Virtual Corpora at the Oxford Text Archive

Martin Wynne, Rowan Wilson and Ylva Berglund
Oxford Text Archive, Oxford University

1 Deeper and wider access to electronic text holdings

The Oxford Text Archive (OTA) has been collecting and distributing electronic texts since 1976. It now hosts the Arts and Humanities Data Service (AHDS) Centre for Literature, Languages and Linguistics, and provides a service as a data centre and a source of expertise and advice on the creation, storage, preservation, distribution and use of digital resources.

The AHDS has a commitment to providing improved resource discovery mechanisms for its collections and to its collections. The OTA is committed to keeping up with current and future developments (Wynne 2002) and to offering deeper and wider access to the archive.

In terms of resource discovery and wider access, various initiatives are underway, which are mentioned now but not covered in depth here, such as participation in the Open Archives Initiative, including the Open Language Archives Community (OLAC) and various portals projects.

In terms of deeper access, we aim to make more texts more easily available and to improve online access to the electronic texts in our collection. This paper focuses on the keystone of the OTA's initiative to provide deeper access, the new online *virtual corpus* and online query system for the holdings of the Oxford Text Archive is described, demonstrated and explained.

2 From magnetic tape to open archives

The Oxford Text Archive is a large repository of electronic texts and corpora. In many ways, the archive has worked in much the same way since its inception. The user consults the catalogue, selects a text or a number of texts and then completes the relevant procedure in order to download the text or texts to their computer. The main development in terms of resource delivery in the past 27 years is that many of the resources can now be downloaded directly from the website, rather than being sent by post on magnetic media or downloaded by ftp. Users are then left to their own devices to find software to analyse the texts and try to extract information from them.

In order to make the archive more useful and usable for linguistics researchers, learners and teachers, a system for the querying of texts online has now been developed at the Oxford Text Archive. It is possible for the user to construct a corpus of texts from the archive for downloading or querying online. Concordances and collocation information can be generated and displayed. Texts for this user-defined corpus can be selected on the basis of any of the metadata categories in the resource description, and the text selection can then be refined by simply picking and choosing from a list of texts generated by the initial query.

3 Virtual corpora and online delivery

Online concordancing is not new. Many sites and corpus projects offer this facility. Neither is the ability to select the texts on the fly and thus construct a virtual corpus a new innovation. The novelty of the new OTA service lies in the application of both a flexible text selection functionality and an online query system to a large online archive.

A system of virtual corpus building based on fine-grained and detailed text-level metadata is implemented at the Max Planck Institut in Nijmegen (Wittenberg and Broeder 2002). This differs from the OTA system in that it is primarily for reassembling the contents of language corpora, rather than a means of using an archive of hitherto discrete and non-interoperable texts for virtual corpus building.

There is also some use of the term *virtual corpus* referring to the use of the web as a corpus. Indeed, a logical extension of the concept of virtual corpus building as applied to the Oxford Text Archive could be to apply the concept to texts held in disparate locations on the web. There are certainly initiatives in this direction, one of the most significant of which is the Webcorp system (Renouf 2001). However, researchers currently find several important problems with the use of the web as corpus, particularly in the case of English language texts. One problem is the identification of the language of a page, or text which is part of a page. Another is the amount of non-native speaker writing, and the difficulty of identifying it. Similar and related problems are encountered with regional varieties. There are also strong biases in terms of the types of text which are present on the web, and in terms of the relative frequencies of different text types. For example, computing manuals and pornography are strongly represented, and formal correspondence is rare, in contrast to the relative importance of these text types in the language in general. There is also the problems of the so-called 'invisible web', whereby much textual content is served on the fly to browsers and is not available to the search engines and spider programs which would typically be expected to build corpora on the web.

It is also often difficult to identify any information at all about texts, or at best it is difficult to have any confidence about the accuracy of information. To put it simply, textual metadata is sparse, unreliable and implemented in non-standard ways.

The problem of metadata is addressed by proposals for the semantic web. It will be interesting to see what use linguists can make of this for corpus building and online linguistic analysis. However, even if satisfactory standards for encoding metadata for content on the web can be formulated, there will still be questions about the reliability of the implementation of these standards in any given case.

The semantic web can be seen as one step towards using the web as a corpus, or of making corpora from texts on the web, by establishing interoperable textual metadata. Another step would be to distribute the text processing tasks. Interestingly there is a project to realise the distribution of data processing on the internet, in a initiative known as the Grid:

> *The Grid* is the name that describes the next significant development in Internet computing. A term first coined in the mid '90s to describe a vision for a distributed computing infrastructure for advanced science projects, the Grid was first properly explained by Ian Foster and Carl Kesselman in their book *The Grid: Blueprint for a New Computing Infrastructure*(Morgan Kaufmann, 1999; ISBN 1-55860-475-8). In this vision, the Grid will be to *all* computational resources what the World Wide Web presently is to documents containing information. Grid users will have at their disposal distributed high performance computers able to access and process terabytes of data stored in global databases, plus the appropriate tools to control these resources. The Grid is similar to Tim Berners-Lee's vision of the*Semantic Web*, an information space of meaning built upon the World Wide Web that he invented, but it extends that vision.
>
> At present, the Grid is in a similar state of infancy to that in which the World Wide Web was 10 years ago, when it was little more than "a very slow way to ftp files between distant computers". However, while the World Wide Web is simple to define in terms of a language for organizing information (HTML) and an information transport protocol for delivering it (HTTP), the form that the Grid will take is more complex, and the path of its development unclear. There are even debates about whether there will be different kinds of Grids for different purposes, one concerned primarily with number-crunching tasks on distributed supercomputers, another involved in integrating disparate information resources, private corporate Grids, etc. But that the Grid revolution is almost upon us is without doubt.
>
> (Oxford e.science centre website, http://e-science.ox.ac.uk/)

At the OTA we have the advantage of being closely linked with the latest developments in the Grid and e-science (the term coined for research done using the Grid through a regional e-science centre based alongside the OTA in the Research Technologies Services of the Oxford University Computing Services. Interestingly, much of the early work on setting up the Grid and e-science projects is concerned with authentification, certification and the identification of trusted sites, users and data. It seems that both with e-science and with attempts to do linguistic analysis on the web, a crucial issue is the identification of reliable quality data sources, and this is constraining the more utopian visions of the possibilities for global distributed computing.

At the current juncture then, our ambitions are restricted to working with mature and proven technologies, and therefore we aim to concentrate on providing virtual corpora at the OTA based on texts in our archive. The OTA is a trusted repository of quality texts created for academic research, and so there is some guarantee of text quality. It will however be interesting to watch and experiment ourselves with attempts to widen the concept to texts and tools in distributed locations in cyberspace.

