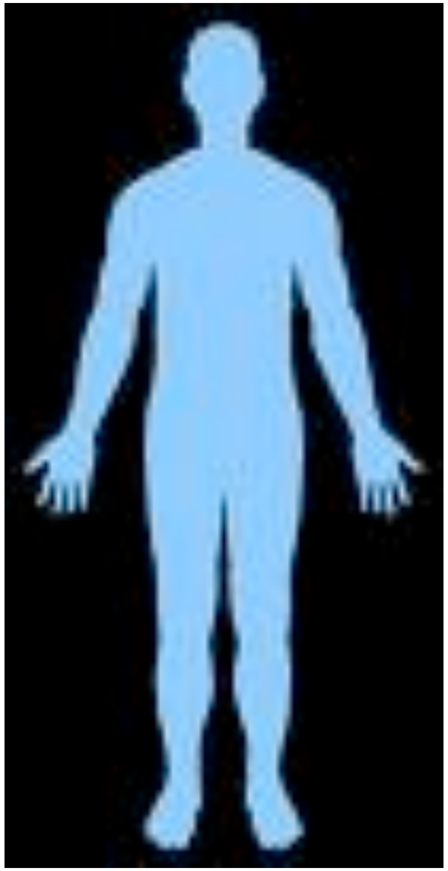
Spoken and written language corpora

Janne Bondi Johannessen
(UiO)

PhD training course: Infrastructural tools for the study of linguistic variation
Fefor Høifjellshotell, Gudbrandsdalen,
Norway, 2.-6. June, 2009

Corpus



- The main part of a bodily structure or organ.
- A large collection of writings of a specific kind or on a specific subject.
- A collection of writings or recorded remarks used for linguistic analysis.
  - (http://www.thefreedictionary.com)

What is a corpus?

- Text
- Size
- Format
- Representation
- Use
- Language
- Medium
- Annotation
- Technology

Corpus or web?

- We will discuss this later.

Corpus history: text studies

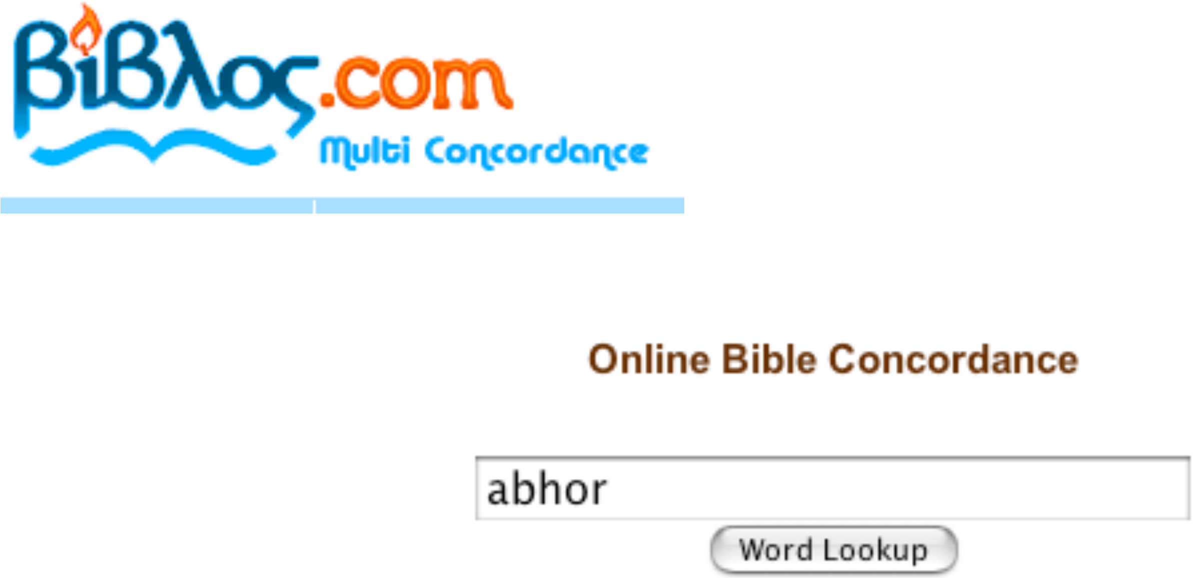
- Verbal concordances of the Bible are the invention of the Dominican friars. The text which served as basis of their work was naturally that of the Vulgate, the Bible of the Middle Ages.

