Class 2

A glimpse of corpus linguistics

Sources of data in linguistic analysis

- introspection
- individual citations
- corpus data

Definition of a corpus

"A corpus is a body of text assembled according to explicit design criteria for a specific purpose"

(Atkins and Clear 1992:5, after Granger 1998:7)

Purpose

- compilation of a dictionary
  - general English or ESP?
  - large and comprehensive or small and including most frequent words/terms?
  - etc.
- linguistic research
  - medium (spoken or written language or both?)
  - feature of language (frequent or infrequent)?
  - etc.

Design criteria

- size
- content
- balance and representativeness
- permanence

(Hunston 2002)

Types of corpora

- general vs. special language
- reference vs. monitor
- spoken vs. written
- whole texts vs. samples
- synchronic vs. diachronic
- monolingual vs. mulitilingual
- parallel vs. comparable

Historical Background

- Pre-electronic Corpora (biblical & literary studies, early dictionaries, etc.)
- 1st-generation Major Corpora (Brown, LOB, LLC, Kolhapur, Wellington, etc.)
- 2nd-generation Mega-corpora (Bank of English, British National Corpus, ICE-GB, American National Corpus, etc.)

(Kennedy 1998:13)

1-generation corpora

- The Brown Corpus
  http://khnt.aksis.uib.no/icame/manuals/brown/
- The LOB Corpus
  http://khnt.hit.uib.no/icame/manuals/lobman/
- The Survey of English Usage
  http://www.ucl.ac.uk/english-usage/about/index.htm
- The London-Lund Corpus (spoken)
  http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM

2-generation corpora

- The British National Corpus (BNC)
  http://www.natcorp.ox.ac.uk/
- Corpus of Contemporary American English (COCA)
  http://corpus.byu.edu/coca/
- International Corpus of English (ICE)
  http://www.ucl.ac.uk/english-usage/ice/index.htm
  http://www.ucl.ac.uk/english-usage/projects/ice-gb/index.htm

Archive / Library / Corpus

- **Archive**: a repository of readable electronic texts not linked in any coordinated way, e.g. the Oxford Text Archive
- **Electronic Text Library (or ETL, Fr. 'textothèque')**: a collection of electronic texts in standardized format with certain conventions relating to content, etc., but without rigorous selectional constraints.
- **Corpus**: a subset of an ETL, built according to explicit design criteria for a specific purpose, e.g. the Corpus Révolutionnaire (Bibliothèque Beaubourg, Paris), the Cobuild Corpus.

Atkins, S, Clear, J, & Ostler (1992) after David Lee

Text Archives

- The Oxford Text Archive
  http://ota.ahds.ac.uk/
- ICAME
  http://helmer.hit.uib.no/icame.html
- TRACTOR
  http://www.tractor.bham.ac.uk/
- Project Guttenberg
  http://www.promo.net/pg/index.html
- The Books Online
  http://digital.library.upenn.edu/books/

Useful Link

- David Lee's Bookmarks for Corpus-based Linguistics

  http://tiny.cc/corpora

Activity 1

- Using David Lee's Bookmarks for CBL find three corpora in a language other than English (preferably a language you know). Note down their names, URL addresses and short descriptions (size, content, availability)
- Using the webpages for Gutenberg Project and The Books Online check what books are available in this language. Choose one author and produce a list of resources by this author available on the Internet.

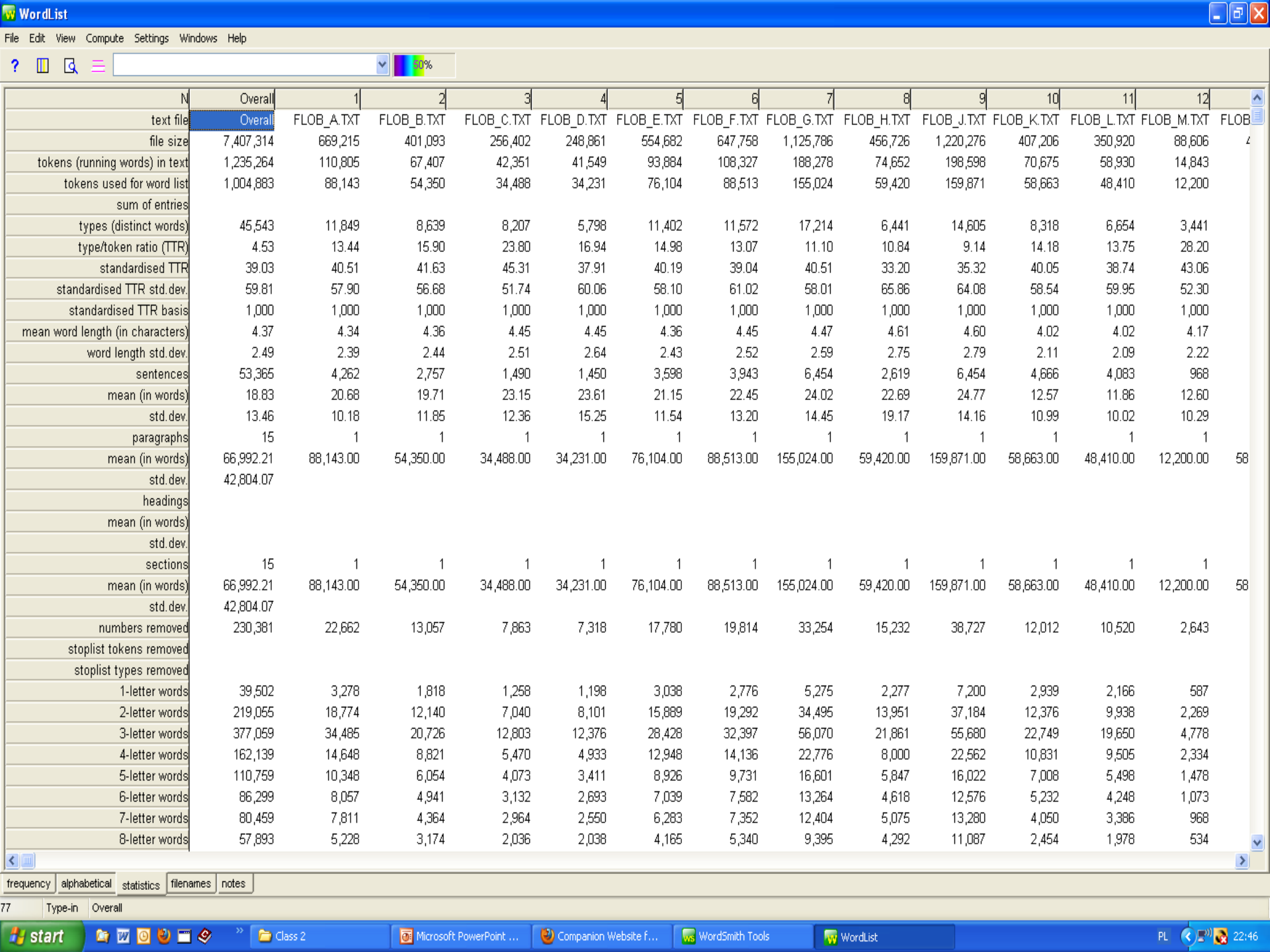
Consulting corpora

- A corpus by itself is useless unless you have a computer programme which allows you to analyze it.
- Programmes for analysing corpora are called concordancers.
- Some concordancers are purpose-built and work only with one particular corpus, and some others are designed to analyze any corpus.