4 Making virtual corpora at the OTA

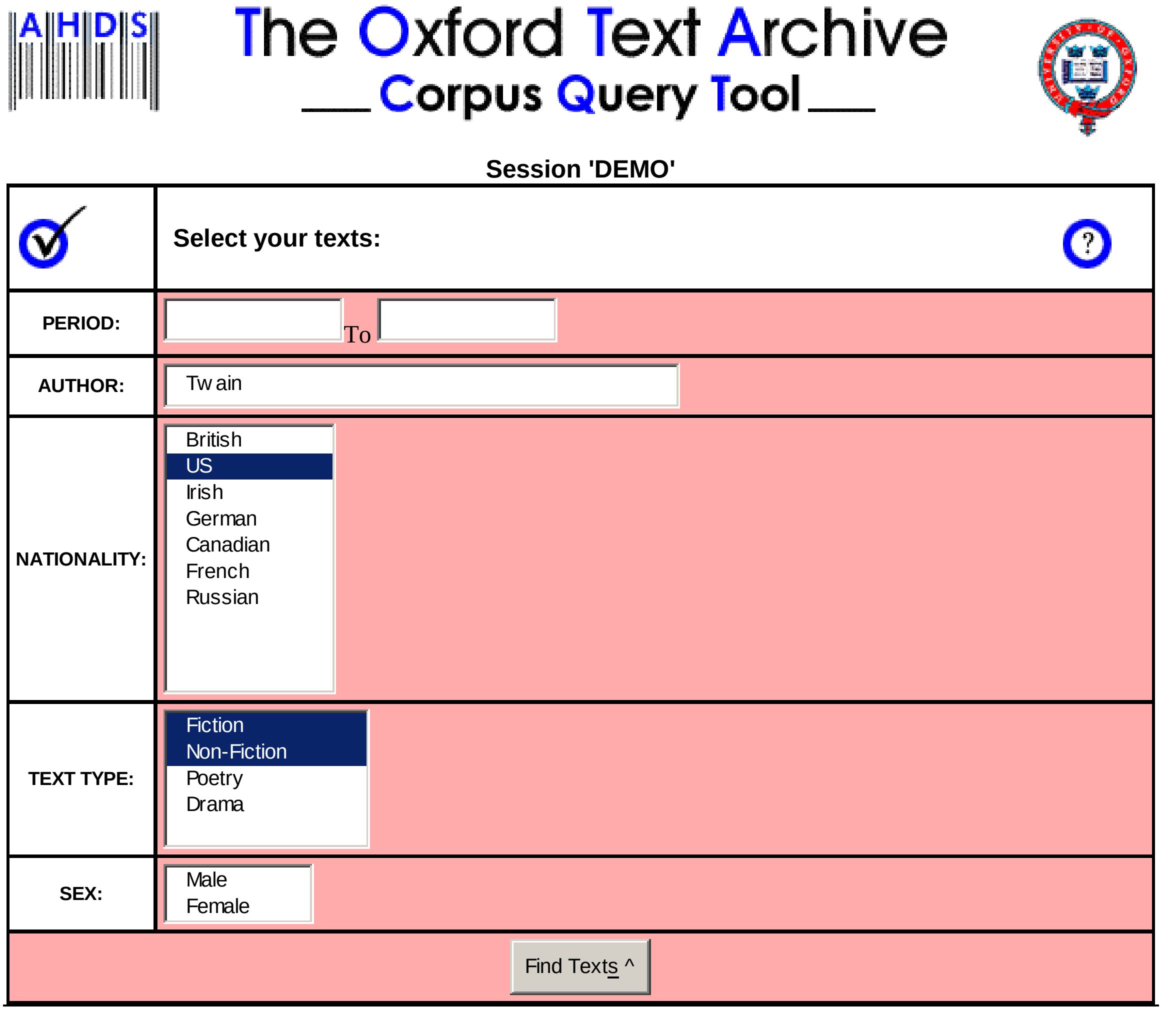
A specific challenge of providing a *virtual corpus* service using the holdings of the Oxford Text Archive lies in the heterogenous nature of texts. There are more than two thousand texts in the archive and they have been collected and documented over a period of more than twenty-five years. They reflect a multitude of different practices in the encoding of the texts, in the construction of collections of texts, and in the documentation of the resources. The size and diversity of the archive makes it a potentially extremely rich linguistic resource. It is however necessary for the text selection and analysis software to be able to deal in a consistent fashion with the different texts and metadata.

The first stage in the development of the virtual corpus system at the OTA was a selection of the texts which were to be used. Selection criteria were established with the aim of making it possible to construct a variety of corpora based on literary and non-literary prose in English from the 18th to the early 20th centuries. The texts selected were therefore all prose (excluding drama and poetry) in English, both fiction and non-fiction. A handful of texts which fall outside of these categories were included so that the text selection functionality could be tested, and so there are some earlier and some later texts, some poetry and drama, but not enough to make a representative corpus of these text types. Within the set of texts fitting the selection criteria, specific texts were chosen on the basis of textual integrity and markup. Less reliable texts were excluded. In terms of text encoding and markup, texts with no textual markup and with SGML (or XML) markup were selected. Linguistic and textual annotation encoded in non-standard ways was stripped out, in order to ensure that only the text was searched by the linguistic analysis software. SGML markup was left in because this can easily be identified automatically and differentiated from the text.

As texts were selected, some additional information was added to the database of metadata. Categories identifying the sex and nationality of authors, original language (if the text is a translation), text type and the date of original publication of the printed work on which the electronic edition is based. These categories were chosen on the basis that they would be useful for linguistic researchers to attempting to control variation or identify correspondences. These categories were not used as selection criteria for the texts included in the system at this stage.

The text selection and analysis software is implemented using CGI scripts written in Perl. These scripts access, manipulate and display metadata about the texts, and sections of the texts themselves. All back-end data is stored in a series of tables in a MySQL database. The Perl DBI module is used to pass queries from the script to the MySQL database, and to transmit the resulting tables back to the PERL script for display. The corpora generated by users' metadata queries are stored in temporary tables within the database. This allows comprehensive indexes to be generated and delivers increased performance. A house-keeping script is run periodically to handle the retiring of these temporary tables.

In the following section, the functionality of the system is demonstrated by working through an example query.



Session 'DEMO'

✓	Select your texts: ?
PERIOD:	To
AUTHOR:	Tw ain
NATIONALITY:	British [✓ US] Irish German Canadian French Russian
TEXT TYPE:	[✓ Fiction] [✓ Non-Fiction] Poetry Drama
SEX:	Male Female