Corpus history: text studies

- The first concordance, completed in 1230, was undertaken under the guidance of Hugo de Sancto Charo, afterwards a cardinal, assisted, it is said, by 500 fellow-Dominicans. It contained no quotations, and was purely an index to passages where a word was found.
- The first concordance to be printed appeared in 1470

(CATHOLIC ENCYCLOPEDIA: Concordances of the Bible)

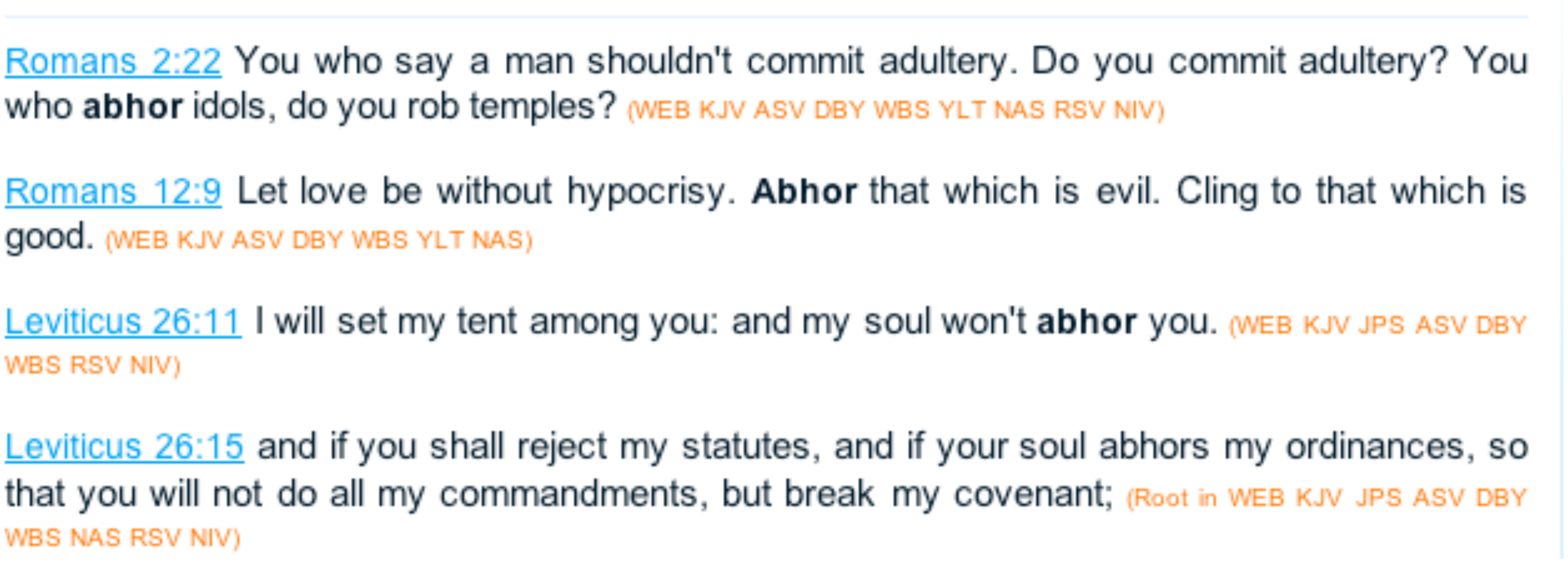
Bible concordance:
http://bibletab.com/



Online Bible Concordance

abhor

Bible result



Romans 2:22 You who say a man shouldn't commit adultery. Do you commit adultery? You who **abhor** idols, do you rob temples? (WEB KJV ASV DBY WBS YLT NAS RSV NIV)

Romans 12:9 Let love be without hypocrisy. **Abhor** that which is evil. Cling to that which is good. (WEB KJV ASV DBY WBS YLT NAS)

Leviticus 26:11 I will set my tent among you: and my soul won't **abhor** you. (WEB KJV JPS ASV DBY WBS RSV NIV)

Leviticus 26:15 and if you shall reject my statutes, and if your soul abhors my ordinances, so that you will not do all my commandments, but break my covenant; (Root in WEB KJV JPS ASV DBY WBS NAS RSV NIV)

