

README

Anna Čermáková
Centre for Corpus Research
College of Arts and Law
University of Birmingham
a.cermakova@bham.ac.uk

Anthony Hennessey
Statistics and Probability
School of Mathematical Sciences
University of Nottingham
anthony.hennessey@nottingham.ac.uk

Jamie Lentin
Shuttle Thread
Manchester
jamie.lentin@shuttlethread.com

Michaela Mahlberg
Centre for Corpus Research
College of Arts and Law
University of Birmingham
m.a.mahlberg@bham.ac.uk

Viola Wiegand
Centre for Corpus Research
College of Arts and Law
University of Birmingham
v.wiegand@bham.ac.uk

Contents

1	Corpora	1
1.1	ChiLit - Children's Literature	1
1.1.1	Included texts	1
2	Cleaning of corpora texts	4
2.1	Converting to 7-bit ASCII	5
3	Maintaining this corpora repository	6
3.1	.bib file	6
3.2	Adding a new text to a corpus	6
3.3	Adding a new corpus	6
3.4	Repository Tags	6
3.5	Updating README.pdf	6

1 Corpora

1.1 ChiLit - Children's Literature

From work by Anna Čermáková.

1.1.1 Included texts

- Anstey, F. (1882). *Vice Versa; or, A Lesson to Fathers* (D. Clarke, M. Pettit & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/26853>
- Anstey, F. (1900). *The Brass Bottle*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/30689>
- Ballantyne, R. M. (1858). *The Coral Island: A Tale of the Pacific Ocean* (D. Price, Ed.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/646>
- Barrie, J. M. (1911). *Peter Pan*. Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/16>
- Burnett, F. H. (1911). *The Secret Garden*. Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/113>
- Carroll, L. (1865). *Alice's Adventures in Wonderland*. Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/11>
- Carroll, L. (1871). *Through the Looking-Glass*. Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/12>
- Crockett, S. R. (1897). *The Surprising Adventures of Sir Toady Lion with Those of General Napoleon Smith*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/39340>
- Ewing, J. H. G. (1869). *Mrs. Overthway's Remembrances*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/17772>
- Ewing, J. H. G. (1883). *Jackanapes*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/20351>
- Falkner, J. M. (1898). *Moonfleet*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/10743>

- Farrar, F. W. (1858). *Eric; Or, Little by Little*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/12083>
- Farrow, G. E. (1898). *Adventures in Wallypug-Land*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/52393>
- Grahame, K. [K.]. (1895). *The Golden Age*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/291>
- Grahame, K. [K.]. (1898). *Dream Days*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/35187>
- Grahame, K. [Kenneth]. (1908). *The Wind in the Willows* (M. Lough, Ed.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/289>
- Haggard, H. R. (1885). *King Solomon's Mines*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/2166>
- Haggard, H. R. (1887). *Allan Quatermain*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/711>
- Henty, G. A. (1882). *Winning His Spurs. A Tale of the Crusades*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/12308>
- Henty, G. A. (1884). *With Clive in India; Or, The Beginnings of an Empire*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/18833>
- Hughes, T. (1857). *Tom Brown's Schooldays*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/1480>
- Ingelow, J. (1869). *Mopsa the Fairy*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/32867>
- Jefferies, R. (1881). *Wood Magic, a Fable*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/25299>
- Kingsley, C. [C.]. (1870). *Madam How and Lady Why; Or, First Lessons in Earth Lore for Children*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/1697>
- Kingsley, C. [Charles]. (1863). *The Water-Babies* (D. Price, Ed.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/1018>
- Kipling, R. [R.]. (1899). *Stalky and Co.* Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/3006>
- Kipling, R. [Rudyard]. (1894). *The Jungle Book* (Anonymous Volunteer & D. Widger, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/236>
- Lang, A. (1889). *Prince Prigio*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/21935>
- MacDonald, G. (1871). *At the Back of the North Wind* (M. Ward, Ed.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/225>
- MacDonald, G. (1872). *The Princess and the Goblin* (S. Shell, Emmy & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/34339>
- Mare, W. D. I. (1910). *The Three Mulla-mulgars*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/32620>
- Marryat, F. [F.]. (1841). *Masterman Ready; Or, The Wreck of the "Pacific"*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/21552>
- Marryat, F. [F.]. (1844). *The Settlers in Canada*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/22496>
- Marryat, F. [Frederick]. (1847). *The Children of the New Forest* (J. Sutherland, C. Franks & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/6471>
- Martineau, H. (1841a). *Feats on the Fiord*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/35892>
- Martineau, H. (1841b). *The Crofton Boys*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/23265>
- Martineau, H. (1841c). *The Peasant and the Prince*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/23275>
- Martineau, H. (1841d). *The Settlers at Home*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/23264>
- Meade, L. T. (1886). *A World of Girls: The Story of a School*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/43147>
- Molesworth, N. M. (1877). *The Cuckoo Clock*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/15569>