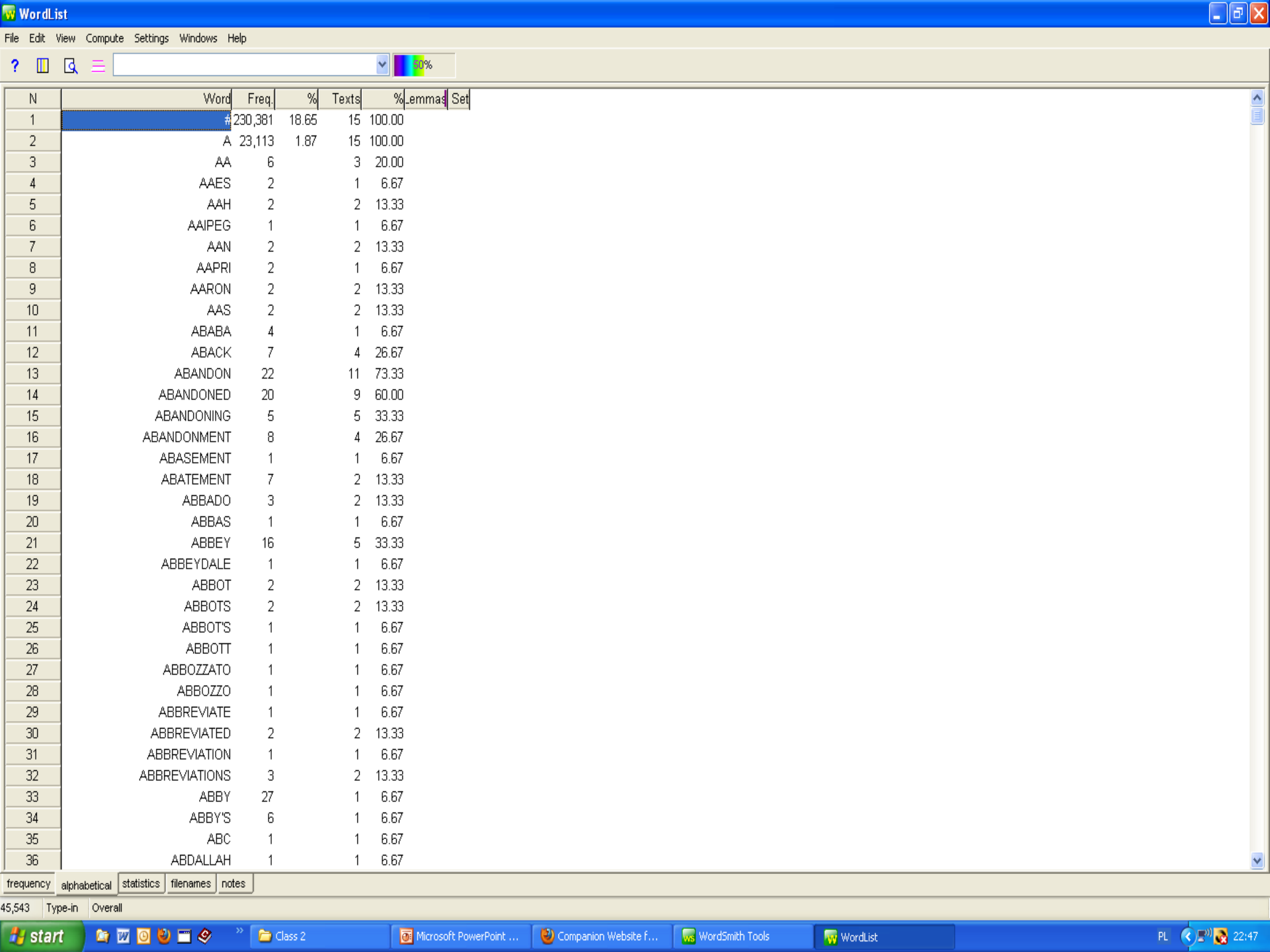
Basic types of analysis

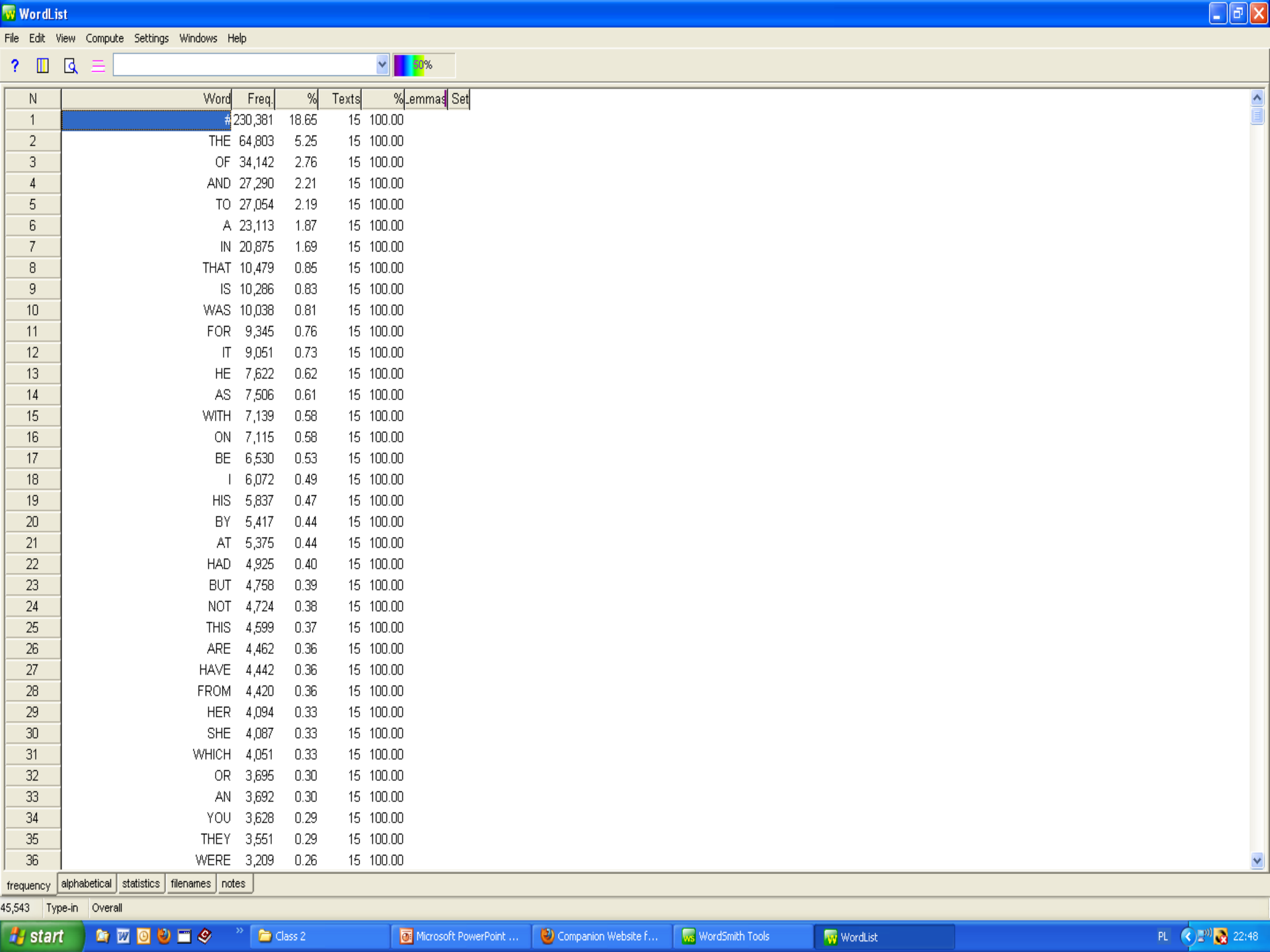
- basic statistics
  - number of words, sentences, paragraphs, n-letter words, etc. in a text or group of texts
- wordlists
  - alphabetical
  - frequency
- concordances = key words in context (KWIC)
  - sorting

Example of basic statistics


WordList
File Edit View Compute Settings Windows Help
50%
N Overall 1 2 3 4 5 6 7 8 9 10 11 12
text file Overall FLOB_A.TXT FLOB_B.TXT FLOB_C.TXT FLOB_D.TXT FLOB_E.TXT FLOB_F.TXT FLOB_G.TXT FLOB_H.TXT FLOB_J.TXT FLOB_K.TXT FLOB_L.TXT FLOB_M.TXT
file size 7,407,314 669,215 401,093 256,402 248,861 554,682 647,758 1,125,786 456,726 1,220,276 407,206 350,920 88,606
tokens (running words) in text 1,235,264 110,805 67,407 42,351 41,549 93,884 108,327 188,278 74,652 198,598 70,675 58,930 14,843
tokens used for word list 1,004,883 88,143 54,350 34,488 34,231 76,104 88,513 155,024 59,420 159,871 58,663 48,410 12,200
sum of entries
types (distinct words) 45,543 11,849 8,639 8,207 5,798 11,402 11,572 17,214 6,441 14,605 8,318 6,654 3,441
type/token ratio (TTR) 4.53 13.44 15.90 23.80 16.94 14.98 13.07 11.10 10.84 9.14 14.18 13.75 28.20
standardised TTR 39.03 40.51 41.63 45.31 37.91 40.19 39.04 40.51 33.20 35.32 40.05 38.74 43.06
standardised TTR std.dev. 59.81 57.90 56.68 51.74 60.06 58.10 61.02 58.01 65.86 64.08 58.54 59.95 52.30
standardised TTR basis 1,000 1,000 1,000 1,000 1,000 1,000 1,000 1,000 1,000 1,000 1,000 1,000 1,000
mean word length (in characters) 4.37 4.34 4.36 4.45 4.45 4.36 4.45 4.47 4.61 4.60 4.02 4.02 4.17
word length std.dev. 2.49 2.39 2.44 2.51 2.64 2.43 2.52 2.59 2.75 2.79 2.11 2.09 2.22
sentences 53,365 4,262 2,757 1,490 1,450 3,598 3,943 6,454 2,619 6,454 4,666 4,083 968
mean (in words) 18.83 20.68 19.71 23.15 23.61 21.15 22.45 24.02 22.69 24.77 12.57 11.86 12.60
std.dev. 13.46 10.18 11.85 12.36 15.25 11.54 13.20 14.45 19.17 14.16 10.99 10.02 10.29
paragraphs 15 1 1 1 1 1 1 1 1 1 1 1 1
mean (in words) 66,992.21 88,143.00 54,350.00 34,488.00 34,231.00 76,104.00 88,513.00 155,024.00 59,420.00 159,871.00 58,663.00 48,410.00 12,200.00
std.dev. 42,804.07
headings
mean (in words)
std.dev.
sections 15 1 1 1 1 1 1 1 1 1 1 1 1
mean (in words) 66,992.21 88,143.00 54,350.00 34,488.00 34,231.00 76,104.00 88,513.00 155,024.00 59,420.00 159,871.00 58,663.00 48,410.00 12,200.00
std.dev. 42,804.07
numbers removed 230,381 22,662 13,057 7,863 7,318 17,780 19,814 33,254 15,232 38,727 12,012 10,520 2,643
stoplist tokens removed
stoplist types removed
1-letter words 39,502 3,278 1,818 1,258 1,198 3,038 2,776 5,275 2,277 7,200 2,939 2,166 587
2-letter words 219,055 18,774 12,140 7,040 8,101 15,889 19,292 34,495 13,951 37,184 12,376 9,938 2,269
3-letter words 377,059 34,485 20,726 12,803 12,376 28,428 32,397 56,070 21,861 55,680 22,749 19,650 4,778
4-letter words 162,139 14,648 8,821 5,470 4,933 12,948 14,136 22,776 8,000 22,562 10,831 9,505 2,334
5-letter words 110,759 10,348 6,054 4,073 3,411 8,926 9,731 16,601 5,847 16,022 7,008 5,498 1,478
6-letter words 86,299 8,057 4,941 3,132 2,693 7,039 7,582 13,264 4,618 12,576 5,232 4,248 1,073
7-letter words 80,459 7,811 4,364 2,964 2,550 6,283 7,352 12,404 5,075 13,280 4,050 3,386 968
8-letter words 57,893 5,228 3,174 2,036 2,038 4,165 5,340 9,395 4,292 11,087 2,454 1,978 534
frequency alphabetical statistics filenames notes
77 Type-in Overall
start
Class 2
Microsoft PowerPoint ...
Companion Website f...
WordSmith Tools
WordList
PL
22:46

Examples of wordlists

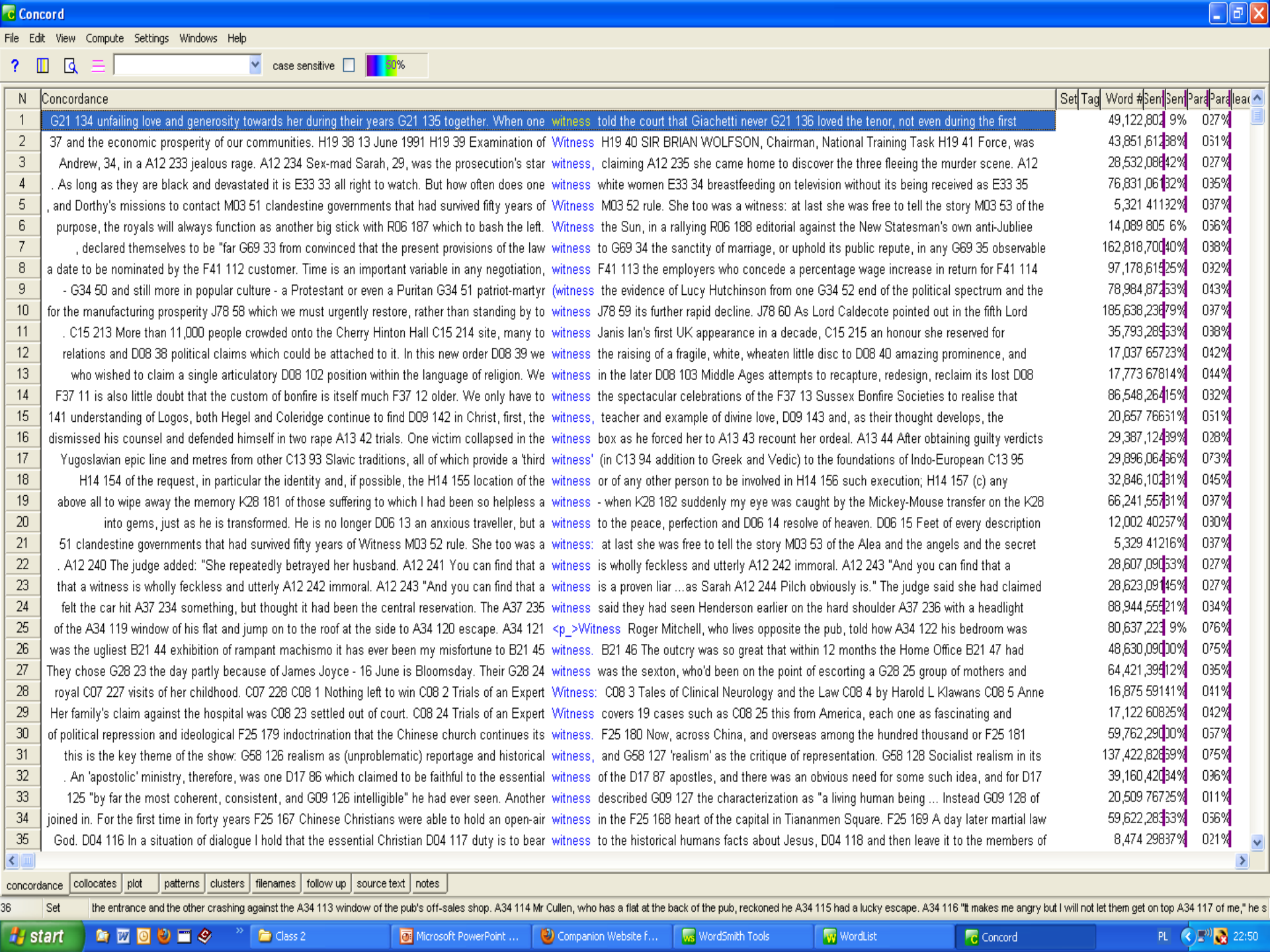

WordList
File Edit View Compute Settings Windows Help
80%
N Word Freq. % Texts % Lemmas Set
1 # 230,381 18.65 15 100.00
2 A 23,113 1.87 15 100.00
3 AA 6 3 20.00
4 AAES 2 1 6.67
5 AAH 2 2 13.33
6 AAIPEG 1 1 6.67
7 AAN 2 2 13.33
8 AAPRI 2 1 6.67
9 AARON 2 2 13.33
10 AAS 2 2 13.33
11 ABABA 4 1 6.67
12 ABACK 7 4 26.67
13 ABANDON 22 11 73.33
14 ABANDONED 20 9 60.00
15 ABANDONING 5 5 33.33
16 ABANDONMENT 8 4 26.67
17 ABASEMENT 1 1 6.67
18 ABATEMENT 7 2 13.33
19 ABBADO 3 2 13.33
20 ABBAS 1 1 6.67
21 ABBEY 16 5 33.33
22 ABBEYDALE 1 1 6.67
23 ABBOT 2 2 13.33
24 ABBOTS 2 2 13.33
25 ABBOT'S 1 1 6.67
26 ABBOTT 1 1 6.67
27 ABBOZZATO 1 1 6.67
28 ABBOZZO 1 1 6.67
29 ABBREVIATE 1 1 6.67
30 ABBREVIATED 2 2 13.33
31 ABBREVIATION 1 1 6.67
32 ABBREVIATIONS 3 2 13.33
33 ABBY 27 1 6.67
34 ABBY'S 6 1 6.67
35 ABC 1 1 6.67
36 ABDALLAH 1 1 6.67
frequency alphabetical statistics filenames notes
45,543 Type-in Overall
start
Class 2
Microsoft PowerPoint ...
Companion Website f...
WordSmith Tools
WordList
PL
22:47