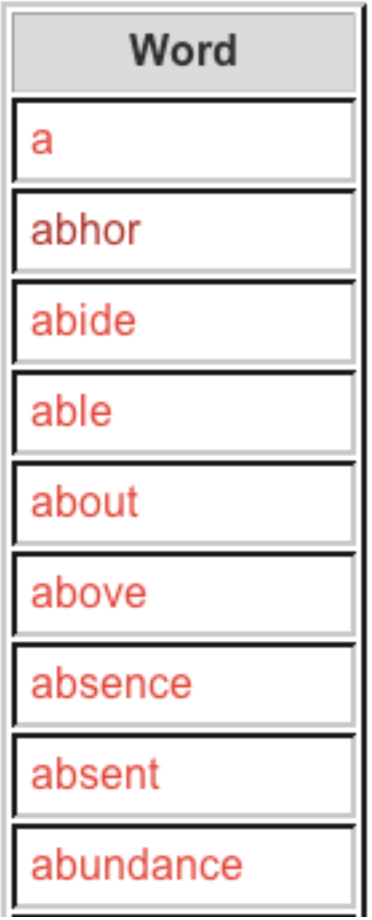
Corpus history: Shakespeare Concordance

- Samuel Ayscough (1745-1804), Librarian and Index-maker, known as 'The Prince of Indexers'.
- Ayscough is also remembered as the writer of the first concordance to Shakespeare. Entitled *An Index to the Remarkable Passages and Words Made Use of by Shakespeare; Calculated to Point out the Different Meanings to Which the Words are Applied,* it was published by John Stockdale in 1790. (*Wikipedia*)

Corpus history: Shakespeare Concordance

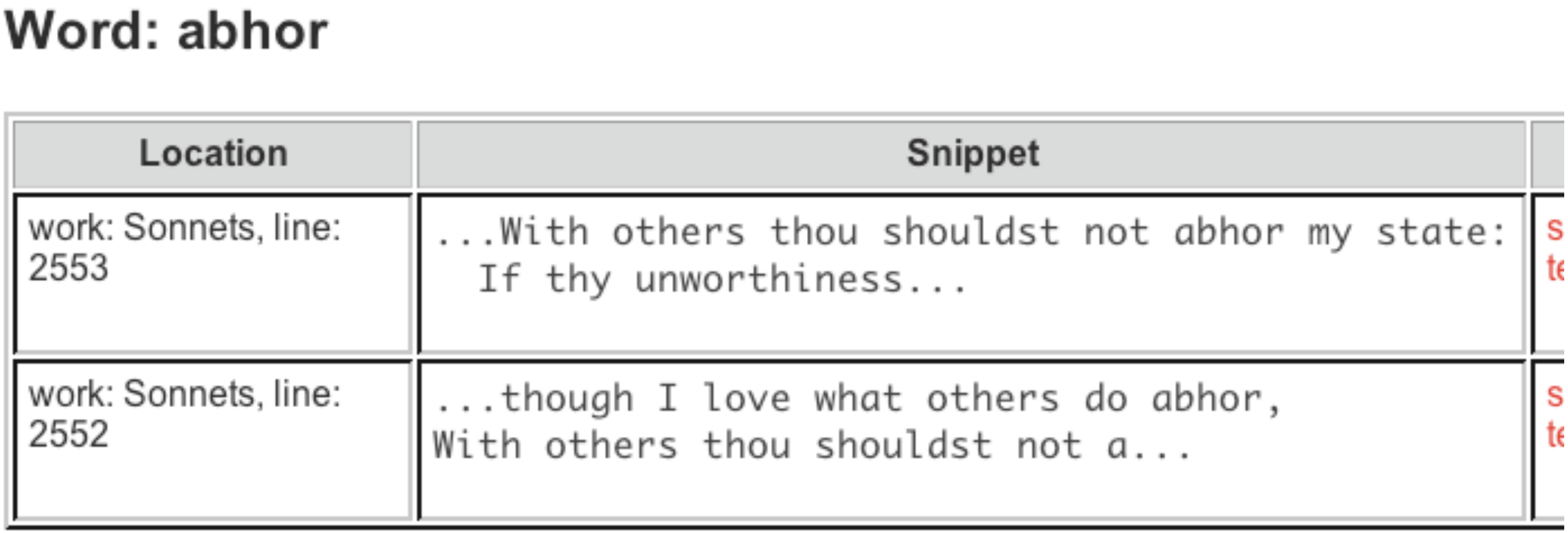
- In the year after her marriage, Mary Cowden Clarke began her valuable Shakespeare concordance, which was eventually issued in eighteen monthly parts (1844-1845), and in volume form in 1845 as *The Complete Concordance to Shakespeare, being a Verbal Index to all the Passages in the Dramatic Works of the Poet.*
- This work superseded the Copious Index to ... Shakespeare (1790) of Samuel Ayscough, and the Complete Verbal Index ... (1805-1807) of Francis Twiss. (Wikipedia)

Shakespeare Concordance: http://demo.openshakespeare.org/concordance/



Word
a
abhor
abide
able
about
above
absence
absent
abundance

Shakespeare result



Word: abhor

Location	Snippet	
work: Sonnets, line: 2553	...With others thou shouldst not abhor my state: If thy unworthiness...	s te
work: Sonnets, line: 2552	...though I love what others do abhor, With others thou shouldst not a...	s te

