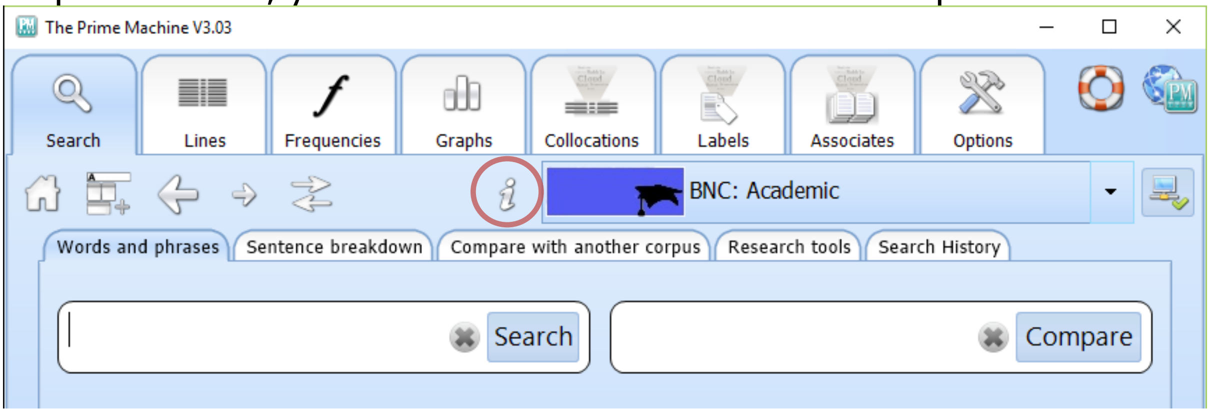
Corpora Available (Offsite)

This document sets out the corpora currently available for offsite access through *The Prime Machine.*

Staff and students at XJTLU can connect to the server using the tPM Home Network, and gain access to a wider range of corpora and other resources. They should refer to **tPM Help 003b Corpora available on tPM Home Network**, which can be found on the XJTLU Staff and Students pages of www.theprimemachine.com.

Information about the currently selected corpus

You can view basic information about the currently selected corpus including its major text categories by clicking on the ℹ symbol to the left of the drop-down menu. If you select the "Compare with another corpus" tab first, you will see information about both corpora.



Fundamental Features

The Prime Machine counts punctuation in its total counts, so figures presented here and in the software itself will look higher than information you will find on corpus websites.

In earlier versions of the tPM databases, spoken corpora contained the name (or speaker label) for each speaker as part of the running text, meaning additional "words" appear at the beginning of each new turn. In dialogues, the Cards for these corpora often only show one "sentence" with names or labels appearing above and below. Currently, the spoken corpora are being updated. Where **v2** appears next to a spoken corpus, it means it is still waiting to be re-processed. When **v3** appears next to a spoken corpus, the #Speaker# label will no longer be treated as running text, and you will typically see wider context on the Cards.

1. The British National Corpus

Corpus		Size	Details
	The British National Corpus	115M	Includes all of the sub-corpora listed below, as one large corpus. Major categories are set up following the classifications (e.g. Academic, Unpublished, Spoken). BNC XML Edition (2007-02-08). For details see: http://www.natcorp.ox.ac.uk/XMLedition/
The BNC has also been loaded as separate sub-corpora, following classifications in Lee (2001), which were provided in the raw BNC XML corpus.			
	BNC: Unpublished	5M	This is a collection of unpublished texts, including essays which have not been printed, letters, non-academic prose, etc. The Major Categories are: Fiction, Letters, Non-Academic (non-fiction), Non-printed essays and Other.
	BNC: Spoken	14M	This is a collection of spoken texts. It is a combination of "Conversation" and "Other Spoken" sub-corpora. The Major Categories are Broadcast, Conversation, Interviews, Lectures, Speeches and Other. ***tPM Spoken V2***
	BNC: Other Publications	21M	This is a collection of other publications, including adverts, letters, magazines, instructions and official proceedings. The Major Categories are: Advertising, Parliamentary proceedings, Official/company documents, Instructional, Personal letters, Professional/Business letters, Popular magazines and Other.
	BNC: Non-Academic	28M	This is a collection of non-academic texts on a variety of topics. The Major Categories are Biography, Commerce, Humanities & Arts, Medicine, Natural Science, Politics, Law & Education, Social Science, Technology & Engineering and Religion.
	BNC: Newspapers	11M	This is a collection of newspaper articles from broadsheet, tabloid and local newspapers. The Major Categories are: Arts, Commerce, Editorial, Report, Science, Social, Sports, Tabloid and Misc. (Miscellaneous).
	BNC: Fiction	20M	This is a collection of fiction, grouped into the Major Categories: Drama, Poetry and Prose.
	BNC: Academic	18M	This is a collection of academic texts, grouped into Major Categories according to a variety of academic disciplines: Humanities & Arts, Medicine, Natural Science, Politics, Law & Education, Social Science and Technology & Engineering.

Other notes:

- Texts from the British National Corpus show source information from Burnard (2007).
- Labels have been edited to try to improve readability of features of the texts and producers (authors/speakers). More information can be found in Jeaco (2015).

2. Hindawi Academic Corpora

Hindawi is a publisher of academic journals. It provides access to these texts for data mining, and the following corpora have been created using Hindawi's open access full-text corpus for text mining research (http://www.hindawi.com/corpus/).