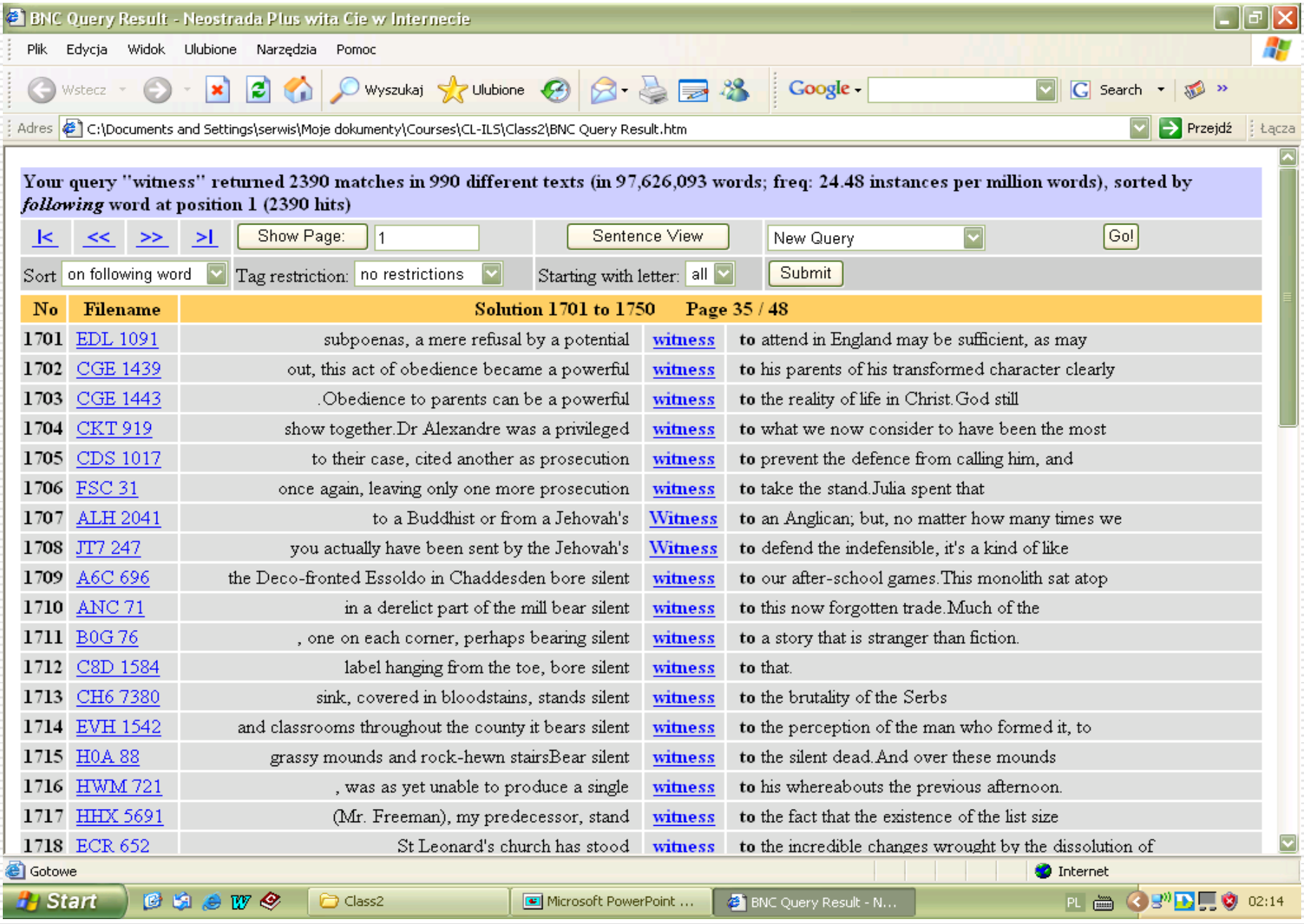

WordList
File Edit View Compute Settings Windows Help
50%
N Word Freq. % Texts % Lemmas Set
1 # 230,381 18.65 15 100.00
2 THE 64,803 5.25 15 100.00
3 OF 34,142 2.76 15 100.00
4 AND 27,290 2.21 15 100.00
5 TO 27,054 2.19 15 100.00
6 A 23,113 1.87 15 100.00
7 IN 20,875 1.69 15 100.00
8 THAT 10,479 0.85 15 100.00
9 IS 10,286 0.83 15 100.00
10 WAS 10,038 0.81 15 100.00
11 FOR 9,345 0.76 15 100.00
12 IT 9,051 0.73 15 100.00
13 HE 7,622 0.62 15 100.00
14 AS 7,506 0.61 15 100.00
15 WITH 7,139 0.58 15 100.00
16 ON 7,115 0.58 15 100.00
17 BE 6,530 0.53 15 100.00
18 I 6,072 0.49 15 100.00
19 HIS 5,837 0.47 15 100.00
20 BY 5,417 0.44 15 100.00
21 AT 5,375 0.44 15 100.00
22 HAD 4,925 0.40 15 100.00
23 BUT 4,758 0.39 15 100.00
24 NOT 4,724 0.38 15 100.00
25 THIS 4,599 0.37 15 100.00
26 ARE 4,462 0.36 15 100.00
27 HAVE 4,442 0.36 15 100.00
28 FROM 4,420 0.36 15 100.00
29 HER 4,094 0.33 15 100.00
30 SHE 4,087 0.33 15 100.00
31 WHICH 4,051 0.33 15 100.00
32 OR 3,695 0.30 15 100.00
33 AN 3,692 0.30 15 100.00
34 YOU 3,628 0.29 15 100.00
35 THEY 3,551 0.29 15 100.00
36 WERE 3,209 0.26 15 100.00
frequency alphabetical statistics filenames notes
45,543 Type-in Overall
start Class 2 Microsoft PowerPoint ... Companion Website f... WordSmith Tools WordList PL 22:48