Corpus history: lexicography

- Published on 15 April 1755 and written by Samuel Johnson, ***A Dictionary of the English Language***
  - *An archive of 150 000 illustrative citations on slips of paper for the 40 000 head words.*
- Oxford English Dictionary
  - 71 years of building a corpus of the literary canon of English from 1000 AD to 1928 (the 12th and final volume)
  - The 2nd edition of 1984: 447 000 word forms, 2.4 mill. quotations

Corpus history: Dialect

- Lexical variation
- The English Dialect Dictionary (1898-1905)
- The Existing Phonology of English Dialects (1889)

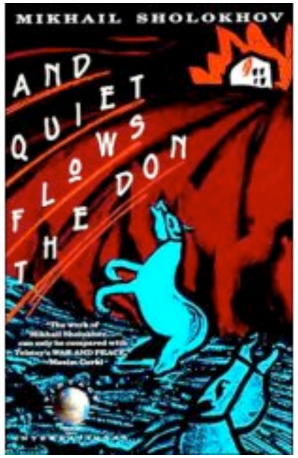
Corpus history: Language Education

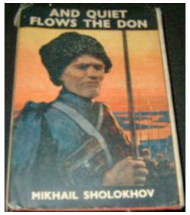
- Thorndike (1921) compiled a corpus of 4.5 mill. words from 41 different sources
- To make frequency lists
- To teach native Americans English

Corpus history: Grammar

- The very famous Danish linguist Otto Jespersen (1860-1943)
  - Used informal corpora for his linguistic descriptions
- Survey of English Usage (Randolph Quirk, 1959-) a pre-electronic systematic corpus to be used as a basis for grammar description (A Comprehensive Grammar of the English Language, Quirk et al 1985)

Types of corpus use: Literature


MIKHAIL SHOLOKHOV
AND QUIET FLOWS THE DON
"The work of Mikhail Sholokhov can only be compared with Tolstoy's WAR AND PEACE." —Maxim Gorki



Accusations of plagiarism: the Russian Nobel laureate Mikhail Aleksandrovich Sholokhov in 1974 was accused of having stolen large chunks of texts for his book *And Quiet Flows the Don* from the late author Fjodor Krjukov.

Types of corpus use: Literature

- An inter-Nordic research team was formed in 1975, captained by the late Geir Kjetsaa, a professor of Russian at the University of Oslo, for disentangling the **Don** mystery.
- Quantitative data were gathered and organised
  - relating to word lengths
  - frequencies of certain words and phrases
  - sentence lengths
  - Grammatical characteristics, etc.
  - These data were extracted from three corpora
    - » (Nils Lid Hjort)

Tamil concordance: Index of Kamparamayanam (Govindankutty 1973)

- 12500 stanzas
- Each stanza: 4 lines
- The complete index: 3500 typed pages

- Method: small cards, annotation for meaning and grammatical category
- Six years hard work for G. And his ”long-suffering” colleagues (Kennedy 1998)

The Brown Corpus
(Brown University Standard Corpus of Present-day American English)

- The first computer corpus compiled for linguistic research
- Synchronic American language printed in 1961
- Taken from a large number of text categories (informative and imaginative prose), reasonably ”representative”
- 1 mill words
- LOB Corpus
  - (London-Oslo-Bergen Corpus)
  - British counterpart of the Brown Corpus

Brown Corpus

- ·A. PRESS: REPORTAGE (44 texts)
- ·B. PRESS: EDITORIAL (27 texts)
- ·C. PRESS: REVIEWS (17 texts)
- ·D. RELIGION (17 texts)
- ·E. SKILL AND HOBBIES (36 texts)
- ·F. POPULAR LORE (48 texts)
- ·G. BELLES-LETTRES (75 texts)
- ·H. MISCELLANEOUS: GOVERNMENT & HOUSE ORGANS (30 texts)
- ·J. LEARNED (80 texts)
- ·K: FICTION: GENERAL (29 texts)
- ·L: FICTION: MYSTERY (24 texts)
- ·M: FICTION: SCIENCE (6 texts)
- ·N: FICTION: ADVENTURE (29 texts)
- ·P.FICTION: ROMANCE (29 texts)
- ·R. HUMOR (9 texts)

Modern Corpus History

- The Brown Corpus has also spawned a number of similarly structured corpora: the LOB Corpus (1960s British English), Kolhapur (Indian English), Wellington (New Zealand English), Australian Corpus of English (Australian English), and the FLOB Corpus (1990s British English). Other corpora: include the International Corpus of English, and the British National Corpus, a 100 million word collection of a range of spoken and written texts, created in the 1990s by a consortium of publishers, universities (Oxford and Lancaster) and the British Library. For contemporary American English, work has stalled on the American National Corpus, but the 360 million word BYU Corpus of American English (1990-present) is now available.