Corpus	Size	Details of Major Categories
Hindawi Mathematics	12.4M	Analysis, Applied Mathematics, Differential & Difference Equations, Discrete Mathematics, General Mathematics and Statistics & Probability.
Hindawi Chemistry	6.2M	Analytical Chemistry, General Chemistry, Inorganic Chemistry, Organic Chemistry, and Physical Chemistry.
Hindawi Physics	6.2M	Acoustics, Condensed Matter Physics, Electricity & Magnetism, General Physics, Heat & Thermodynamics, High Energy Physics, Mechanics, Statics & Dynamics, Nuclear Physics and Optics & Light.
Hindawi Earth Science and Environment	4.7M	Agriculture & the Environment, Astronomy, Ecology, General Earth Sciences, Geomatics, Geophysics, Meteorology & Climatology and Oceanography.
Hindawi Engineering	23.7M	Aerospace & Defence Engineering, Chemical Engineering, Civil Engineering, Electrical, Electronic & Communications Engineering, Engineering General, Manufacturing Engineering, Materials Science & Engineering, Mechanical Engineering & Related Industries and Nanotechnology.
Hindawi Computer Science	9.8M	Computer Applications, Computer Systems Organization, Computing Methodologies, Computing Milieux, Hardware, Information Systems and Software.
Hindawi Social Sciences	4.2M	Journals from a number of different academic disciplines each have their own Major Category: Anthropology, Developmental Psychology, General Education, Human Geography, International Economics and Urban Development.
Hindawi Biological Sciences	23.0M	Journals were grouped for Major Categories through discussion with a colleague from the discipline. Groups are: Anatomy, Biochemistry & Biophysics, Biotechnology, Cell & Molecular Biology, Entomology, Evolution, Genetics, Marine Biology, Microbiology & Microbes and Zoology.

Other notes:

- The Hindawi Social Sciences corpus was created using a newer version of the archive, downloaded on 16/01/2017. Other Hindawi corpora were created from the archive downloaded on 6/11/2013.
- The *intute* website (unfortunately no longer available) was used to identify groupings for journals, either by matching listed journals or through finding similar topics in the journal’s introduction. More details in Jeaco (2015).

3. Fiction and Non-Literary Collections

These corpora were created by using texts obtained from Project Gutenberg: http://www.gutenberg.org/. All of these texts can be obtained directly from the Gutenberg website. Many thanks to the XJTLU research assistances for their help in constructing these corpora.

Corpus		**Size**	Details	**Major Categories**
🟥	**Fiction Collection 12x7**	17M	7 complete novels from each of 12 novelists from roughly Victorian times: Jane Austen, Mary Elizabeth Braddon, Wilkie Collins, Charles Dickens, Maria Edgeworth, George Eliot, Elizabeth Gaskell, Thomas Hardy, Frederick Marryat, Walter Scott, William Makepeace Thackeray and Anthony Trollope.	Author names
🟥	**Fiction Collection 12x7 USA**	10M	7 American novels from each of 12 novelists from roughly the same period as 12x7. The novelists are: James Fenimore Cooper, Hamlin Garland, Nathaniel Hawthorne, William Dean Howells, Henry James, Sinclair Lewis, Jack London, Herman Melville, Upton Sinclair, Harriet Beecher Stowe, Mark Twain and Edith Wharton.	Author names
🟥	**Fiction Collection 37x1**	7M	37 complete Victorian novels, written by 37 different authors, following the list from Mahlberg (2013).	None
🟥	**Fiction Collection 37x1 USA**	4M	37 complete American novels, written by 37 different authors from roughly the same period as 37x1.	None
🟥	**Gothic Fiction Collection**	2.7M	Complete novels or short stories from 23 authors, considered to be Gothic Fiction.	None
🟨	**Non-Literary Collection**	9.4M	Non-literary (Non-Fiction) texts from a variety of (British) authors, from the following text types: Letters, Travel, History, Trials, Speeches & Sermons, Periodicals, Biography, Manuals & Handbooks and Essays.	Following text types listed to the left.
	Non-Literary Collection USA		Non-literary (Non-Fiction) texts from a variety of American authors, from the following text types: Letters, Travel, History, Trials, Speeches & Sermons, Periodicals, Biography, Manuals and Handbooks & Essays.	Following text types listed to the left.

Other notes:

- The novels for the Fiction Collections have been split into separate chapters, and each chapter has been loaded into the corpus as a text.
- Corpora with no Major Categories use *The British National Corpus* as the reference corpus for the calculation of Key Words (and Key Associates).

More details

Approximate size is given in number of words (tokens, including punctuation). M = million.

Some of the background to corpus construction and refactoring for The Prime Machine can be found in the doctoral thesis (Jeaco, 2015, pages 63-66 and 253-261).

References

Burnard, L. (2007). BNC User Reference Guide: List of Sources. Retrieved 12 August, 2013, from http://sara.natcorp.ox.ac.uk/docs/URG/bibliog.html

Jeaco, S. (2015). The Prime Machine: a user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language. Unpublished Ph.D. dissertation, University of Liverpool. Retrieved from https://livrepository.liverpool.ac.uk/2014579/

Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology, 5*(3), 37-72.

Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*: New York ; Routledge, 2013.

Last Updated: Tuesday, March 27, 2018