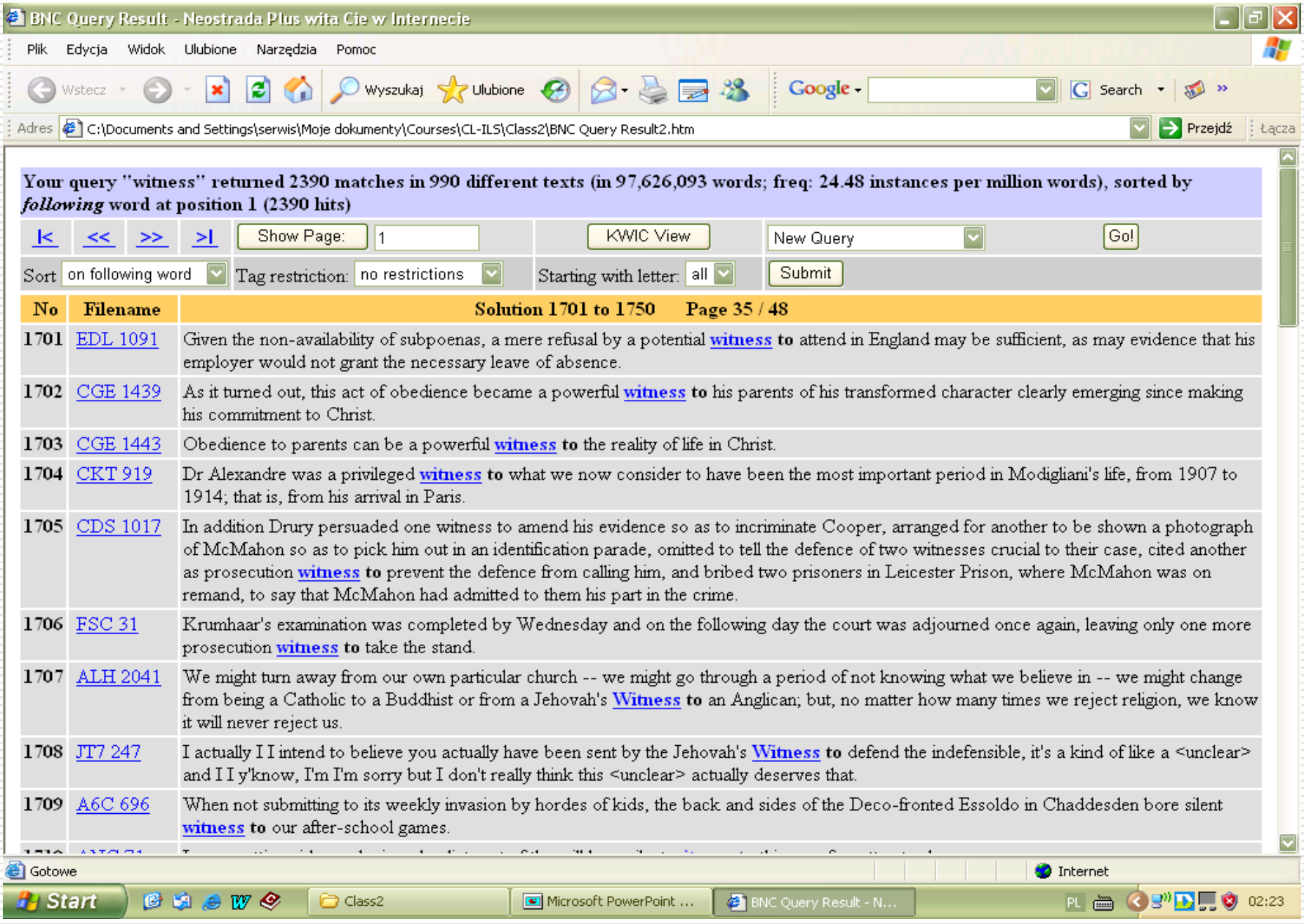
Frequency lists

- http://ucrel.lancs.ac.uk/bncfreq/

Examples of concordances


Concord
File Edit View Compute Settings Windows Help
case sensitive
N Concordance Set Tag Word # Sen Sen Para Para lea
1 G21 134 unfailing love and generosity towards her during their years G21 135 together. When one witness told the court that Giachetti never G21 136 loved the tenor, not even during the first 49,122,802 9% 027%
2 37 and the economic prosperity of our communities. H19 38 13 June 1991 H19 39 Examination of Witness H19 40 SIR BRIAN WOLFSON, Chairman, National Training Task H19 41 Force, was 43,851,612 38% 051%
3 Andrew, 34, in a A12 233 jealous rage. A12 234 Sex-mad Sarah, 29, was the prosecution's star witness, claiming A12 235 she came home to discover the three fleeing the murder scene. A12 28,532,088 42% 027%
4 . As long as they are black and devastated it is E33 33 all right to watch. But how often does one witness white women E33 34 breastfeeding on television without its being received as E33 35 76,831,061 32% 035%
5 , and Dorthy's missions to contact M03 51 clandestine governments that had survived fifty years of Witness M03 52 rule. She too was a witness: at last she was free to tell the story M03 53 of the 5,321 411 32% 037%
6 purpose, the royals will always function as another big stick with R06 187 which to bash the left. Witness the Sun, in a rallying R06 188 editorial against the New Statesman's own anti-Jubliee 14,089 805 6% 036%
7 , declared themselves to be "far G69 33 from convinced that the present provisions of the law witness to G69 34 the sanctity of marriage, or uphold its public repute, in any G69 35 observable 162,818,700 40% 038%
8 a date to be nominated by the F41 112 customer. Time is an important variable in any negotiation, witness F41 113 the employers who concede a percentage wage increase in return for F41 114 97,178,615 25% 032%
9 - G34 50 and still more in popular culture - a Protestant or even a Puritan G34 51 patriot-martyr (witness the evidence of Lucy Hutchinson from one G34 52 end of the political spectrum and the 78,984,872 53% 043%
10 for the manufacturing prosperity J78 58 which we must urgently restore, rather than standing by to witness J78 59 its further rapid decline. J78 60 As Lord Caldecote pointed out in the fifth Lord 185,638,236 79% 037%
11 . C15 213 More than 11,000 people crowded onto the Cherry Hinton Hall C15 214 site, many to witness Janis Ian's first UK appearance in a decade, C15 215 an honour she reserved for 35,793,289 53% 038%
12 relations and D08 38 political claims which could be attached to it. In this new order D08 39 we witness the raising of a fragile, white, wheaten little disc to D08 40 amazing prominence, and 17,037 657 23% 042%
13 who wished to claim a single articulatory D08 102 position within the language of religion. We witness in the later D08 103 Middle Ages attempts to recapture, redesign, reclaim its lost D08 17,773 678 14% 044%
14 F37 11 is also little doubt that the custom of bonfire is itself much F37 12 older. We only have to witness the spectacular celebrations of the F37 13 Sussex Bonfire Societies to realise that 86,548,264 15% 032%
15 141 understanding of Logos, both Hegel and Coleridge continue to find D09 142 in Christ, first, the witness, teacher and example of divine love, D09 143 and, as their thought develops, the 20,657 766 31% 051%
16 dismissed his counsel and defended himself in two rape A13 42 trials. One victim collapsed in the witness box as he forced her to A13 43 recount her ordeal. A13 44 After obtaining guilty verdicts 29,387,124 39% 028%
17 Yugoslavian epic line and metres from other C13 93 Slavic traditions, all of which provide a 'third witness' (in C13 94 addition to Greek and Vedic) to the foundations of Indo-European C13 95 29,896,064 66% 073%
18 H14 154 of the request, in particular the identity and, if possible, the H14 155 location of the witness or of any other person to be involved in H14 156 such execution; H14 157 (c) any 32,846,102 31% 045%
19 above all to wipe away the memory K28 181 of those suffering to which I had been so helpless a witness - when K28 182 suddenly my eye was caught by the Mickey-Mouse transfer on the K28 66,241,557 31% 037%
20 into gems, just as he is transformed. He is no longer D06 13 an anxious traveller, but a witness to the peace, perfection and D06 14 resolve of heaven. D06 15 Feet of every description 12,002 402 57% 030%
21 51 clandestine governments that had survived fifty years of Witness M03 52 rule. She too was a witness: at last she was free to tell the story M03 53 of the Alea and the angels and the secret 5,329 412 16% 037%
22 . A12 240 The judge added: "She repeatedly betrayed her husband. A12 241 You can find that a witness is wholly feckless and utterly A12 242 immoral. A12 243 "And you can find that a 28,607,090 53% 027%
23 that a witness is wholly feckless and utterly A12 242 immoral. A12 243 "And you can find that a witness is a proven liar ...as Sarah A12 244 Pilch obviously is." The judge said she had claimed 28,623,091 45% 027%
24 felt the car hit A37 234 something, but thought it had been the central reservation. The A37 235 witness said they had seen Henderson earlier on the hard shoulder A37 236 with a headlight 88,944,555 21% 034%
25 of the A34 119 window of his flat and jump on to the roof at the side to A34 120 escape. A34 121 <p_>Witness Roger Mitchell, who lives opposite the pub, told how A34 122 his bedroom was 80,637,223 9% 076%
26 was the ugliest B21 44 exhibition of rampant machismo it has ever been my misfortune to B21 45 witness. B21 46 The outcry was so great that within 12 months the Home Office B21 47 had 48,630,090 00% 075%
27 They chose G28 23 the day partly because of James Joyce - 16 June is Bloomsday. Their G28 24 witness was the sexton, who'd been on the point of escorting a G28 25 group of mothers and 64,421,395 12% 035%
28 royal C07 227 visits of her childhood. C07 228 C08 1 Nothing left to win C08 2 Trials of an Expert Witness: C08 3 Tales of Clinical Neurology and the Law C08 4 by Harold L Klawans C08 5 Anne 16,875 591 41% 041%
29 Her family's claim against the hospital was C08 23 settled out of court. C08 24 Trials of an Expert Witness covers 19 cases such as C08 25 this from America, each one as fascinating and 17,122 608 25% 042%
30 of political repression and ideological F25 179 indoctrination that the Chinese church continues its witness. F25 180 Now, across China, and overseas among the hundred thousand or F25 181 59,762,290 00% 057%
31 this is the key theme of the show: G58 126 realism as (unproblematic) reportage and historical witness, and G58 127 'realism' as the critique of representation. G58 128 Socialist realism in its 137,422,828 59% 075%
32 . An 'apostolic' ministry, therefore, was one D17 86 which claimed to be faithful to the essential witness of the D17 87 apostles, and there was an obvious need for some such idea, and for D17 39,160,420 34% 096%
33 125 "by far the most coherent, consistent, and G09 126 intelligible" he had ever seen. Another witness described G09 127 the characterization as "a living human being ... Instead G09 128 of 20,509 767 25% 011%
34 joined in. For the first time in forty years F25 167 Chinese Christians were able to hold an open-air witness in the F25 168 heart of the capital in Tiananmen Square. F25 169 A day later martial law 59,622,283 53% 056%
35 God. D04 116 In a situation of dialogue I hold that the essential Christian D04 117 duty is to bear witness to the historical humans facts about Jesus, D04 118 and then leave it to the members of 8,474 298 37% 021%
concordance collocates plot patterns clusters filenames follow up source text notes
36 Set the entrance and the other crashing against the A34 113 window of the pub's off-sales shop. A34 114 Mr Cullen, who has a flat at the back of the pub, reckoned he A34 115 had a lucky escape. A34 116 "It makes me angry but I will not let them get on top A34 117 of me," he s
start Class 2 Microsoft PowerPoint ... Companion Website f... WordSmith Tools WordList Concord PL 22:50


BNC Query Result - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz Wyszukaj Ulubione Google Search
Adres C:\Documents and Settings\serwis\Moje dokumenty\Courses\CL-ILS\Class2\BNC Query Result.htm
Przejdź Łącza
Your query "witness" returned 2390 matches in 990 different texts (in 97,626,093 words; freq: 24.48 instances per million words), sorted by following word at position 1 (2390 hits)
|< << >> >| Show Page: 1 Sentence View New Query Go!
Sort on following word Tag restriction: no restrictions Starting with letter: all Submit
No Filename Solution 1701 to 1750 Page 35 / 48
1701 EDL 1091 subpoenas, a mere refusal by a potential witness to attend in England may be sufficient, as may
1702 CGE 1439 out, this act of obedience became a powerful witness to his parents of his transformed character clearly
1703 CGE 1443 .Obedience to parents can be a powerful witness to the reality of life in Christ.God still
1704 CKT 919 show together.Dr Alexandre was a privileged witness to what we now consider to have been the most
1705 CDS 1017 to their case, cited another as prosecution witness to prevent the defence from calling him, and
1706 FSC 31 once again, leaving only one more prosecution witness to take the stand.Julia spent that
1707 ALH 2041 to a Buddhist or from a Jehovah's Witness to an Anglican; but, no matter how many times we
1708 JT7 247 you actually have been sent by the Jehovah's Witness to defend the indefensible, it's a kind of like
1709 A6C 696 the Deco-fronted Essoldo in Chaddesden bore silent witness to our after-school games.This monolith sat atop
1710 ANC 71 in a derelict part of the mill bear silent witness to this now forgotten trade.Much of the
1711 B0G 76 , one on each corner, perhaps bearing silent witness to a story that is stranger than fiction.
1712 C8D 1584 label hanging from the toe, bore silent witness to that.
1713 CH6 7380 sink, covered in bloodstains, stands silent witness to the brutality of the Serbs
1714 EVH 1542 and classrooms throughout the county it bears silent witness to the perception of the man who formed it, to
1715 H0A 88 grassy mounds and rock-hewn stairsBear silent witness to the silent dead.And over these mounds
1716 HWM 721 , was as yet unable to produce a single witness to his whereabouts the previous afternoon.
1717 HHX 5691 (Mr. Freeman), my predecessor, stand witness to the fact that the existence of the list size
1718 ECR 652 St Leonard's church has stood witness to the incredible changes wrought by the dissolution of
Gotowe
Internet
Start
Class2
Microsoft PowerPoint ...
BNC Query Result - N...
PL
02:14


BNC Query Result - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz
Wyszukaj
Ulubione
Google
Search
Adres
C:\Documents and Settings\serwis\Moje dokumenty\Courses\CL-ILS\Class2\BNC Query Result2.htm
Przejdź
Łącza
Your query "witness" returned 2390 matches in 990 different texts (in 97,626,093 words; freq: 24.48 instances per million words), sorted by following word at position 1 (2390 hits)
|< << >> >|
Show Page: 1
KWIC View
New Query
Go!
Sort on following word
Tag restriction: no restrictions
Starting with letter: all
Submit
No Filename Solution 1701 to 1750 Page 35 / 48
1701 EDL 1091 Given the non-availability of subpoenas, a mere refusal by a potential witness to attend in England may be sufficient, as may evidence that his employer would not grant the necessary leave of absence.
1702 CGE 1439 As it turned out, this act of obedience became a powerful witness to his parents of his transformed character clearly emerging since making his commitment to Christ.
1703 CGE 1443 Obedience to parents can be a powerful witness to the reality of life in Christ.
1704 CKT 919 Dr Alexandre was a privileged witness to what we now consider to have been the most important period in Modigliani's life, from 1907 to 1914; that is, from his arrival in Paris.
1705 CDS 1017 In addition Drury persuaded one witness to amend his evidence so as to incriminate Cooper, arranged for another to be shown a photograph of McMahon so as to pick him out in an identification parade, omitted to tell the defence of two witnesses crucial to their case, cited another as prosecution witness to prevent the defence from calling him, and bribed two prisoners in Leicester Prison, where McMahon was on remand, to say that McMahon had admitted to them his part in the crime.
1706 FSC 31 Krumhaar's examination was completed by Wednesday and on the following day the court was adjourned once again, leaving only one more prosecution witness to take the stand.
1707 ALH 2041 We might turn away from our own particular church -- we might go through a period of not knowing what we believe in -- we might change from being a Catholic to a Buddhist or from a Jehovah's Witness to an Anglican; but, no matter how many times we reject religion, we know it will never reject us.
1708 JT7 247 I actually I I intend to believe you actually have been sent by the Jehovah's Witness to defend the indefensible, it's a kind of like a <unclear> and I I y'know, I'm I'm sorry but I don't really think this <unclear> actually deserves that.
1709 A6C 696 When not submitting to its weekly invasion by hordes of kids, the back and sides of the Deco-fronted Essoldo in Chaddesden bore silent witness to our after-school games.
Gotowe
Internet
Start
Class2
Microsoft PowerPoint ...
BNC Query Result - N...
PL
02:23

More sophisticated functionalities

- extraction of n-grams
- extractions of collocations
- keyword analysis

Key terms

- **compile:** collect and put together (for example, texts for a *corpus*).
- **collocation:** words appearing together.
- **collocate (v,n):** to appear together, or words that appear together. (In the *collocations* 'apple tree', 'apple pie', and 'Adam's apple', 'apple' *collocates* with 'tree', 'pie', and 'Adam's'. They are *collocates*.)
- **colligation:** patterns based on syntactic groups rather than individual words, e.g . *interested in*, *like* + *ving* (Barnbrook 1996).

Key terms (2)

- **token:** individual or running word
- **type:** word form. "I see a cat and a dog" contains seven *tokens* but only six types (the type 'a' occurrs twice).
- **lemma:** the set of different forms of a word, such as the inflected forms of a verb, e.g. 'sing', 'sang', 'sung' are one lemma, 'boy', 'boys' another.
- **lemmatisation:** the process of converting the words in a text into *lemmas*. We often talk of *lemmatised wordlists*.

Key terms (3)

- **string:** combination of letters or other characters that you are searching for in a corpus, e.g 'a nice day'.
- **hit:** When your search *string* is found in the *corpus*, it is referred to as a hit or *match*.
- **match:** When your search *string* is found in the *corpus*, it is referred to as a match or *hit*. You can say: "The search has returned 47 hits/matches."

Key terms (4)

- **concordance (line):** A word/phrase you have searched for (hit/match) displayed with its surrounding *context*.
- **context:** the words surrounding a *hit*.
- **KWIC** (Key-Word In Context)**:** a form of displaying a *concordance* where the *hit* is shown in the centre of a page with a certain amount of characters of *context* on both sides. This display is useful for observing characteristic patterns.