Corpora in Norway

- University of Oslo (The Text Laboratory and Lexicography)
- University of Bergen (Aksis)
- University of Tromsø (Sami)
- NTNU

Corpora developed at the Text Laboratory (UiO)

- Oslo-korpuset av taggede norske tekster, bokmål
- Oslo-korpuset av taggede norske tekster, nynorsk
- The French Newspaper Corpus
- KAL
- Usenet-korpuset
- Leksikografisk bokmålskorpus LBK
- (+ "Distribution and meaning of nominal forms in Norwegian - A cross-linguistic perspective", NTNU)

- Norske talespråkskorpus
- Nordic Dialect Corpus
- NoTa-Oslo
- Big Brother-korpuset
- Bosnisk-korpuset
- Sidaama-korpuset
- Oslo Multilingual Corpus
- OPUS
- LOGONs norsk-engelske turistkorpus
- The Sofie Treebank

Some corpora linked from the Text Laboratory

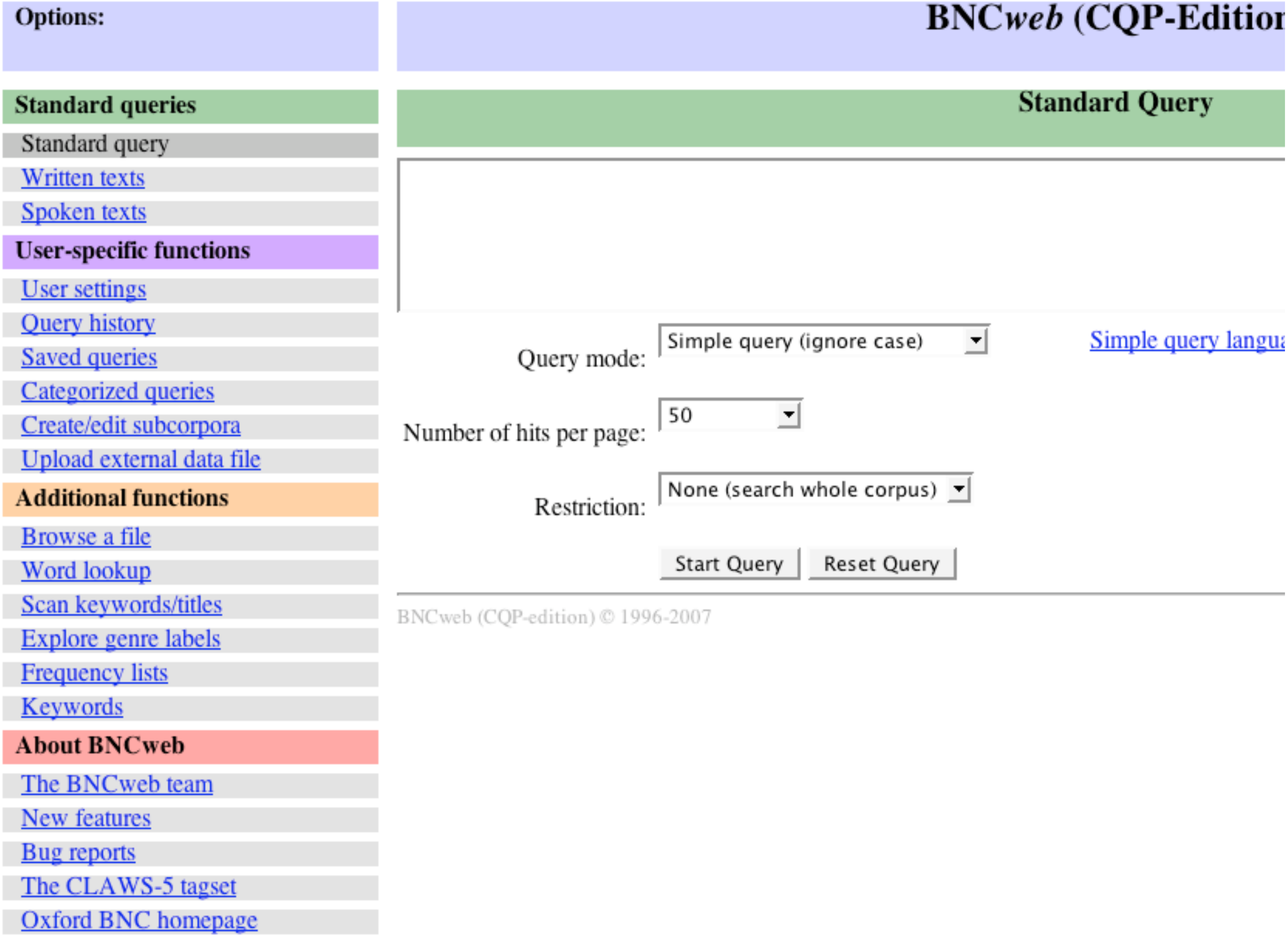
- Aftenposten,
- Atekst
- Bibelen (Bibelselskapet)Bibelen (Menighetsfakultetet)
- British National Corpus - Homepage British National Corpus - The Zurich BNC *web*
- Query SystemBySoc: Dansk Talesprog
- BYU Corpus of American English
- Cola-prosjektet
- Corpus Gesproken Nederlands
- The English-Norwegian Parallel Corpus (ENPC)
- Færoyskt TekstaSavn (FTS) Gothenburg Spoken Language Corpus
- Norsk aviskorpus (UiB: Aksis)
- (+ "Distribution and meaning of nominal formsin Norwegian - A cross-linguistic perspective". NTNU)
- Nynorskkorpuset til Norsk ordbok 2014 (UiO)
- Korpus 2000 - "dansk sprog omkring årtusindskiftet”
- Norske tekster (UiB: Aksis)
- Språkbankens korpus av svenske tekster (Göteborg Universitet, Språkbanken)
- Talesøk - Søkbare norske talemålsinnspillinger (UiB: Aksis)
- UNO - Språkkontakt og ungdomsspråk i Norden
- VISL-korpuset av danske tekster (VISL - Visual Interactive Syntax Learning)