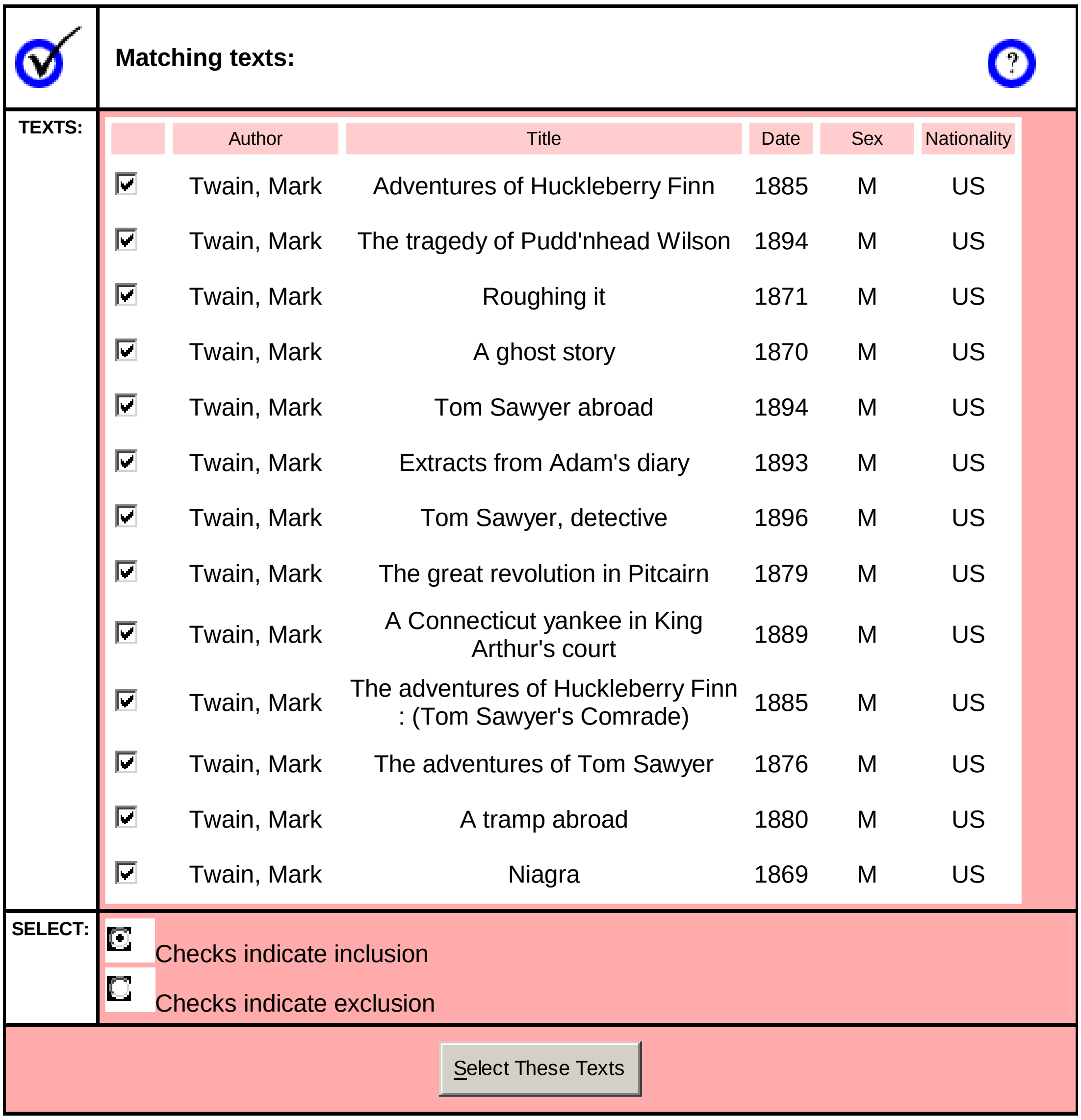
Find Texts ^

Figure 1: the text selection form

5 Extracting information from virtual corpora

In this section, a small piece of lexical research is carried out. It has been observed that Mark Twain frequently uses the word ‘presently’ in his prose. A comparison of his writings with the Brown Corpus (a one million word representative corpus of written American English from the 1960s) using the *keywords* function of Wordsmith tools shows that the word ‘presently’ shows strongly as a keyword in Twain’s writing. It occurs 438 times in 1.3 million words of Twain’s writing and just 35 times in 1 million words in the corpus. However, the mismatch between the type of language in the texts under investigation and the reference corpus in this case means that these results may merely reflect differences between fiction and non-fiction (since most of Brown is non-fiction), or change over time, since Twain’s work is from the late 19th and early 20th centuries and the Brown corpus is from the 1960s. It would be more useful to compare Twain’s writing with a corpus representative of his type of writing, from a similar time and place.

What is therefore desirable is a corpus of Mark Twain’s writing, plus a representative corpus of literary writing of his American contemporaries, and some tools for the simple lexical analysis of these corpora. The rest of this subsections details how this can be achieved with the OTA Virtual Corpus system.



Matching texts:

TEXTS:

	Author	Title	Date	Sex	Nationality
☑	Twain, Mark	Adventures of Huckleberry Finn	1885	M	US
☑	Twain, Mark	The tragedy of Pudd'nhead Wilson	1894	M	US
☑	Twain, Mark	Roughing it	1871	M	US
☑	Twain, Mark	A ghost story	1870	M	US
☑	Twain, Mark	Tom Sawyer abroad	1894	M	US
☑	Twain, Mark	Extracts from Adam's diary	1893	M	US
☑	Twain, Mark	Tom Sawyer, detective	1896	M	US
☑	Twain, Mark	The great revolution in Pitcairn	1879	M	US
☑	Twain, Mark	A Connecticut yankee in King Arthur's court	1889	M	US
☑	Twain, Mark	The adventures of Huckleberry Finn : (Tom Sawyer's Comrade)	1885	M	US
☑	Twain, Mark	The adventures of Tom Sawyer	1876	M	US
☑	Twain, Mark	A tramp abroad	1880	M	US
☑	Twain, Mark	Niagra	1869	M	US

SELECT:

● Checks indicate inclusion

○ Checks indicate exclusion

Figure 2: the text listing generated from the ‘Twain’ query

5.1 Making a Mark Twain corpus

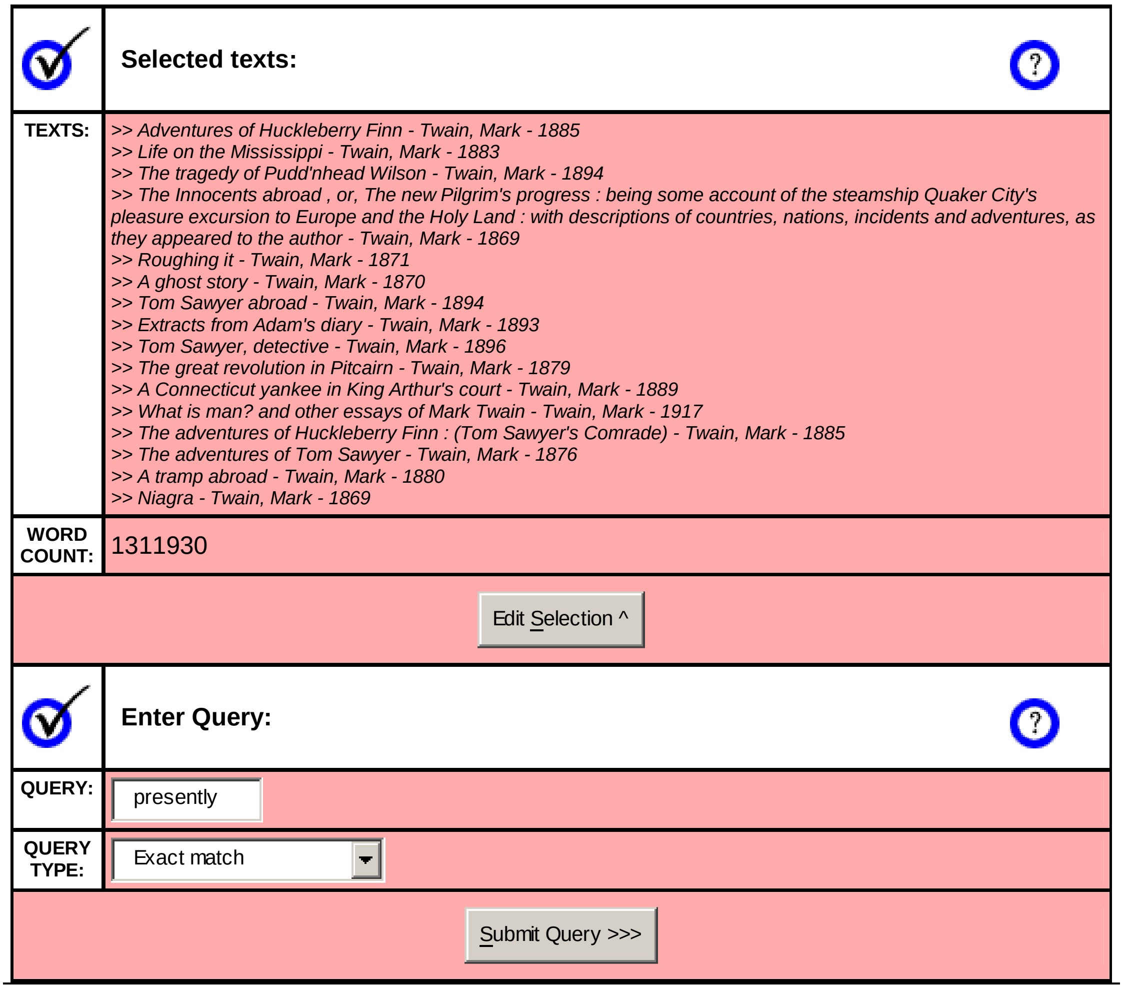
Figure 1 shows the input form for selecting texts based on metadata categories. In this simple form, the user has a limited set of categories in which values may be selected or input. In a later stage of the development of the software we will also make available an ‘advanced’ form through which all metadata categories may be queried.