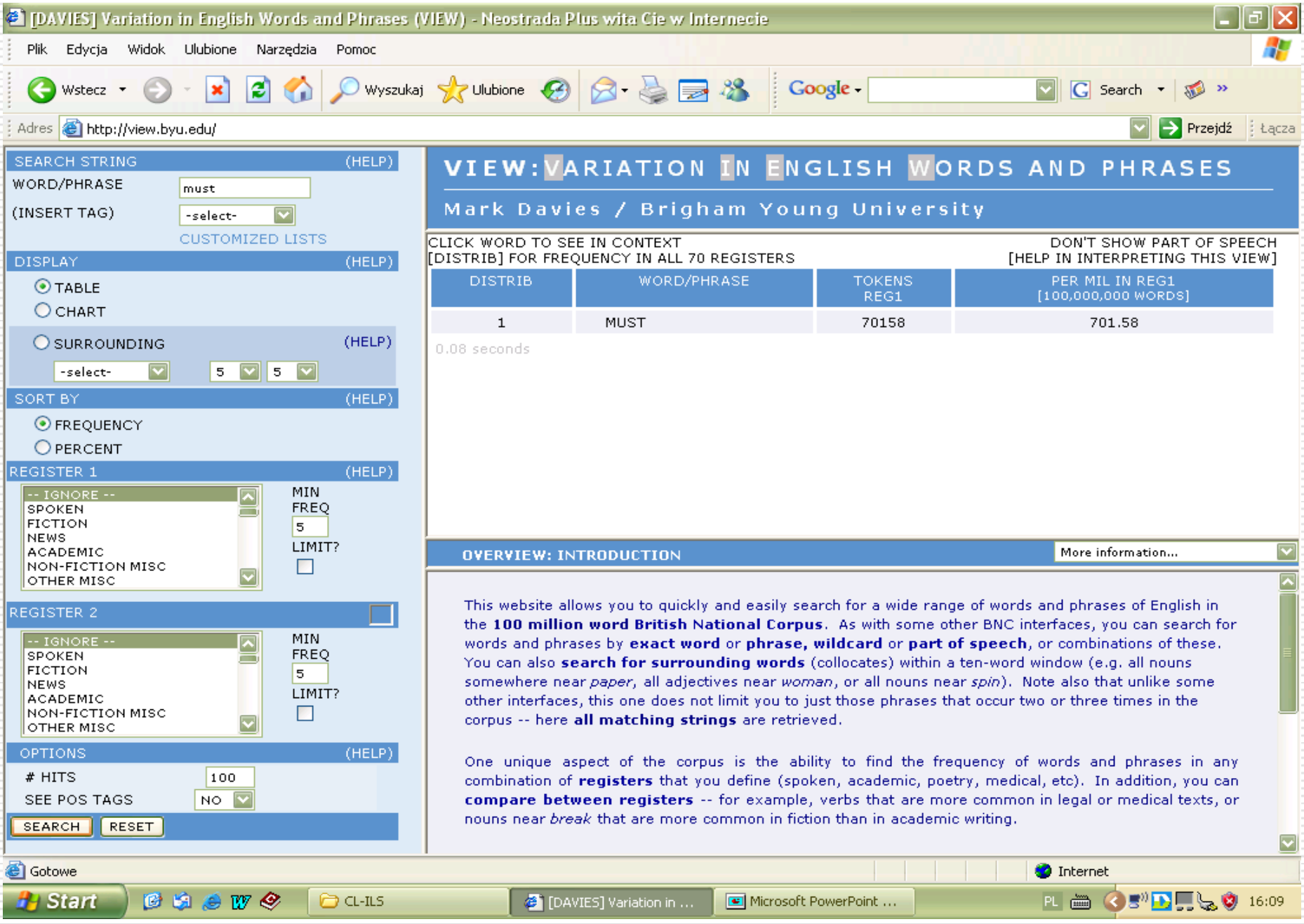
Advantages and disadvantages of corpora

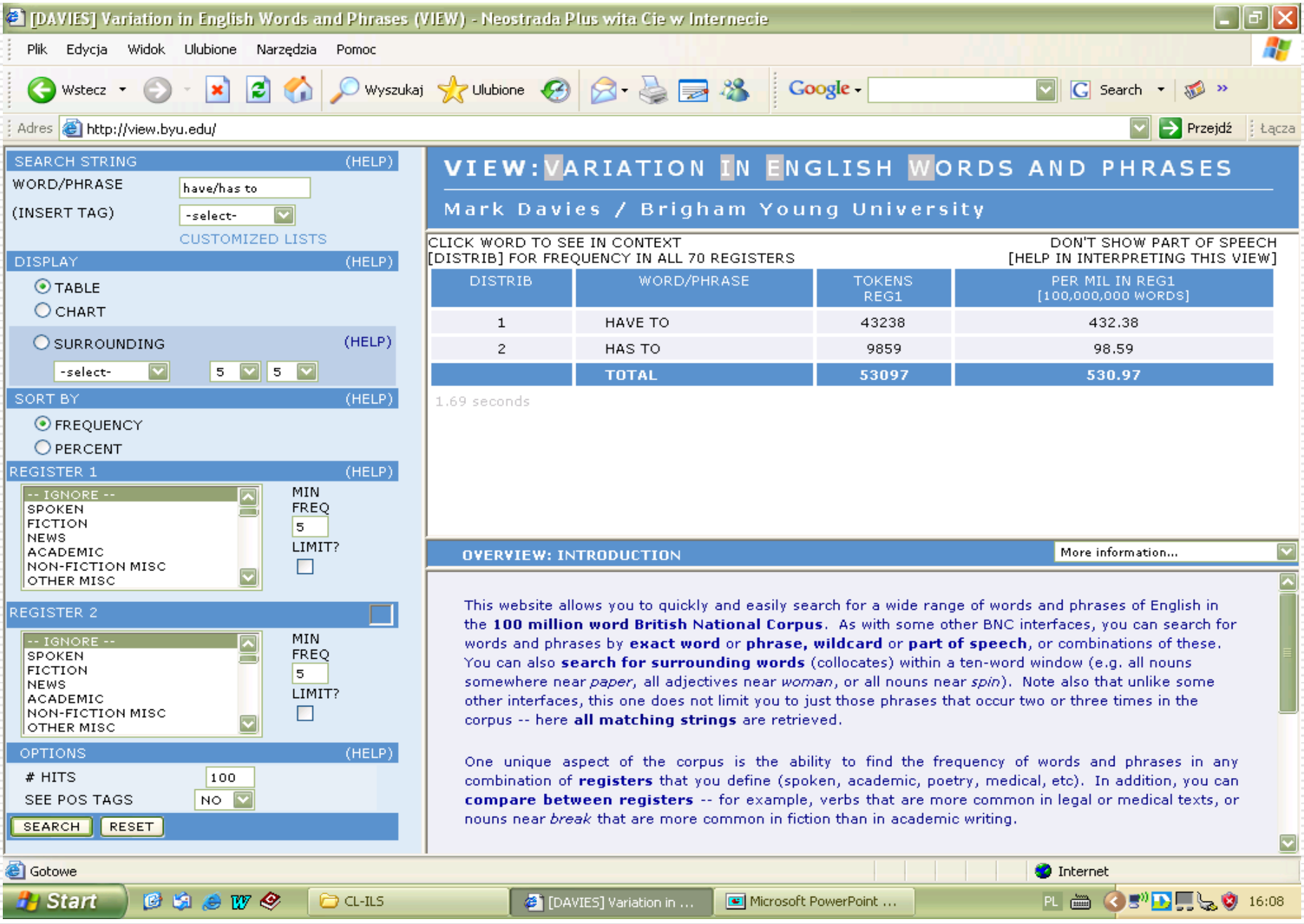
- information on frequency, meanings (e.g. pragmatic meaning, semantic prosody), collocation and phraseology
- information on what is probable but not what is possible in language

  (If a particular word or collocation cannot be found in a corpus, this does not mean they do not exist in lanaguge, they are just very rare.)

Example 1 – *must vs. have to*

- Work in groups and answer the following questions
  - What is the difference in meaning between *must* and *have to*?
  - What are the differences in the frequencies of the two verbs in English?
  - Are they the same across registers?

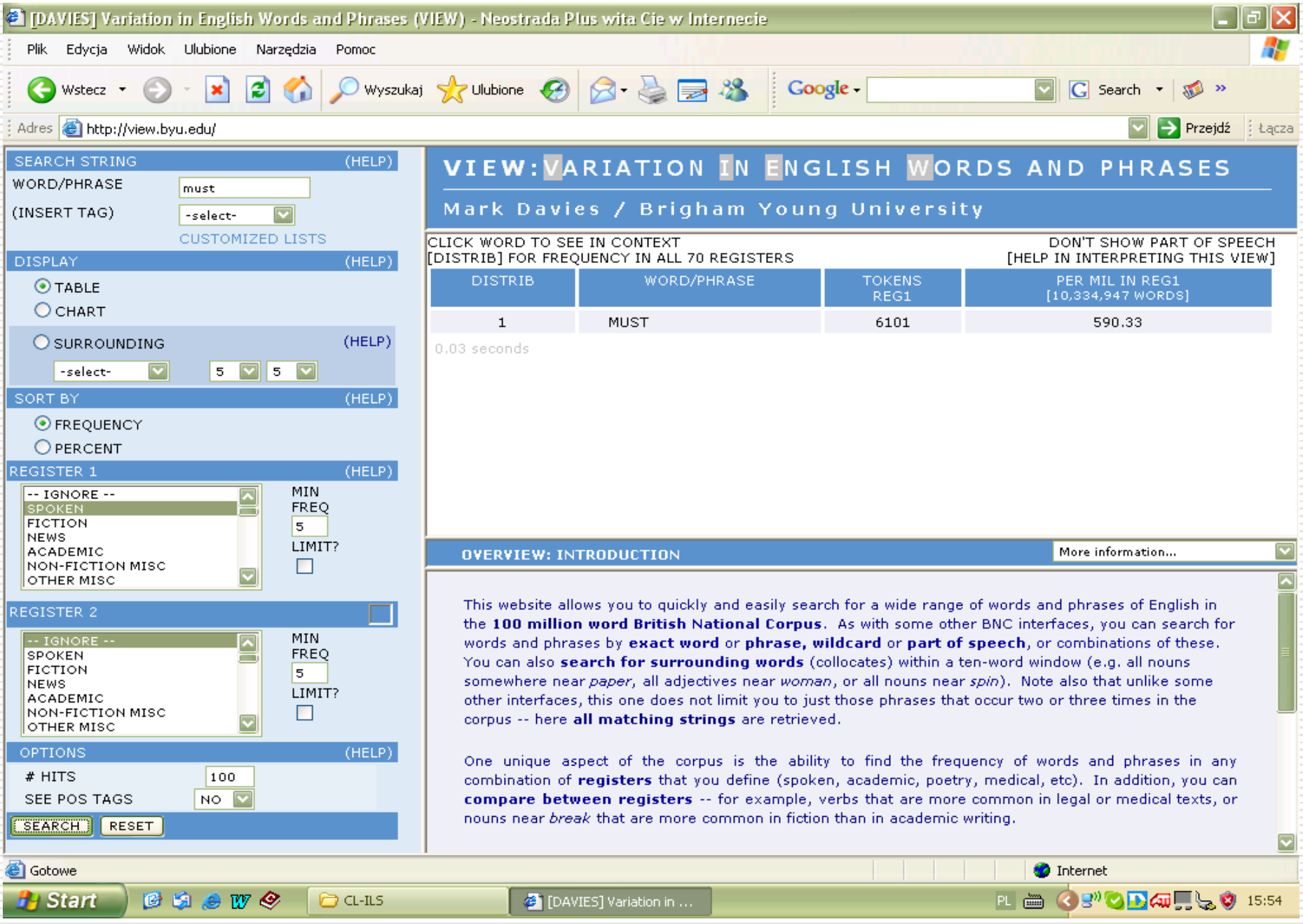

[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz Wyszukaj Ulubione Google Search
Adres http://view.byu.edu/
Przejdź Łącza
SEARCH STRING (HELP)
WORD/PHRASE must
(INSERT TAG) -select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select- 5 5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ 5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ 5
LIMIT?
OPTIONS (HELP)
HITS 100
SEE POS TAGS NO
SEARCH RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB WORD/PHRASE TOKENS REG1 PER MIL IN REG1 [100,000,000 WORDS]
1 MUST 70158 701.58
0.08 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Gotowe
Internet
Start
CL-ILS
[DAVIES] Variation in ...
Microsoft PowerPoint ...
PL
16:09

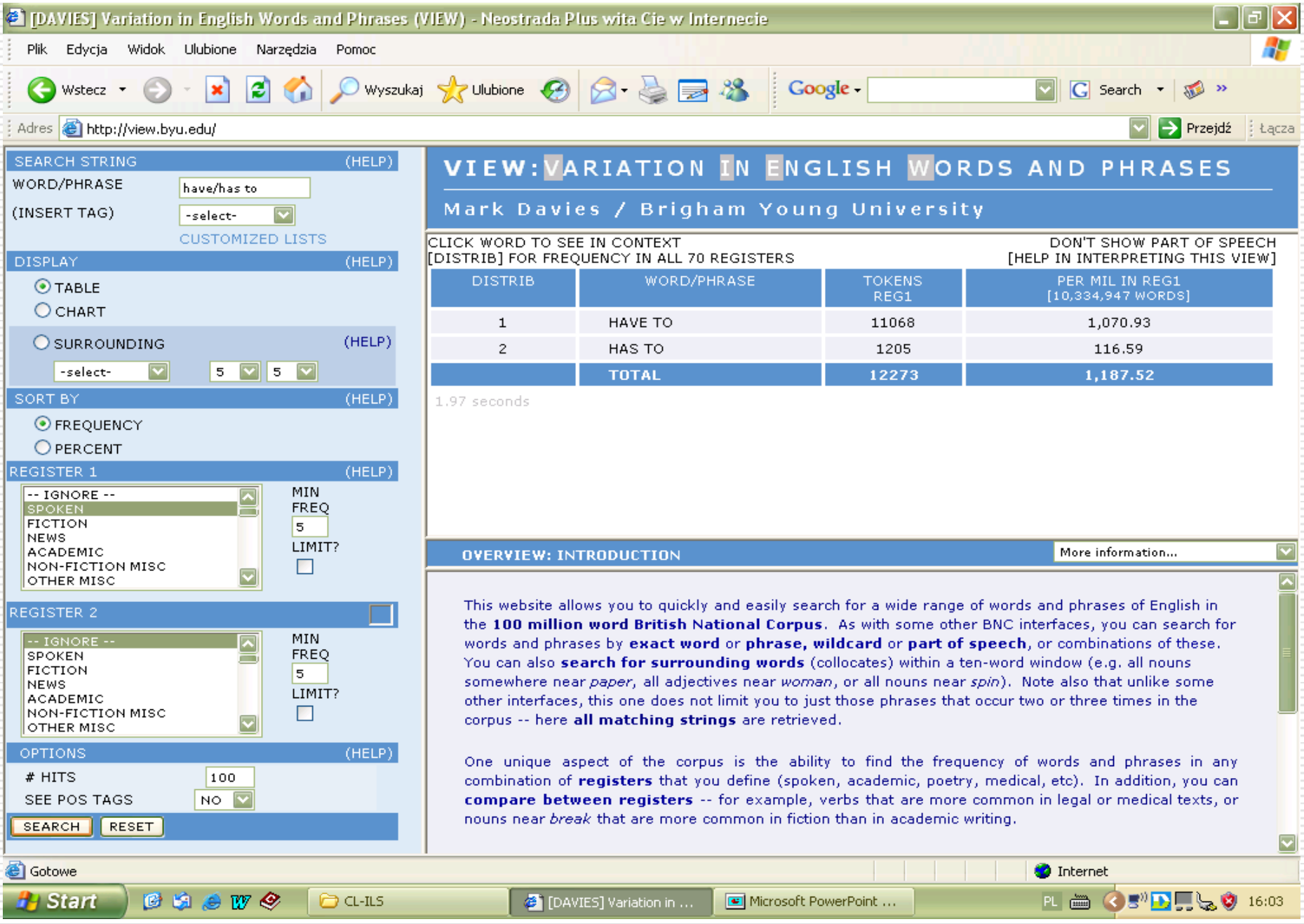

[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz Wyszukaj Ulubione Google Search
Adres http://view.byu.edu/
Przejdź Łącza
SEARCH STRING (HELP)
WORD/PHRASE have/has to
(INSERT TAG) -select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select- 5 5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ 5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ 5
LIMIT?
OPTIONS (HELP)
HITS 100
SEE POS TAGS NO
SEARCH RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB | WORD/PHRASE | TOKENS REG1 | PER MIL IN REG1 [100,000,000 WORDS]
1 | HAVE TO | 43238 | 432.38
2 | HAS TO | 9859 | 98.59
TOTAL | 53097 | 530.97
1.69 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Gotowe
Internet
Start
CL-ILS
[DAVIES] Variation in ...
Microsoft PowerPoint ...
PL
16:08