Types of electronic corpora

- General corpora
  - Balanced
- Specialised corpora
  - Training +Test
  - Dialect + Regional
  - Parallel
- Types of text
  - Sample or Full
  - Dynamic (Monitor) or Static
- Time
  - Synchronic or diachronic
- Medium of corpus
  - Written or Spoken

Annotation of corpus

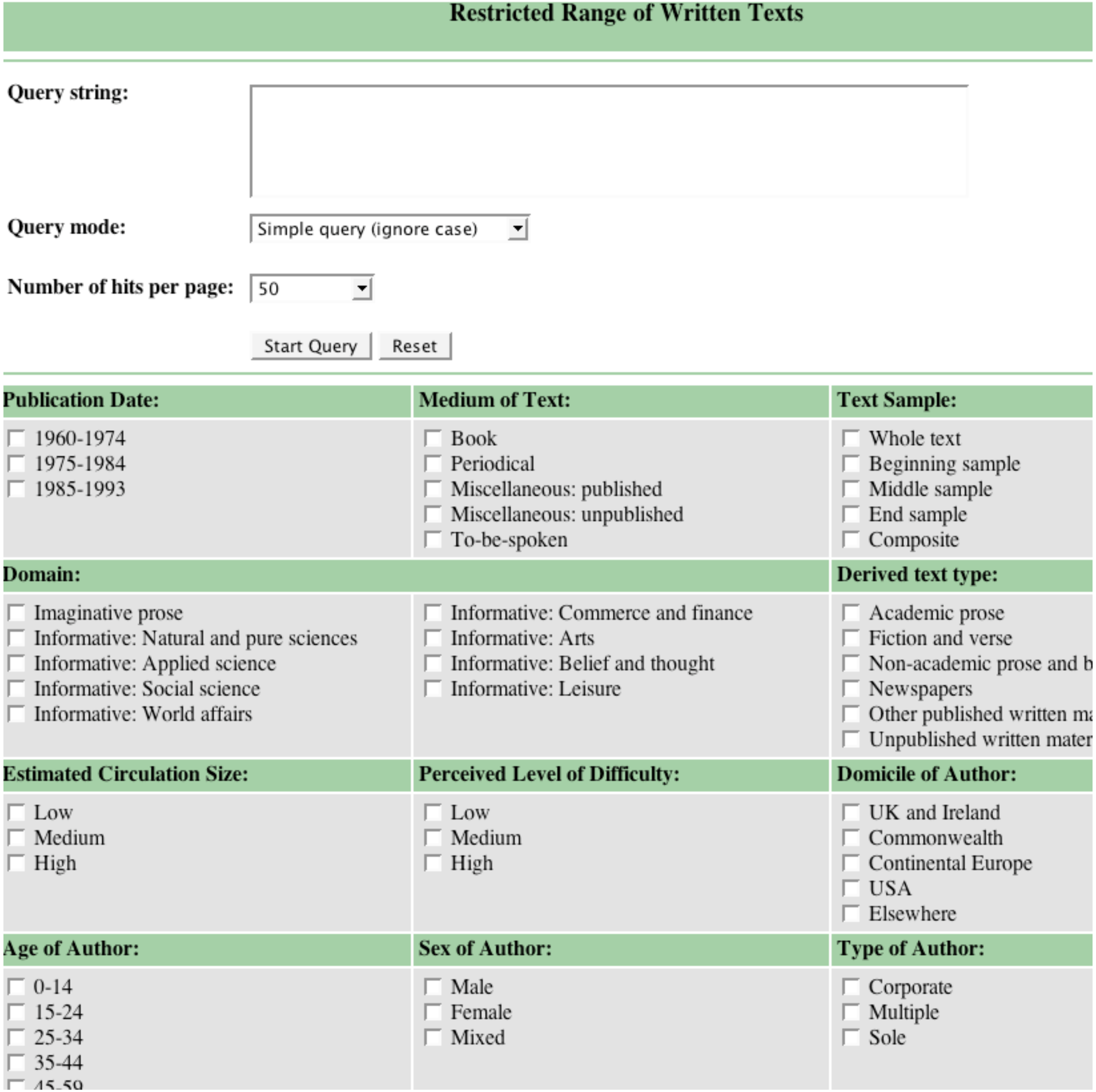
- POS tagging
- Parsing
- Alignment
- Semantic (sense) tagging
- Source data:
  - Author
  - Work
  - Date
  - Place