In this example, a corpus of writings of Mark Twain is created by specifying the author’s surname, which should be sufficient for an accurate search in this case.

Clicking on ‘Find texts’ will load the page shown in Figure 2. This shows a listing of the texts which match the values entered in the text selection form. The user now has an opportunity to view and edit the list, or perhaps go back and change or refine the selection. At this stage the user may find that there are not sufficient texts matching the specified criteria to make a useful corpus.

There are also multiple copies of some texts in the archive, so the user needs to be aware of this and to unselect repeated entries.

Clicking on ‘Select these texts’ causes the user-defined corpus to be constructed from the selected texts. This involves building a database of all the words in the corpus, and so takes some time. The resulting corpus will however remain available to the user for the length of the current session, and so the user will not need to keep rebuilding.

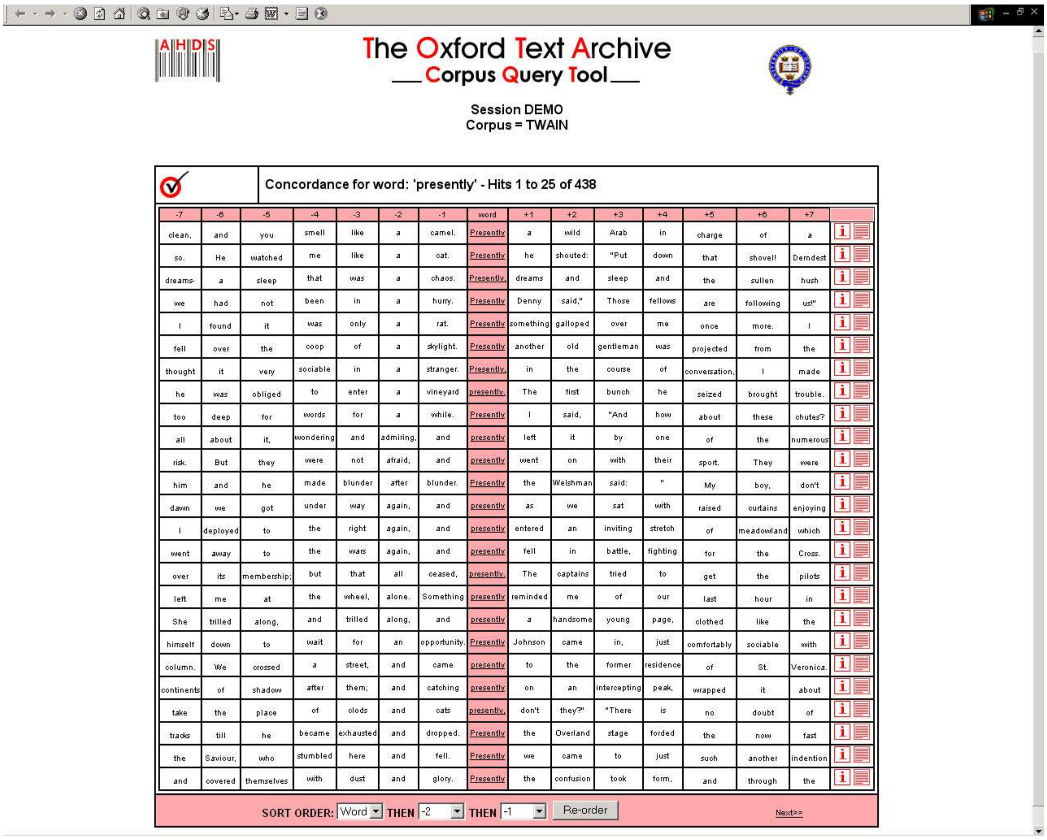


✓	**Selected texts:** ?
TEXTS:	*>> Adventures of Huckleberry Finn - Twain, Mark - 1885* *>> Life on the Mississippi - Twain, Mark - 1883* *>> The tragedy of Pudd'nhead Wilson - Twain, Mark - 1894* *>> The Innocents abroad , or, The new Pilgrim's progress : being some account of the steamship Quaker City's pleasure excursion to Europe and the Holy Land : with descriptions of countries, nations, incidents and adventures, as they appeared to the author - Twain, Mark - 1869* *>> Roughing it - Twain, Mark - 1871* *>> A ghost story - Twain, Mark - 1870* *>> Tom Sawyer abroad - Twain, Mark - 1894* *>> Extracts from Adam's diary - Twain, Mark - 1893* *>> Tom Sawyer, detective - Twain, Mark - 1896* *>> The great revolution in Pitcairn - Twain, Mark - 1879* *>> A Connecticut yankee in King Arthur's court - Twain, Mark - 1889* *>> What is man? and other essays of Mark Twain - Twain, Mark - 1917* *>> The adventures of Huckleberry Finn : (Tom Sawyer's Comrade) - Twain, Mark - 1885* *>> The adventures of Tom Sawyer - Twain, Mark - 1876* *>> A tramp abroad - Twain, Mark - 1880* *>> Niagra - Twain, Mark - 1869*
WORD COUNT:	1311930
	Edit Selection ^
✓	**Enter Query:** ?
QUERY:	presently
QUERY TYPE:	Exact match
	Submit Query >>>

Figure 3: entering the search query

A word count of the new corpus is given, and the user has an option to go back and edit the text selection again. The user can choose whether to download the corpus of selected texts or to submit online queries to the corpus[1]. Downloading the corpus enables the user to carry out their analysis of the corpus unrestrained by the limited functionality of the OTA online service. This is considered an essential aspect of the functionality, as the authors do not wish to try to prescribe or predict all of the functions and options users might wish to have available for the analysis of the corpus. It is expected that some users will have their own tools which they want to run locally to process the data. They may also have their own text collections which they wish to integrate or compare with their virtual corpus.

