lemma="(be|have|do)"]`).

Cliticized forms have also changed considerably in corpus frequency: *'m* from 5,884 to 7,035 (+19%), *'s* from 135,217 to 150,881 (+9%), *'ve* from 4,866 to 5,684 (+16%), and *n't* from 35,664 to 40,234 (+12%). On the other hand, including the non-exhaustive, heuristic `<page>` tag strategy decreased the occurrence of the lemma *page* from 23,349 to 6,157 (–74%).

6 References

Burnard, Lou. 2007. Reference Guide for the British National Corpus (XML Edition). *British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services*. http://www.natcorp.ox.ac.uk/docs/URG/.

De Smet, Hendrik. 2005. A corpus of Late Modern English texts. *ICAME Journal* 29. 69–82.

Diller, Hans-Jürgen, Hendrik De Smet & Jukka Tyrkkö. 2011. A European database of descriptors of English electronic texts. *The European English Messenger* 19. 21–35.

Evert, Stefan. 2010a. CQP query language tutorial (CWB Version 3.0). http://cwb.sourceforge.net/.

Evert, Stefan. 2010b. Corpus Encoding Tutorial. http://cwb.sourceforge.net/.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference*, 1–21. University of Birmingham. http://www.stefan-evert.de/PUB/EvertHardie2011.pdf.

Flach, Susanne. *CQP — A practical guide*. (Draft version 0.2). http://bit.ly/sflach.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

7 Appendix

7.1 Tagset (positional attributes)

POS	CLASS	DESCRIPTION
CC	**CONJ**	coordinating conjunction
CD	**ADJ**	cardinal number
DT	**ART**	determiner
EX	**PRON**	existential there
FW	**UNC**	foreign word
IN	**PREP**	preposition
JJ	**ADJ**	adjective
JJR	**ADJ**	adjective, comparative
JJS	**ADJ**	adjective, superlative
MD	**VERB**	modal
NN	**SUBST**	noun, singular or mass
NNS	**SUBST**	noun plural
NP	**SUBST**	proper noun, singular
NPS	**SUBST**	proper noun, plural
PDT	**ADJ**	predeterminer
POS	**UNC**	possessive ending
PP	**PRON**	personal pronoun
PP$	**PRON**	possessive pronoun
RB	**ADV**	adverb
RBR	**ADV**	adverb, comparative
RBS	**ADV**	adverb, superlative
RP	**ADV**	particle
SENT	**PUNC**	end punctuation
SYM	**SYM**	symbol
TO	**PREP**	to go, to him
UH	**INTJ**	interjection
VB	**VERB**	verb, base form
VBD	**VERB**	verb, past
VBG	**VERB**	verb, gerund/participle
VBN	**VERB**	verb, past participle
VBZ	**VERB**	verb, pres, 3rd p. sing
VBP	**VERB**	verb, pres, non-3rd p.
WDT	**ART**	wh-determiner
WP	**PRON**	wh-pronoun
WP$	**PRON**	possessive wh-pronoun
WRB	**ADV**	wh-abverb
XX0	**ADV**	not, n’t
CURR	**SYM**	£, $
PUNCTUATION TAGS (different to Pennset!):		
PUN	**PUNC**	punctuation which is not SENT
LQUO	**PUNC**	opening quotes
RQUO	**PUNC**	closing quotes
BRL	**PUNC**	left brackets
BRR	**PUNC**	right brackets
LS	**PUNC**	list item

7.2 Attributes & values (structural attributes)

Attribute	Value	Description
`text_file`	*CLMET text identifier*	CLMET text identifier
`text_id`	*CLMET text id*	1, 2, 3, 4, 5, …
`text_period`	1710-1780 1781-1750 1851-1920	
`text_year`	1788 ?1775 1759-67	Unambiguous year conjectured range
`text_decade`	1710s, 1720s, 1730s X1859-60 1770s?	Unambiguous decade Text spanning decades, preceded by X Unsure dating, if year value is ?1770
`text_quartcent`	1700–1724 1725–1749 … X1750–1778	25 year intervals for clear years If spanning quarter centuries
`text_author`	*Name of author*	
`text_gender`	M F	
`text_author_birth`	*Year*	
`text_title`	*Title of text*	
`text_genre`	Narrative fiction Other Narrative non-fiction Treatise LET Drama	
`text_subgrenre`	fict treat x hist let bio drama hist / treat travel / treat bio / travel bio / treat fict / treat fict / travel	
`text_notes`	*Notes by editors*	
`file_source`	*URL of source text*	
`file_downloaded`	*Date of download*	
`file_comments`	*corpus compiler's comment*	
`page_type`	*Page number or comment*	e.g. [Page ii] (heuristic)

7.3 Word & token count

The following tables summarize and compare the numerical corpus make-up across the two versions (count data for CLMET3.0 was taken from the official corpus documentation, count data for CLMET3.1 calculated using the CWB count tools):

PERIOD	#authors	#texts	CLMET3.1 (all)	CLMET3.1 (excl. punct.)
1710-1780	51	88	12,155,135	10,460,119
1780-1850	70	99	13,268,542	11,473,445
1850-1920	91	146	14,834,182	12,942,471
TOTAL	212	333	40,257,859	34,876,035

GENRE:	CLMET3.0			CLMET3.1		
	1710-1780	**1780-1850**	**1850-1920**	**1710-1780**	**1780-1850**	**1850-1920**
Narrative fiction	4,642,670	4,830,718	6,311,301	5,405,645	5,780,352	7,561,339
Narrative non-fiction	1,863,855	1,940,245	958,410	2,145,946	2,261,485	1,097,487
Drama	407,885	347,493	607,401	523,318	441,040	763,352
Letters	1,016,745	714,343	479,724	1,208,219	842,795	554,046
Treatise	1,114,521	1,692,992	1,782,124	1,263,090	1,927,272	2,030,210
Other	1,434,755	1,759,796	2,481,247	1,635,846	2,047,513	2,851,805
TOTAL	**10,480,431**	**11,285,587**	**12,620,207**	**12,182,064**	**13,300,457**	**14,858,239**

7.4 CQP data model

Details of the CQP data model are described in Evert (2010a; 2010b). The following illustrates the first few (and last 2) lines of **clmet.vrt**. Positional attributes, i.e. token-level annotation, are represented as tab-separated information. Structural attributes variably apply to the text, sentence, or token-level (i.e. **<text>**, **<file>**, **<s>**, **<p>**, and **<page>** tags).

```
<text id="1" file="CLMET3_1_1_1.txt" period="1710-1780" quartcent="1700-1724"
    decade="1710s" year="1710" genre="Treatise" subgenre="treat"
    author="Berkeley, George" gender="M" author_birth="1685" notes="">
<file source="http://ota.ahds.ac.uk/text/4634.html" downloaded="25-09-2012"
    comments="">
<p>
<s>
A             DT     a             ART
TREATISE      NN     treatise      SUBST
Concerning    VBG    concern       VERB
the           DT     the           ART
PRINCIPLES    NP     principles    SUBST
OF            IN     of            PREP
Human         NP     Human         SUBST
Knowlege      NP     knowlege      SUBST
.             PUN    .             PUNC
</s>
</p>
   ...
</file>
</text>
```

Notes: Strings that TreeTagger identified as <unknown> are assigned their lower-cased string as lemma (here: *knowlege*). Snippet also illustrates the issues of (proper) noun tagging.