must vs. have/has to

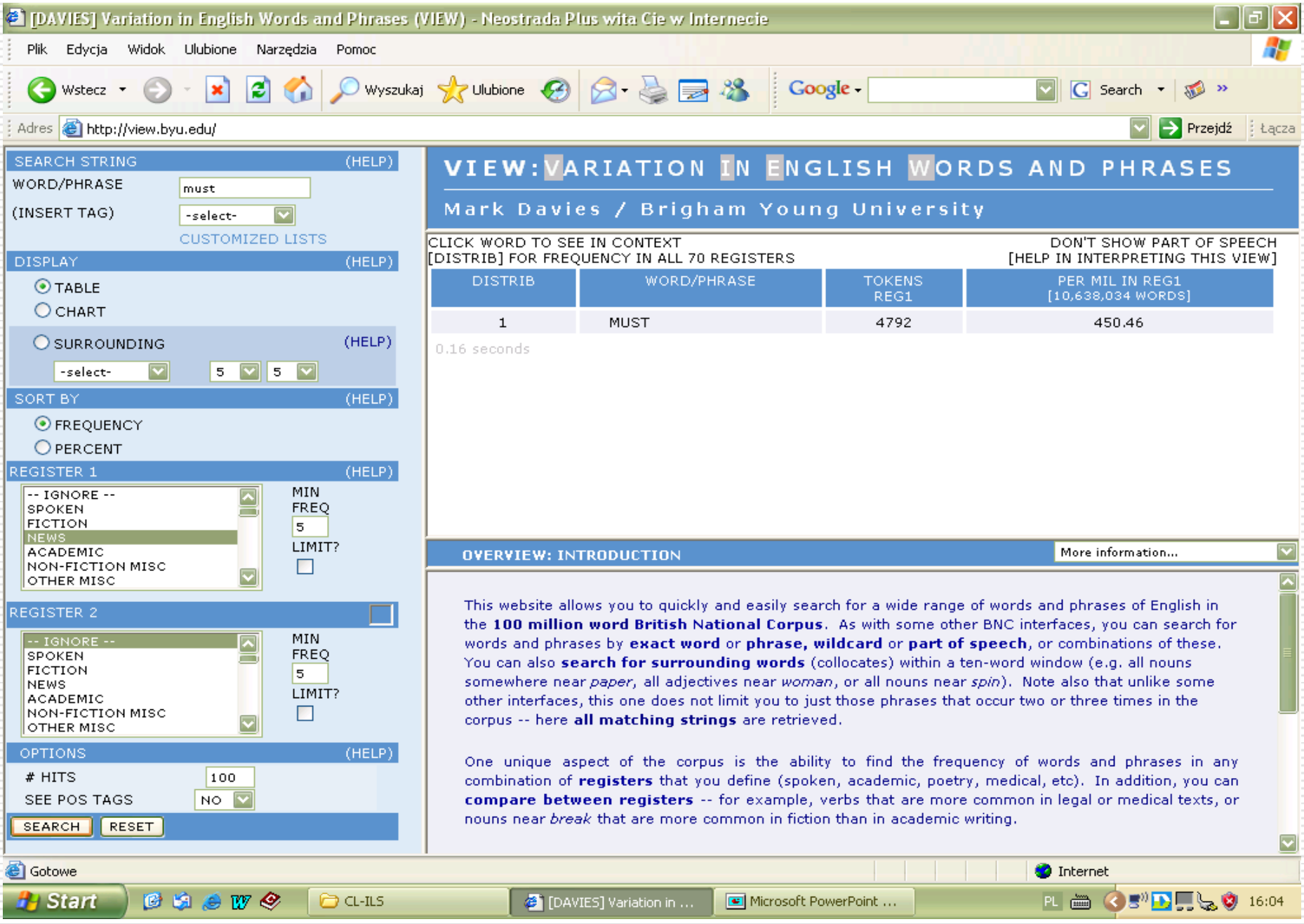
must	have/has to
70,158	53,097

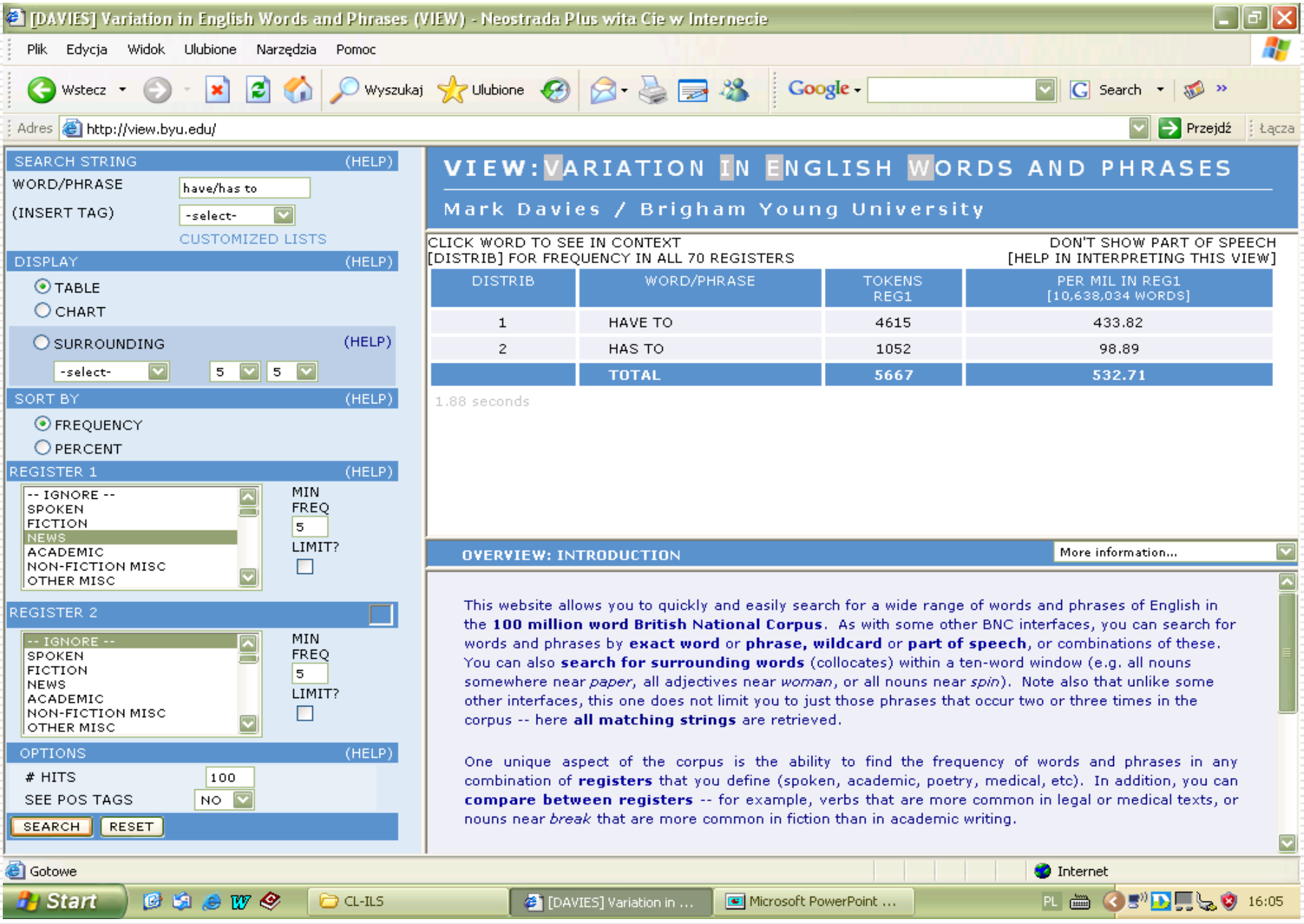
Must and *have to* in spoken English


[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz
Wyszukaj
Ulubione
Google
Search
Adres
http://view.byu.edu/
Przejdź
Łącza
SEARCH STRING (HELP)
WORD/PHRASE
must
(INSERT TAG)
-select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select-
5
5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
OPTIONS (HELP)
HITS
100
SEE POS TAGS
NO
SEARCH
RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB
WORD/PHRASE
TOKENS REG1
PER MIL IN REG1 [10,334,947 WORDS]
1
MUST
6101
590.33
0.03 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Gotowe
Internet
Start
CL-ILS
[DAVIES] Variation in ...
PL
15:54

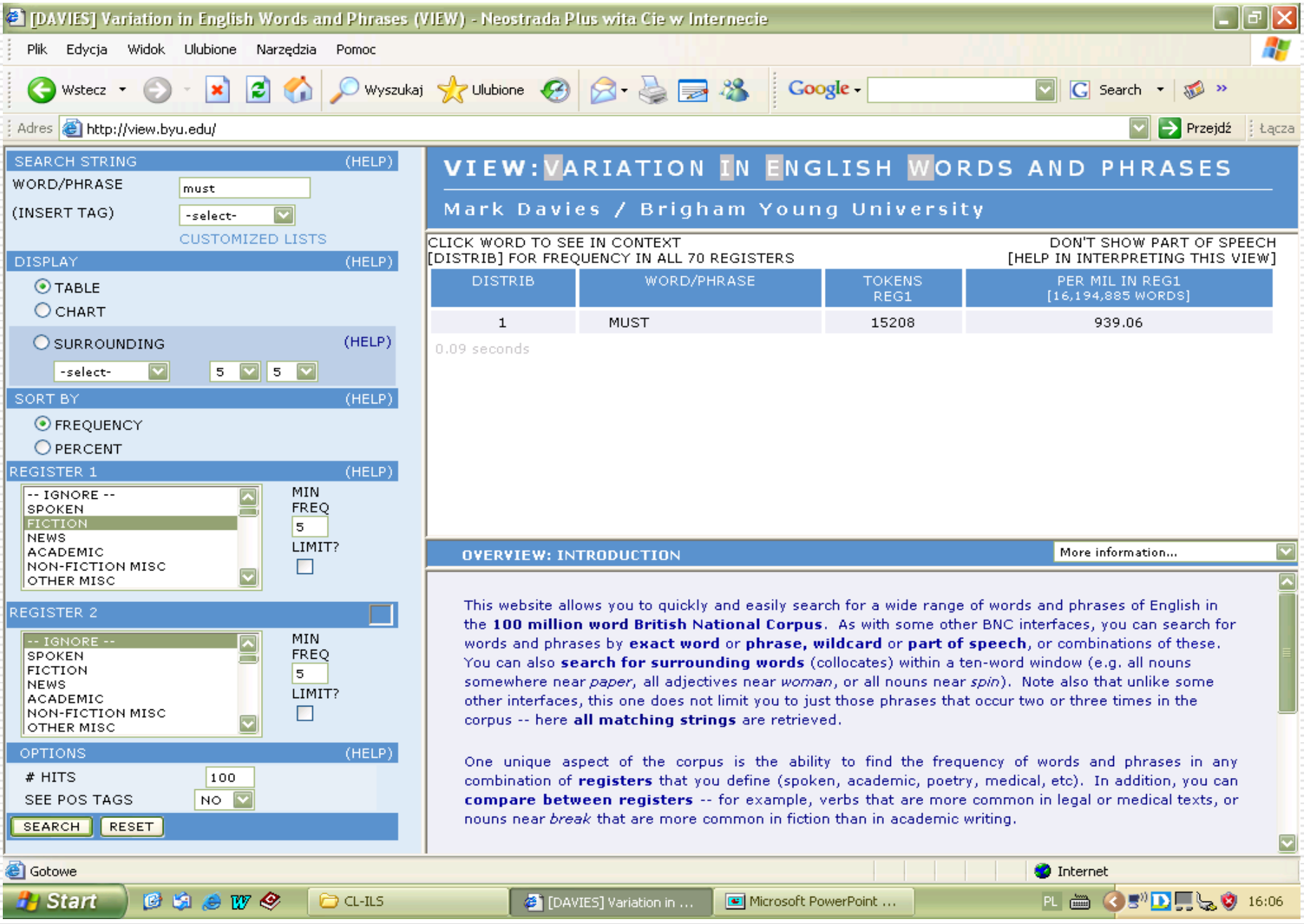

[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz
Wyszukaj
Ulubione
Google
Search
Adres
http://view.byu.edu/
Przejdź
Łącza
SEARCH STRING (HELP)
WORD/PHRASE
have/has to
(INSERT TAG)
-select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select-
5
5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
OPTIONS (HELP)
HITS
100
SEE POS TAGS
NO
SEARCH
RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB
WORD/PHRASE
TOKENS REG1
PER MIL IN REG1 [10,334,947 WORDS]
1
HAVE TO
11068
1,070.93
2
HAS TO
1205
116.59
TOTAL
12273
1,187.52
1.97 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Gotowe
Internet
Start
CL-ILS
[DAVIES] Variation in ...
Microsoft PowerPoint ...
PL
16:03