- Molesworth, N. M. (1879). *The Tapestry Room: A Child's Romance*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/17175>
- Molesworth, N. M. (1895). *The Carved Lions*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/39549>
- Nesbit, E. (1899a). *The Book of Dragons*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/23661>
- Nesbit, E. (1899b). *The Story of the Treasure Seekers*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/770>
- Nesbit, E. (1901). *Nine Unlikely Tales*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/49913>
- Nesbit, E. (1905). *The Railway Children* (L. Bowler, Ed.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/1874>
- Nesbit, E. (1906a). *Five Children and It* (J. Isbell & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/17314>
- Nesbit, E. (1906b). *The Story of the Amulet*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/837>
- Potter, B. (1902). *The Tale of Peter Rabbit* (R. Cicconetti, R. Holder & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/14838>
- Potter, B. (1903). *The Tale of Squirrel Nutkin* (R. Cicconetti, Emmy & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/14872>
- Potter, B. (1904a). *The Tale of Benjamin Bunny* (R. Cicconetti & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/14407>
- Potter, B. (1904b). *The Tale of Two Bad Mice* (D. Edwards, Emmy & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/45264>
- Potter, B. (1908). *The Tale of Jemima Puddle-Duck* (R. Cicconetti, Emmy & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/14814>
- Potter, B. (1909). *The Tale of the Flopsy Bunnies* (M. Ciesielski & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/14220>
- Reed, T. B. (1887). *The Fifth Form at Saint Dominic's: A School Story*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/24632>
- Ruskin, J. (1851). *The King of the Golden River; or, the Black Brothers: A Legend of Stiria*. (C. Curnow, J. Hollowell & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/33673>
- Sewell, A. (1877). *Black Beauty*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/271>
- Sinclair, C. (1839). *Holiday House: A Series of Tales* (J. Srna, D. Alexander, D. Wilson & Distributed Proofreading Team, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/32811>
- Stevenson, R. L. [R. L.]. (1886). *Kidnapped*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/421>
- Stevenson, R. L. [Robert Louis]. (1883). *Treasure Island* (J. Boss, J. Hamm & D. Widger, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/120>
- Stretton, H. (1867). *Jessica's First Prayer*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/50104>
- Stretton, H. (1868). *Little Meg's children*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/30555>
- Stretton, H. (1869). *Alone in London*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/12172>
- Strickland, A. (1826). *The Rival Crusoes; or The Shipwreck*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/34849>
- Thackeray, W. M. (1855). *The Rose and the Ring*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/897>
- Tytler, A. F. (1870). *Leila at home*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/49220>
- Wilde, O. (1888). *The Happy Prince, and Other Tales* (D. Price & P. Redmond, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/902>
- Yonge, C. M. [C. M.]. (1853). *The Heir of Redclyffe*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/2505>

Yonge, C. M. [C. M.]. (1856). *The Daisy Chain, or Aspirations*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/3610>

Yonge, C. M. [C. M.]. (1866). *The Dove in the Eagle's Nest*. Retrieved September 10, 2017, from <https://www.gutenberg.org/ebooks/3139>

Yonge, C. M. [Charlotte M.]. (1864). *The Little Duke: Richard the Fearless* (J. Haselow, M. Taylor & D. Price, Eds.). Retrieved June 28, 2017, from <https://www.gutenberg.org/ebooks/3048>

2 Cleaning of corpora texts

The sources were the Gutenberg plain text UTF-8 files.

1. Convert to unix line endings.
2. Remove non-authorial text.
3. Reformat the book title and author to make consistent across all texts.
4. Reformat chapter headings to make consistent across all texts.
5. Manual corrections

Steps 2, 3 and 4 were done manually.