If the user does not opt to download the corpus and do the analysis locally, they can now enter the search query string in the form shown in Figure 3. The query type may also be selected, allowing the use of wildcards and substrings. In this case, we search for the word ‘presently’. Clicking on ‘Submit query’ loads the concordance display shown in Figure 4[2].



The Oxford Text Archive
Corpus Query Tool

Session DEMO
Corpus = TWAIN

Concordance for word: 'presently' - Hits 1 to 25 of 438

-7	-6	-5	-4	-3	-2	-1	word	+1	+2	+3	+4	+5	+6	+7	
clean,	and	you	smell	like	a	camel.	Presently	a	wild	Arab	in	charge	of	a	ℹ ▤
so.	He	watched	me	like	a	cat.	Presently	he	shouted:	"Put	down	that	shovel!	Derndest	ℹ ▤
dreams-	a	sleep	that	was	a	chaos.	Presently,	dreams	and	sleep	and	the	sullen	hush	ℹ ▤
we	had	not	been	in	a	hurry.	Presently	Denny	said,"	Those	fellows	are	following	us!"	ℹ ▤
I	found	it	was	only	a	rat.	Presently	something	galloped	over	me	once	more.	I	ℹ ▤
fell	over	the	coop	of	a	skylight.	Presently	another	old	gentleman	was	projected	from	the	ℹ ▤
thought	it	very	sociable	in	a	stranger.	Presently,	in	the	course	of	conversation,	I	made	ℹ ▤
he	was	obliged	to	enter	a	vineyard	presently.	The	first	bunch	he	seized	brought	trouble.	ℹ ▤
too	deep	for	words	for	a	while.	Presently	I	said,	"And	how	about	these	chutes?	ℹ ▤
all	about	it,	wondering	and	admiring,	and	presently	left	it	by	one	of	the	numerous	ℹ ▤
risk.	But	they	were	not	afraid,	and	presently	went	on	with	their	sport.	They	were	ℹ ▤
him	and	he	made	blunder	after	blunder.	Presently	the	Welshman	said:	"	My	boy,	don't	ℹ ▤
dawn	we	got	under	way	again,	and	presently	as	we	sat	with	raised	curtains	enjoying	ℹ ▤
I	deployed	to	the	right	again,	and	presently	entered	an	inviting	stretch	of	meadowland	which	ℹ ▤
went	away	to	the	wars	again,	and	presently	fell	in	battle,	fighting	for	the	Cross.	ℹ ▤
over	its	membership;	but	that	all	ceased,	presently.	The	captains	tried	to	get	the	pilots	ℹ ▤
left	me	at	the	wheel,	alone.	Something	presently	reminded	me	of	our	last	hour	in	ℹ ▤
She	trilled	along,	and	trilled	along,	and	presently	a	handsome	young	page,	clothed	like	the	ℹ ▤
himself	down	to	wait	for	an	opportunity.	Presently	Johnson	came	in,	just	comfortably	sociable	with	ℹ ▤
column.	We	crossed	a	street,	and	came	presently	to	the	former	residence	of	St.	Veronica.	ℹ ▤
continents	of	shadow	after	them;	and	catching	presently	on	an	intercepting	peak,	wrapped	it	about	ℹ ▤
take	the	place	of	clods	and	cats	presently,	don't	they?"	"There	is	no	doubt	of	ℹ ▤
tracks	till	he	became	exhausted	and	dropped.	Presently	the	Overland	stage	forded	the	now	fast	ℹ ▤
the	Saviour,	who	stumbled	here	and	fell.	Presently	we	came	to	just	such	another	indention	ℹ ▤
and	covered	themselves	with	dust	and	glory.	Presently	the	confusion	took	form,	and	through	the	ℹ ▤

SORT ORDER: Word THEN -2 THEN -1 Re-order Next>>

Figure 4: concordance display for ‘presently’ in the Twain corpus

There are options for sorting the lines, finding the source text for a given line, and expanding the context. Clicking on the key word will give a list of collocates in the corpus. In this case the most frequent lexical collocates appear to be words involved in advancing the narrative (‘began’), as may be expected from the meaning of ‘presently’, and in particular to do with movement (‘came’, ‘away’) and reported speech (‘says’). The way in which these words are used in co-occurrence with the keyword can easily be checked, as the entries in the collocation list are links to concordance lines where the two words co-occur. It is important to go back to the concordance lines in this way and look at the meanings of particular collocations if the interpretation of the collocations listings is not to be mere speculation.

In this case, there are 438 occurrences of ‘presently, out of a total corpus of 1311930 words. This represents a relative frequency of 0.03%.

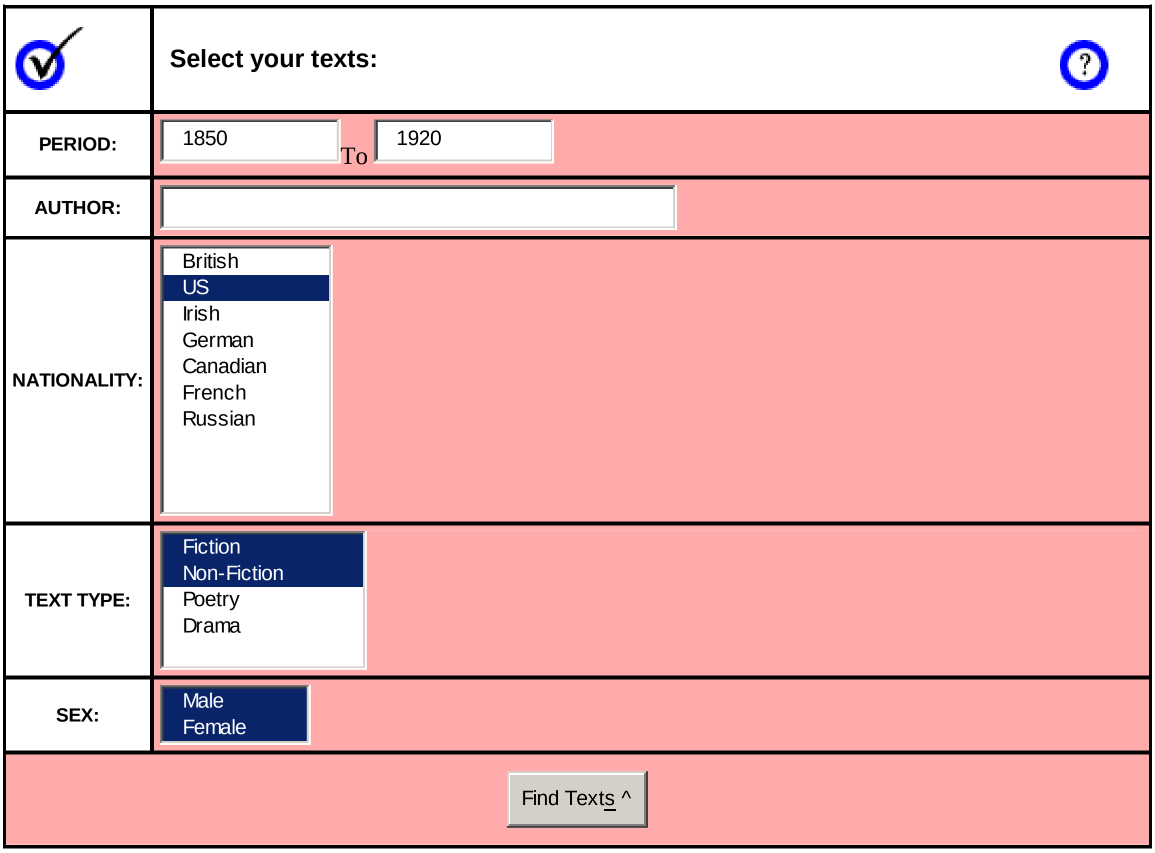
The next section will describe how a reference corpus can be built and will compare the use of the word ‘presently’ in this corpus.