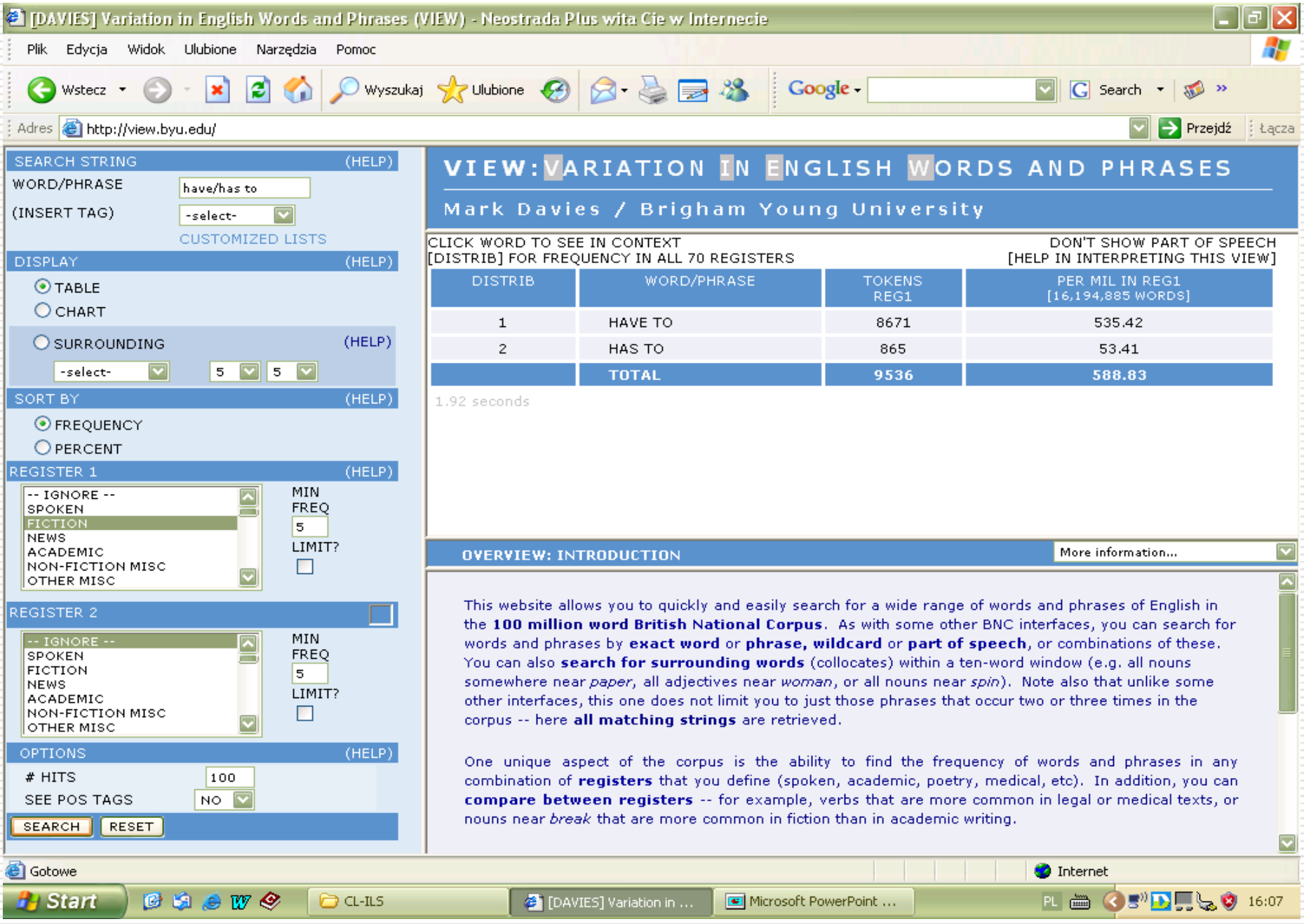
Must and *have to* in the news


[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz
Wyszukaj
Ulubione
Google
Search
Adres
http://view.byu.edu/
Przejdź
Łącza
SEARCH STRING (HELP)
WORD/PHRASE
must
(INSERT TAG)
-select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select-
5
5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
OPTIONS (HELP)
HITS
100
SEE POS TAGS
NO
SEARCH
RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB
WORD/PHRASE
TOKENS REG1
PER MIL IN REG1 [10,638,034 WORDS]
1
MUST
4792
450.46
0.16 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Gotowe
Internet
Start
CL-ILS
[DAVIES] Variation in ...
Microsoft PowerPoint ...
PL
16:04

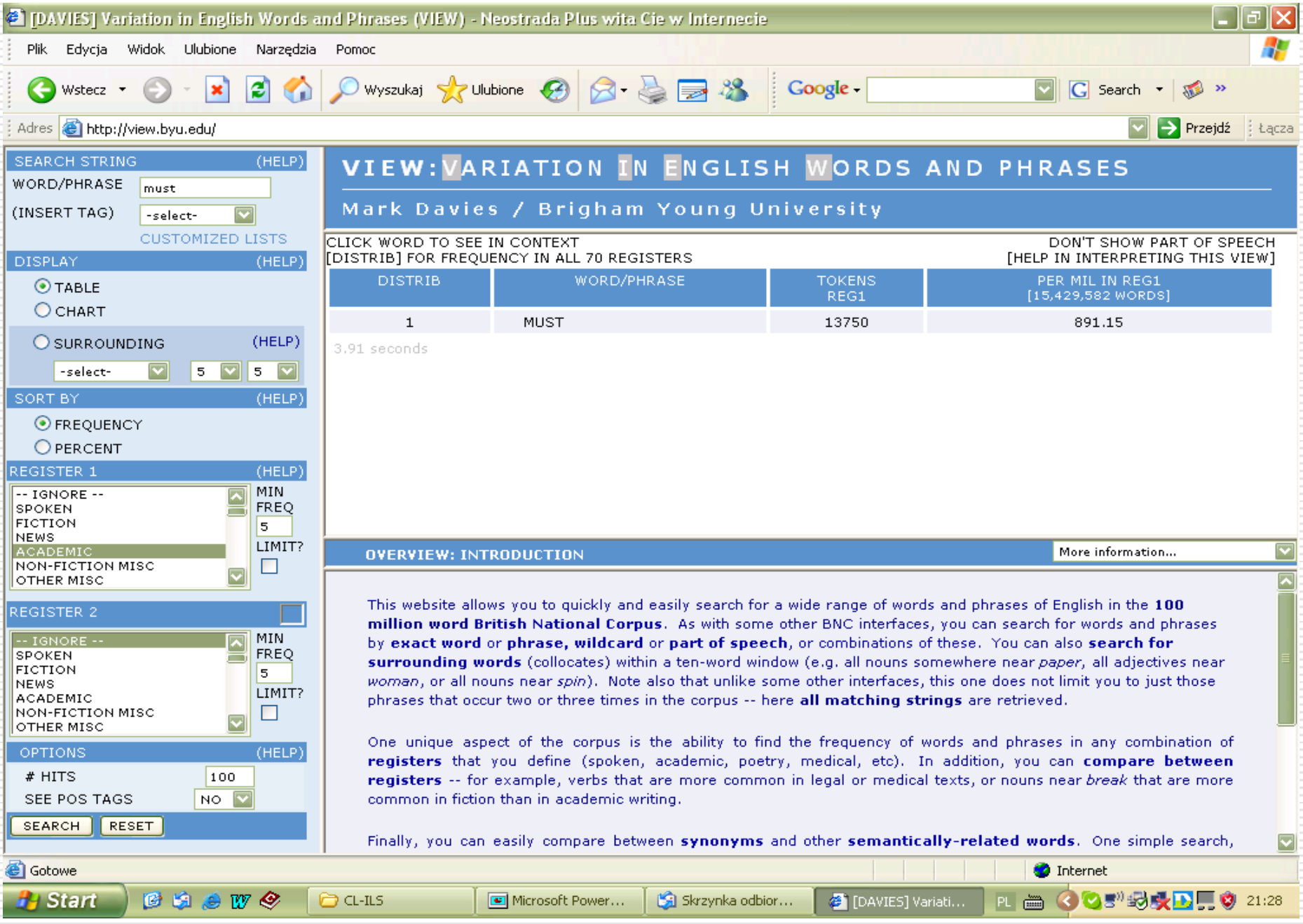

[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz
Wyszukaj
Ulubione
Google
Search
Adres
http://view.byu.edu/
Przejdź
Łącza
SEARCH STRING (HELP)
WORD/PHRASE
have/has to
(INSERT TAG)
-select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select-
5
5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
OPTIONS (HELP)
HITS
100
SEE POS TAGS
NO
SEARCH
RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB
WORD/PHRASE
TOKENS REG1
PER MIL IN REG1 [10,638,034 WORDS]
1
HAVE TO
4615
433.82
2
HAS TO
1052
98.89
TOTAL
5667
532.71
1.88 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Gotowe
Internet
Start
CL-ILS
[DAVIES] Variation in ...
Microsoft PowerPoint ...
PL
16:05