Step 1 was achieved using the following command

```
for f in ChiLit/*.txt; do dos2unix -m $f; done
```

Some specifics of step 2:

- Tables of content are removed.
- Lists of illustrations are removed.
- Any preface text attributed to a person other than the author is removed. When attribution is unclear the text is left.
- Any postface text attributed to a person other than the author is removed. When attribution is unclear the text is left.
- Transcriber notes are removed.
- In the texts illustrations are usually indicated by text enclosed in square brackets. Where this text includes a caption the caption is kept, for example

[Illustration: THE WONDERSTONE.]

becomes

[THE WONDERSTONE.]

Where there is no authorial caption the construct is deleted. All the following example would be deleted

[Illustration]

[Illustration: Chapter Seventeen]

[Illustration: Page 91]

- The book title is put on the first line of the file, without any newlines.
- The book author is put on the second line of the file, without any newlines.
- Chapter headings are formatted as follows: If the chapter heading begins with 'CHAPTER' or 'BOOK' it must be followed by a number or roman numerals and then a dot. The chapter or book number cannot be written in word form. The heading can optionally be followed by a chapter title; the chapter title must not break onto a new line. Here are some examples

CHAPTER 1. The Old Sea-dog at the Admiral Benbow

CHAPTER 2. TRAVELLING COMPANIONS.

CHAPTER 3.

CHAPTER IV. Little Meg's Treat to Her Children

CHAPTER V.

BOOK 1.

BOOK II. Jessica's Mother

Sections beginning with 'INTRODUCTION', 'PREFACE', 'CONCLUSION', 'PROLOGUE', 'PRELUDE' or 'MORAL' are also be treated as separate chapters. These do not require numbers, but do require the dot. Again the heading can optionally be followed by a title; the title must not break onto a new line. Here are some examples

PREFACE.

INTRODUCTION.

PROLOGUE. THE OLYMPIANS

MORAL.--_There is no moral to this chapter._

In all cases there must be no space at the beginning of the line.

- Part headings are on a line before the first chapter of that part, in the same format. Blank lines are allowed between the part heading and the chapter heading. The following example is from *Treasure Island*

PART TWO. The Sea-cook

CHAPTER 7. I Go to Bristol

IT was longer than the squire imagined ere we were ready for the sea,
and none of our first plans--not even Dr. Livesey's, of keeping me

2.1 Converting to 7-bit ASCII.

In addition, the Perl module `Text::Unidecode`¹ can be used to unify the use of hyphens, apostrophes and quotes across the texts with the following command

```
perl -C -MText::Unidecode -n -i -e'print unidecode($_)' */*.txt
```

7-bit ASCII² is a subset of UTF-8³ and so the files may be treated as UTF-8. Note that this will also affect accents and other special characters; for example, Hôtel becomes Hotel, archæologist becomes archaeologist and £60,000 becomes PS60,000.

¹ <http://search.cpan.org/perldoc?Text::Unidecode>

² <https://tools.ietf.org/html/rfc20>

³ <https://tools.ietf.org/html/rfc3629>

3 Maintaining this corpora repository

3.1 .bib file

We currently manage the bibliography in a shared zotero folder. The important fields in the bib entries are:

- The shorttitle field must match the filename of the relevant text file in the corpus folder.
- The keywords field must contain the name of the corpus.
- The title, author and date fields must be present.

Example entry:

```
@book{grahame_wind_1908,
  title = {The Wind in the Willows},
  url = {https://www.gutenberg.org/ebooks/289},
  shorttitle = {willows},          <<=== filename willows.txt
  author = {Grahame, Kenneth},
  editor = {Lough, Mike},
  urldate = {2017-06-28},
  date = {1908},
  keywords = {{ChiLit}}           <<=== corpus id
}
```

3.2 Adding a new text to a corpus

1. Clean the text as described in Section 2.
2. Add entry to the .bib file; see Section 3.1.
3. Generate a new coopora.json file.

```
cd scripts
R --no-restore --no-save <./bib2json.R
```

4. Update repository tags; see Section 3.4.

3.3 Adding a new corpus

1. Edit the scripts/bib2json.R script to include the new corpus.
2. For each new corpus file
 - (a) Clean the text as described in Section 2.
 - (b) Add entry to the .bib file; see Section 3.1.
3. Generate a new coopora.json file.

```
cd scripts
R --no-restore --no-save <./bib2json.R
```

4. Update repository tags; see Section 3.4.

3.4 Repository Tags

TODO

3.5 Updating README.pdf

The README pdf is compiled with latexmk

```
sudo apt install latexmk texlive-latex-extra texlive-bibtex-extra biber
latexmk --outdir=./build --pdf scripts/README.tex && cp build/README.pdf .
```