[1] The option to download the corpus is not implemented in the version shown in Figure 3.

[2] A screen shot has been used for figure 4 to show the concordance output, because the HTML table produced is not the correct shape for this publication. The text is difficult to read the text in this picture, but it is intended only to demonstrate that a KWIC concordance is generated.

5.2 Making a reference corpus

It is first necessary to specify the selection criteria for the reference corpus. In this case a decision was made to select text from the time period 1850 to 1920 (roughly Twain's adult life), by US authors, prose and both fiction and non-fiction. Clicking on 'Find texts' therefore results in a list of the texts in the archive which are deemed to be of a similar type to Mark Twain's writings. The results are shown in Figure 6.



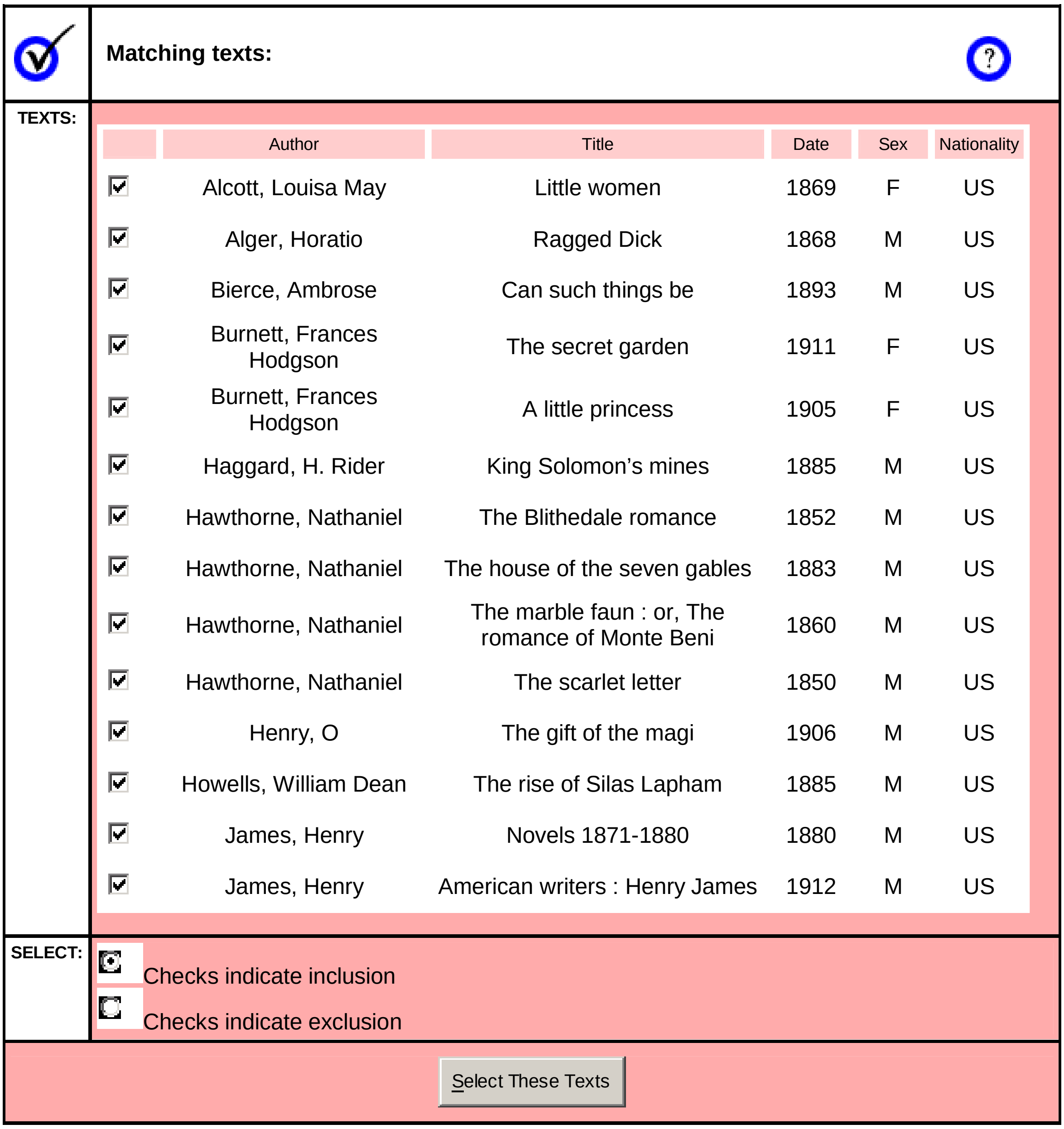
✓	**Select your texts:** ?
PERIOD:	1850 To 1920
AUTHOR:	
NATIONALITY:	British [✓ US] Irish German Canadian French Russian
TEXT TYPE:	[✓ Fiction] [✓ Non-Fiction] Poetry Drama
SEX:	[✓ Male] [✓ Female]

Figure 5: metadata selection for the reference corpus

The listing in figure 6 only shows the first screenful of matching texts. (The full, final selection can be seen in Figure 7.). In this case the user needs to scroll down to see all of the texts. The first thing which it is necessary to do is to unselect the works of Mark Twain, which will also be in the list as they match the specified criteria.

At this stage the user needs to critically analyse the corpus composition. The value of the results of any comparisons will depend on quality of the sample. Texts which exist in more than edition need to be identified if duplication is to be avoided. The user must also use this list of texts in order to identify the biases in the data. There may be too many works by a particular author or of a particular type which make the corpus unrepresentative. It may also be desirable to restrict the size of the corpus to a particular limit, in order to speed up the process, or produce a manageable number of hits, or to have a reference corpus of a similar size to the test corpus for particular statistical comparisons.

Once the user is happy with the selection, clicking on 'Select these texts' will take the user to the online query input stage (Figure 7). The full list of texts selected from the archive can be seen in Figure 7, along with the total number of words in the reference corpus, a little over 2 million.