Query mode: Simple query (ignore case)

Number of hits per page: 50

Restriction: None (search whole corpus)



Restricted Range of Written Texts

Query string:

Query mode: Simple query (ignore case)

Number of hits per page: 50

Publication Date:

- □ 1960-1974
- □ 1975-1984
- □ 1985-1993

Medium of Text:

- □ Book
- □ Periodical
- □ Miscellaneous: published
- □ Miscellaneous: unpublished
- □ To-be-spoken

Text Sample:

- □ Whole text
- □ Beginning sample
- □ Middle sample
- □ End sample
- □ Composite

Domain:

- □ Imaginative prose
- □ Informative: Natural and pure sciences
- □ Informative: Applied science
- □ Informative: Social science
- □ Informative: World affairs
- □ Informative: Commerce and finance
- □ Informative: Arts
- □ Informative: Belief and thought
- □ Informative: Leisure

Derived text type:

- □ Academic prose
- □ Fiction and verse
- □ Non-academic prose and b
- □ Newspapers
- □ Other published written ma
- □ Unpublished written mater

Estimated Circulation Size:

- □ Low
- □ Medium
- □ High

Perceived Level of Difficulty:

- □ Low
- □ Medium
- □ High

Domicile of Author:

- □ UK and Ireland
- □ Commonwealth
- □ Continental Europe
- □ USA
- □ Elsewhere

Age of Author:

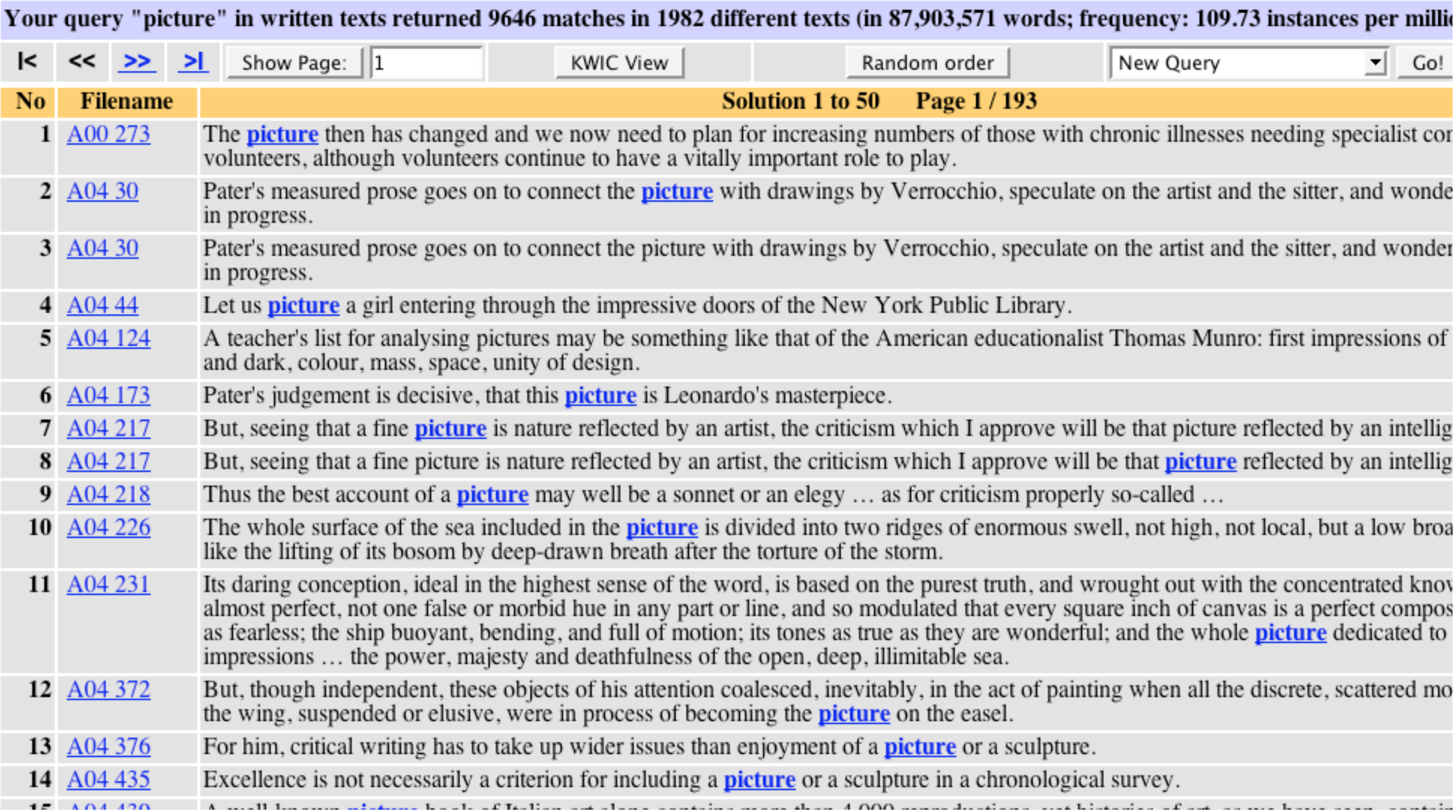
- □ 0-14
- □ 15-24
- □ 25-34
- □ 35-44
- □ 45-59

Sex of Author:

- □ Male
- □ Female
- □ Mixed

Type of Author:

- □ Corporate
- □ Multiple
- □ Sole



Your query "picture" in written texts returned 9646 matches in 1982 different texts (in 87,903,571 words; frequency: 109.73 instances per milli

No	Filename	Solution 1 to 50 Page 1 / 193
1	A00 273	The **picture** then has changed and we now need to plan for increasing numbers of those with chronic illnesses needing specialist co volunteers, although volunteers continue to have a vitally important role to play.
2	A04 30	Pater's measured prose goes on to connect the **picture** with drawings by Verrocchio, speculate on the artist and the sitter, and wonde in progress.
3	A04 30	Pater's measured prose goes on to connect the picture with drawings by Verrocchio, speculate on the artist and the sitter, and wonde in progress.
4	A04 44	Let us **picture** a girl entering through the impressive doors of the New York Public Library.
5	A04 124	A teacher's list for analysing pictures may be something like that of the American educationalist Thomas Munro: first impressions of and dark, colour, mass, space, unity of design.
6	A04 173	Pater's judgement is decisive, that this **picture** is Leonardo's masterpiece.
7	A04 217	But, seeing that a fine **picture** is nature reflected by an artist, the criticism which I approve will be that picture reflected by an intellig
8	A04 217	But, seeing that a fine picture is nature reflected by an artist, the criticism which I approve will be that **picture** reflected by an intellig
9	A04 218	Thus the best account of a **picture** may well be a sonnet or an elegy … as for criticism properly so-called …
10	A04 226	The whole surface of the sea included in the **picture** is divided into two ridges of enormous swell, not high, not local, but a low broa like the lifting of its bosom by deep-drawn breath after the torture of the storm.
11	A04 231	Its daring conception, ideal in the highest sense of the word, is based on the purest truth, and wrought out with the concentrated know almost perfect, not one false or morbid hue in any part or line, and so modulated that every square inch of canvas is a perfect compos as fearless; the ship buoyant, bending, and full of motion; its tones as true as they are wonderful; and the whole **picture** dedicated to impressions … the power, majesty and deathfulness of the open, deep, illimitable sea.
12	A04 372	But, though independent, these objects of his attention coalesced, inevitably, in the act of painting when all the discrete, scattered mo the wing, suspended or elusive, were in process of becoming the **picture** on the easel.
13	A04 376	For him, critical writing has to take up wider issues than enjoyment of a **picture** or a sculpture.
14	A04 435	Excellence is not necessarily a criterion for including a **picture** or a sculpture in a chronological survey.
15	A04 439	A well-known **picture** book of Italian art alone contains more than 4,000 reproductions, yet historian of art, as we have seen, contain