Matching texts:

TEXTS:

	Author	Title	Date	Sex	Nationality
☑	Alcott, Louisa May	Little women	1869	F	US
☑	Alger, Horatio	Ragged Dick	1868	M	US
☑	Bierce, Ambrose	Can such things be	1893	M	US
☑	Burnett, Frances Hodgson	The secret garden	1911	F	US
☑	Burnett, Frances Hodgson	A little princess	1905	F	US
☑	Haggard, H. Rider	King Solomon’s mines	1885	M	US
☑	Hawthorne, Nathaniel	The Blithedale romance	1852	M	US
☑	Hawthorne, Nathaniel	The house of the seven gables	1883	M	US
☑	Hawthorne, Nathaniel	The marble faun : or, The romance of Monte Beni	1860	M	US
☑	Hawthorne, Nathaniel	The scarlet letter	1850	M	US
☑	Henry, O	The gift of the magi	1906	M	US
☑	Howells, William Dean	The rise of Silas Lapham	1885	M	US
☑	James, Henry	Novels 1871-1880	1880	M	US
☑	James, Henry	American writers : Henry James	1912	M	US

SELECT:

● Checks indicate inclusion

○ Checks indicate exclusion

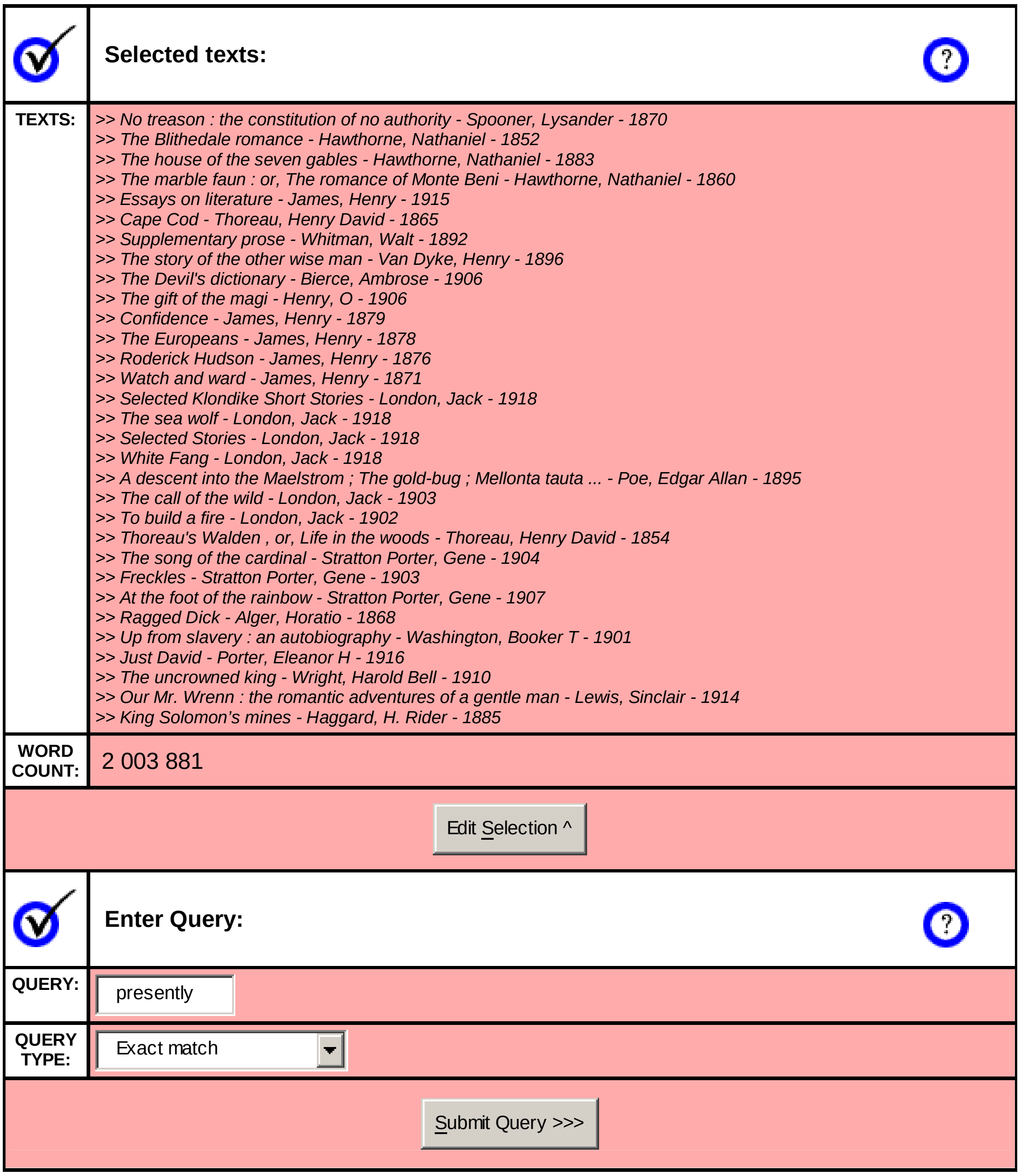
Figure 6: list of texts matching the specified metadata (abbreviated)

At this stage the search term can be entered, so the user now simply types ‘presently’ and clicks on ‘Submit query’, as was done with the Twain corpus. The concordance display is now loaded, and shown in Figure 8.

In this case, there are 201 hits for ‘presently’, which represents a relative frequency of 0.01% (201/2003881). This can be compared with the relative frequency in the Twain corpus of 0.03%, which can be seen to be higher. This would appear to support the hypothesis that Twain uses the word more than the norm, and could be a useful starting point for further investigation.

The usefulness of any reference or representative corpus that can be created using the virtual corpus system in this way will depend on what is available in the archive. The archive has necessarily been built up over the years depending on what texts were offered for deposit, and so no guarantees of coverage for any given text type can be offered. However, it may now be considered a priority for the

OTA to adopt a collections development policy which will aim to fill gaps and expand the coverage so that representative corpora can more readily be constructed.



Selected texts:

TEXTS:

>> No treason : the constitution of no authority - Spooner, Lysander - 1870
>> The Blithedale romance - Hawthorne, Nathaniel - 1852
>> The house of the seven gables - Hawthorne, Nathaniel - 1883
>> The marble faun : or, The romance of Monte Beni - Hawthorne, Nathaniel - 1860
>> Essays on literature - James, Henry - 1915
>> Cape Cod - Thoreau, Henry David - 1865
>> Supplementary prose - Whitman, Walt - 1892
>> The story of the other wise man - Van Dyke, Henry - 1896
>> The Devil's dictionary - Bierce, Ambrose - 1906
>> The gift of the magi - Henry, O - 1906
>> Confidence - James, Henry - 1879
>> The Europeans - James, Henry - 1878
>> Roderick Hudson - James, Henry - 1876
>> Watch and ward - James, Henry - 1871
>> Selected Klondike Short Stories - London, Jack - 1918
>> The sea wolf - London, Jack - 1918
>> Selected Stories - London, Jack - 1918
>> White Fang - London, Jack - 1918
>> A descent into the Maelstrom ; The gold-bug ; Mellonta tauta ... - Poe, Edgar Allan - 1895
>> The call of the wild - London, Jack - 1903
>> To build a fire - London, Jack - 1902
>> Thoreau's Walden , or, Life in the woods - Thoreau, Henry David - 1854
>> The song of the cardinal - Stratton Porter, Gene - 1904
>> Freckles - Stratton Porter, Gene - 1903
>> At the foot of the rainbow - Stratton Porter, Gene - 1907
>> Ragged Dick - Alger, Horatio - 1868
>> Up from slavery : an autobiography - Washington, Booker T - 1901
>> Just David - Porter, Eleanor H - 1916
>> The uncrowned king - Wright, Harold Bell - 1910
>> Our Mr. Wrenn : the romantic adventures of a gentle man - Lewis, Sinclair - 1914
>> King Solomon's mines - Haggard, H. Rider - 1885

WORD COUNT: 2 003 881

Enter Query:

QUERY: presently

QUERY TYPE: Exact match

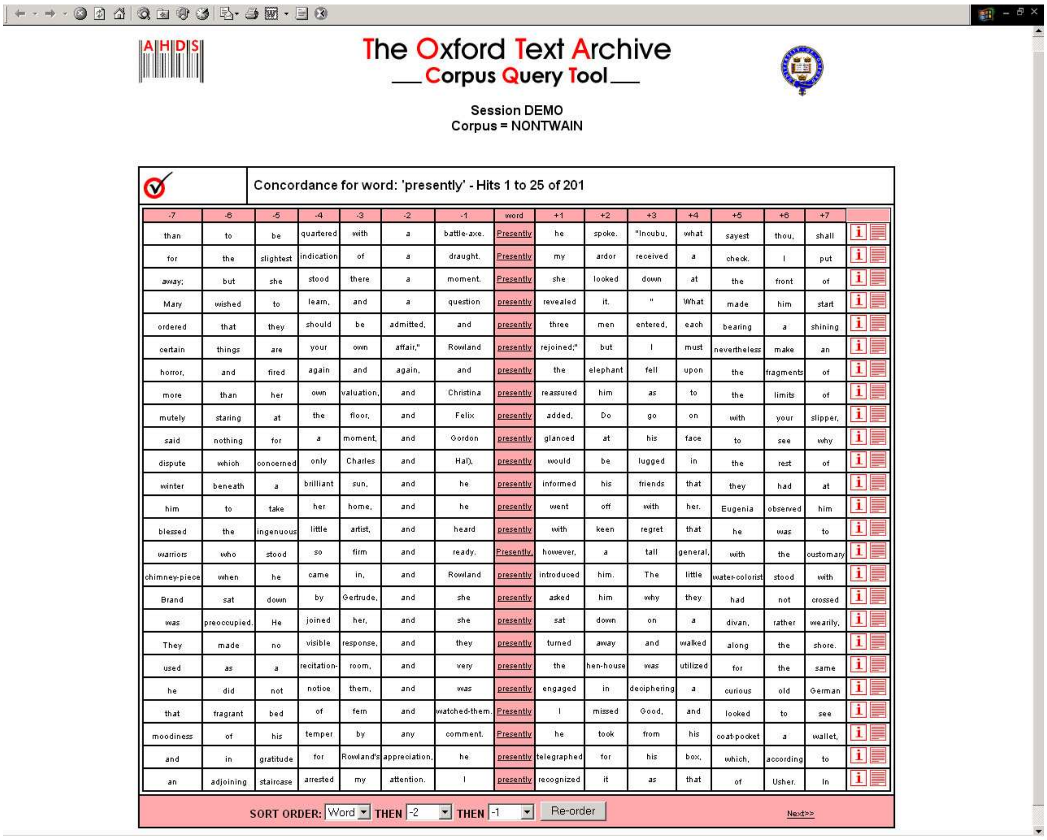
Figure 7: Query entry for the reference corpus

6 Future developments

At the time of writing the user management and public access routines for the virtual corpus system have not yet been implemented, and so cannot be described here. Free public access will be granted to all holdings in the archive which do not have intellectual property rights (IPR) restrictions. However, the load imposed on the server by indexing of corpora will mean that the corpus construction process will not be instantaneous.

There are clear limitations imposed limited number of texts in the corpus, the bias in favour of classic literary texts and the limited functionality of the software. The OTA Virtual Corpus system is not intended to be a replacement for existing or future corpora. Nor is it intended that the online query tools will be elaborated and expanded to a great degree, but they should only facilitate simple investigations for teachers and learners, or serve as the initial basis for hypothesis forming and testing in research.

Online access to corpora through the same interface will be offered. There will also be the possibility to use text level metadata of the contents of corpora in order to pick and choose which texts to use and include in a virtual corpus, as described in (Wittenberg and Broeder 2002).



Session DEMO
Corpus = NONTWAIN

Concordance for word: 'presently' - Hits 1 to 25 of 201

-7	-6	-5	-4	-3	-2	-1	word	+1	+2	+3	+4	+5	+6	+7
than	to	be	quartered	with	a	battle-axe.	Presently	he	spoke.	"Incubu,	what	sayest	thou,	shall
for	the	slightest	indication	of	a	draught.	Presently	my	ardor	received	a	check.	I	put
away;	but	she	stood	there	a	moment.	Presently	she	looked	down	at	the	front	of
Mary	wished	to	learn,	and	a	question	presently	revealed	it.	"	What	made	him	start
ordered	that	they	should	be	admitted,	and	presently	three	men	entered,	each	bearing	a	shining
certain	things	are	your	own	affair,"	Rowland	presently	rejoined;"	but	I	must	nevertheless	make	an
horror,	and	fired	again	and	again,	and	presently	the	elephant	fell	upon	the	fragments	of
more	than	her	own	valuation,	and	Christina	presently	reassured	him	as	to	the	limits	of
mutely	staring	at	the	floor,	and	Felix	presently	added,	Do	go	on	with	your	slipper,
said	nothing	for	a	moment,	and	Gordon	presently	glanced	at	his	face	to	see	why
dispute	which	concerned	only	Charles	and	Hal),	presently	would	be	lugged	in	the	rest	of
winter	beneath	a	brilliant	sun,	and	he	presently	informed	his	friends	that	they	had	at
him	to	take	her	home,	and	he	presently	went	off	with	her.	Eugenia	observed	him
blessed	the	ingenuous	little	artist,	and	heard	presently	with	keen	regret	that	he	was	to
warriors	who	stood	so	firm	and	ready.	Presently,	however,	a	tall	general,	with	the	customary
chimney-piece	when	he	came	in,	and	Rowland	presently	introduced	him.	The	little	water-colorist	stood	with
Brand	sat	down	by	Gertrude,	and	she	presently	asked	him	why	they	had	not	crossed
was	preoccupied.	He	joined	her,	and	she	presently	sat	down	on	a	divan,	rather	wearily,
They	made	no	visible	response,	and	they	presently	turned	away	and	walked	along	the	shore.
used	as	a	recitation-	room,	and	very	presently	the	hen-house	was	utilized	for	the	same
he	did	not	notice	them,	and	was	presently	engaged	in	deciphering	a	curious	old	German
that	fragrant	bed	of	fern	and	watched-them.	Presently	I	missed	Good,	and	looked	to	see
moodiness	of	his	temper	by	any	comment.	Presently	he	took	from	his	coat-pocket	a	wallet,
and	in	gratitude	for	Rowland's	appreciation,	he	presently	telegraphed	for	his	box,	which,	according	to
an	adjoining	staircase	arrested	my	attention.	I	presently	recognized	it	as	that	of	Usher.	In

SORT ORDER: Word ▾ THEN -2 ▾ THEN -1 ▾ Re-order

Next>>

Figure 8: concordance of ‘presently’ in the reference corpus

7 Bibliography

Renouf, A. (2001). The web as a source of linguistic information. Corpus Linguistics 2001, Lancaster, UK, UCREL.

Wittenberg, P. and D. Broeder (2002). Management of Language Resources using Metadata. Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Gran Canaria, Spain, ELRA.

Wynne, M. (2002). The Language Resources Archive of the 21st Century. Language Resources and Evaluation Conference (LREC), Las Palmas, Gran Canaria, Spain.