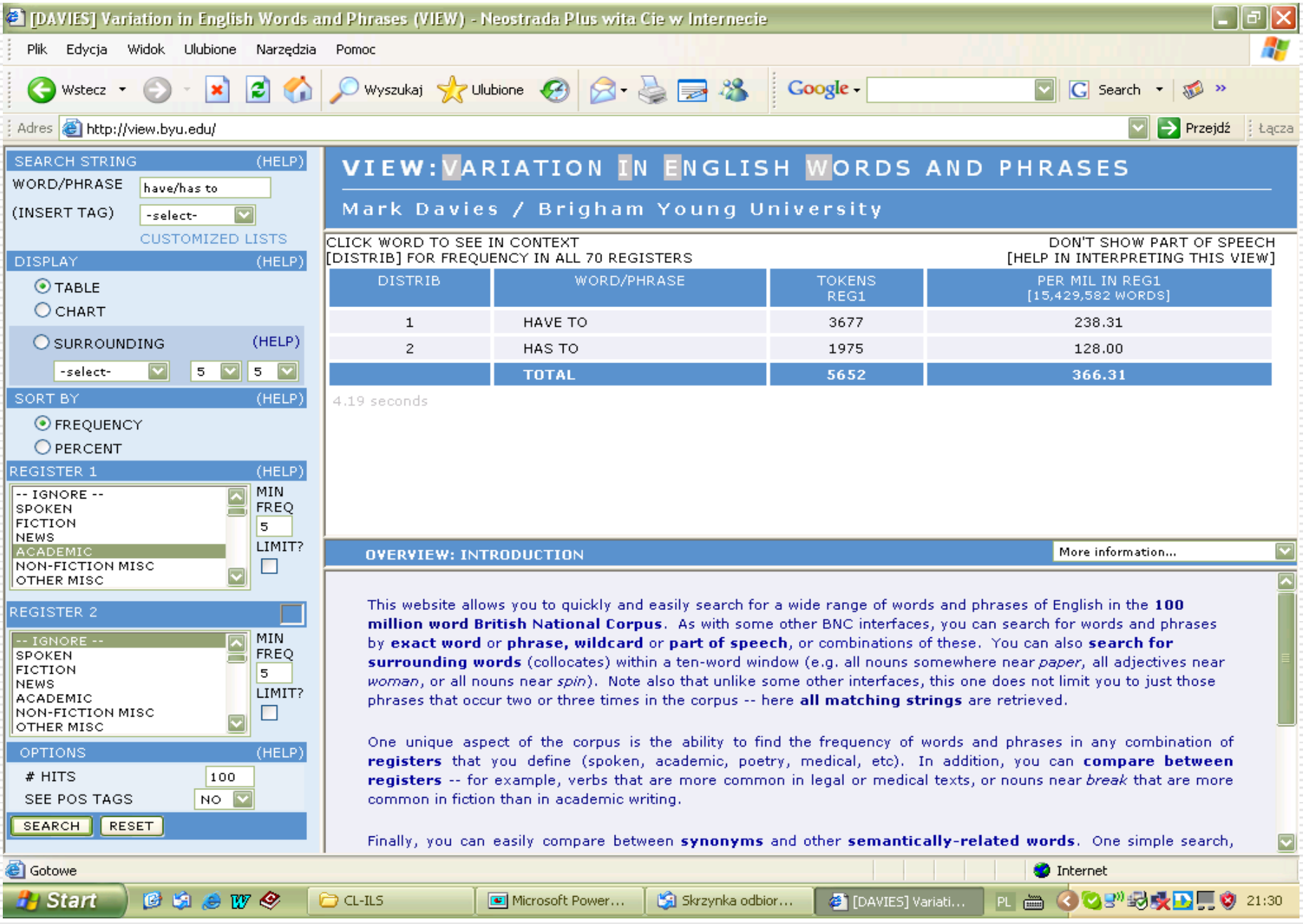
Must and *have to* in fiction


[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz
Wyszukaj
Ulubione
Google
Search
Adres
http://view.byu.edu/
Przejdź
Łącza
SEARCH STRING (HELP)
WORD/PHRASE
must
(INSERT TAG)
-select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select-
5
5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
OPTIONS (HELP)
HITS
100
SEE POS TAGS
NO
SEARCH
RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB
WORD/PHRASE
TOKENS REG1
PER MIL IN REG1 [16,194,885 WORDS]
1
MUST
15208
939.06
0.09 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Gotowe
Internet
Start
CL-ILS
[DAVIES] Variation in ...
Microsoft PowerPoint ...
PL
16:06


[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz
Wyszukaj
Ulubione
Google
Search
Adres
http://view.byu.edu/
Przejdź
Łącza
SEARCH STRING (HELP)
WORD/PHRASE
have/has to
(INSERT TAG)
-select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select-
5
5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
OPTIONS (HELP)
HITS
100
SEE POS TAGS
NO
SEARCH
RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB
WORD/PHRASE
TOKENS REG1
PER MIL IN REG1 [16,194,885 WORDS]
1
HAVE TO
8671
535.42
2
HAS TO
865
53.41
TOTAL
9536
588.83
1.92 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Gotowe
Internet
Start
CL-ILS
[DAVIES] Variation in ...
Microsoft PowerPoint ...
PL
16:07

Must and *have to* in academic English


[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz
Wyszukaj
Ulubione
Google
Search
Adres
http://view.byu.edu/
Przejdź
Łącza
SEARCH STRING (HELP)
WORD/PHRASE must
(INSERT TAG) -select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select- 5 5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ 5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ 5
LIMIT?
OPTIONS (HELP)
HITS 100
SEE POS TAGS NO
SEARCH
RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB
WORD/PHRASE
TOKENS REG1
PER MIL IN REG1 [15,429,582 WORDS]
1
MUST
13750
891.15
3.91 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Finally, you can easily compare between synonyms and other semantically-related words. One simple search,
Gotowe
Internet
Start
CL-ILS
Microsoft Power...
Skrzynka odbior...
[DAVIES] Variati...
PL
21:28


[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie
Plik Edycja Widok Ulubione Narzędzia Pomoc
Wstecz
Wyszukaj
Ulubione
Google
Search
Adres
http://view.byu.edu/
Przejdź
Łącza
SEARCH STRING (HELP)
WORD/PHRASE
have/has to
(INSERT TAG)
-select-
CUSTOMIZED LISTS
DISPLAY (HELP)
TABLE
CHART
SURROUNDING (HELP)
-select-
5
5
SORT BY (HELP)
FREQUENCY
PERCENT
REGISTER 1 (HELP)
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
REGISTER 2
-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC
MIN FREQ
5
LIMIT?
OPTIONS (HELP)
HITS
100
SEE POS TAGS
NO
SEARCH
RESET
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES
Mark Davies / Brigham Young University
CLICK WORD TO SEE IN CONTEXT
[DISTRIB] FOR FREQUENCY IN ALL 70 REGISTERS
DON'T SHOW PART OF SPEECH
[HELP IN INTERPRETING THIS VIEW]
DISTRIB
WORD/PHRASE
TOKENS REG1
PER MIL IN REG1 [15,429,582 WORDS]
1 HAVE TO 3677 238.31
2 HAS TO 1975 128.00
TOTAL 5652 366.31
4.19 seconds
OVERVIEW: INTRODUCTION
More information...
This website allows you to quickly and easily search for a wide range of words and phrases of English in the 100 million word British National Corpus. As with some other BNC interfaces, you can search for words and phrases by exact word or phrase, wildcard or part of speech, or combinations of these. You can also search for surrounding words (collocates) within a ten-word window (e.g. all nouns somewhere near paper, all adjectives near woman, or all nouns near spin). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here all matching strings are retrieved.
One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of registers that you define (spoken, academic, poetry, medical, etc). In addition, you can compare between registers -- for example, verbs that are more common in legal or medical texts, or nouns near break that are more common in fiction than in academic writing.
Finally, you can easily compare between synonyms and other semantically-related words. One simple search,
Gotowe
Internet
Start
CL-ILS
Microsoft Power...
Skrzynka odbior...
[DAVIES] Variati...
PL
21:30

must vs. have/has to across registers

	must	have/has to
general	701.58	530.97
spoken	590.33	1,187.52
news	450.46	532.71
fiction	939.06	588.83
academic	891.15	366.31

[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie

Plik Edycja Widok Ulubione Narzędzia Pomoc

Wstecz Wyszukaj Ulubione Google Search

Adres http://view.byu.edu/ Przejdź Łącza

SEARCH STRING (HELP)

WORD/PHRASE must

(INSERT TAG) -select-

CUSTOMIZED LISTS

DISPLAY (HELP)

TABLE

CHART

SURROUNDING (HELP)

-select- 5 5

SORT BY (HELP)

FREQUENCY

PERCENT

REGISTER 1 (HELP)

-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC

MIN FREQ 5

LIMIT?

REGISTER 2

-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC

MIN FREQ 5

LIMIT?

OPTIONS (HELP)

HITS 100

SEE POS TAGS NO

SEARCH RESET

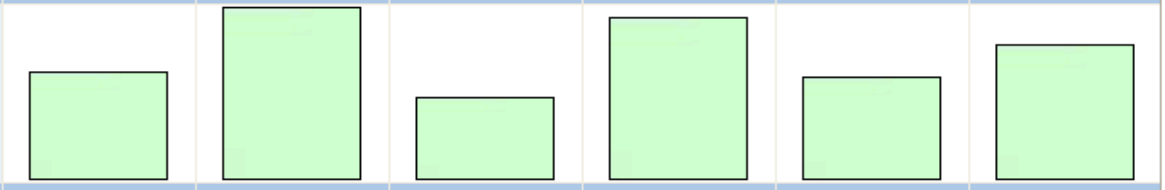
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES

Mark Davies / Brigham Young University

CLICK BARS IN CHART FOR FREQUENCIES FROM SUB-REGISTERS
CLICK REGISTER NAME FOR TABLE OF MATCHING STRINGS

HELP INTERPRETING THIS VIEW
NUMBER BELOW REGISTER FOR KWIC

REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHER MISC
	6101	15208	4792	13750	9331	20976
PER MILLION	590.3	939.1	450.5	891.1	561.0	738.7



OVERVIEW: INTRODUCTION

More information...

This website allows you to quickly and easily search for a wide range of words and phrases of English in the **100 million word British National Corpus**. As with some other BNC interfaces, you can search for words and phrases by **exact word** or **phrase, wildcard** or **part of speech**, or combinations of these. You can also **search for surrounding words** (collocates) within a ten-word window (e.g. all nouns somewhere near *paper*, all adjectives near *woman*, or all nouns near *spin*). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here **all matching strings** are retrieved.

One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of **registers** that you define (spoken, academic, poetry, medical, etc). In addition, you can **compare between registers** -- for example, verbs that are more common in legal or medical texts, or nouns near *break* that are more common in fiction than in academic writing.

Finally, you can easily compare between **synonyms** and other **semantically-related words**. One simple search,

Gotowe Internet

Start CL-ILS Microsoft Power... Skrzynka odbior... [DAVIES] Variati... PL 21:41

[DAVIES] Variation in English Words and Phrases (VIEW) - Neostrada Plus wita Cie w Internecie

Plik Edycja Widok Ulubione Narzędzia Pomoc

Wstecz Wyszukaj Ulubione Google Search

Adres http://view.byu.edu/ Przejdź Łącza

SEARCH STRING (HELP)

WORD/PHRASE have/has to

(INSERT TAG) -select-

CUSTOMIZED LISTS

DISPLAY (HELP)

- (•) TABLE
- ( ) CHART
- ( ) SURROUNDING (HELP)

-select- 5 5

SORT BY (HELP)

- (•) FREQUENCY
- ( ) PERCENT

REGISTER 1 (HELP)

-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC

MIN FREQ 5

LIMIT? [ ]

REGISTER 2 [ ]

-- IGNORE --
SPOKEN
FICTION
NEWS
ACADEMIC
NON-FICTION MISC
OTHER MISC

MIN FREQ 5

LIMIT? [ ]

OPTIONS (HELP)

HITS 100

SEE POS TAGS NO

SEARCH RESET

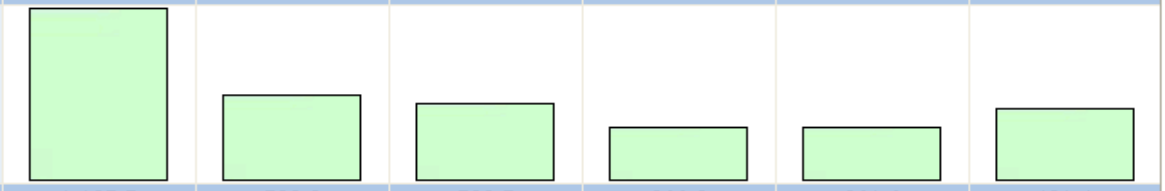
VIEW: VARIATION IN ENGLISH WORDS AND PHRASES

Mark Davies / Brigham Young University

CLICK BARS IN CHART FOR FREQUENCIES FROM SUB-REGISTERS
CLICK REGISTER NAME FOR TABLE OF MATCHING STRINGS

HELP INTERPRETING THIS VIEW

REGISTER	SPOKEN 12273	FICTION 9536	NEWS 5667	ACADEMIC 5652	NONFIC MISC 6015	OTHER MISC 13954
PER MILLION	1,187.5	588.8	532.7	366.3	361.6	491.4



OVERVIEW: INTRODUCTION

More information...

This website allows you to quickly and easily search for a wide range of words and phrases of English in the **100 million word British National Corpus**. As with some other BNC interfaces, you can search for words and phrases by **exact word** or **phrase, wildcard** or **part of speech**, or combinations of these. You can also **search for surrounding words** (collocates) within a ten-word window (e.g. all nouns somewhere near *paper*, all adjectives near *woman*, or all nouns near *spin*). Note also that unlike some other interfaces, this one does not limit you to just those phrases that occur two or three times in the corpus -- here **all matching strings** are retrieved.

One unique aspect of the corpus is the ability to find the frequency of words and phrases in any combination of **registers** that you define (spoken, academic, poetry, medical, etc). In addition, you can **compare between registers** -- for example, verbs that are more common in legal or medical texts, or nouns near *break* that are more common in fiction than in academic writing.

Finally, you can easily compare between **synonyms** and other **semantically-related words**. One simple search,

Internet

Start CL-ILS Microsoft Power... Skrzynka odbior... [DAVIES] Variati... PL 21:40

Example 2 – Word frequencies

- Work in groups and answer the following questions
  - What is the most frequent word in English?
  - Can you make the list of 10 most frequent words in English?
  - Can you guess what proportion of a text is made up of 10 most frequent words in English?
  - How many words do you need to know to read authentic texts for pleasure?

Word	PoS	Freq			
the	Det	61847	6,18%	6,18%	
of	Prep	29391	2,94%	9,12%	
and	Conj	26817	2,68%	11,81%	
a	Det	21626	2,16%	13,97%	
in	Prep	18214	1,82%	15,79%	
to	Inf	16284	1,63%	17,42%	
it	Pron	10875	1,09%	18,51%	
is	Verb	9982	1,00%	19,50%	
to	Prep	9343	0,93%	20,44%	
was	Verb	9236	0,92%	21,36%	
I	Pron	8875	0,89%	22,25%	
for	Prep	8412	0,84%	23,09%	
that	Conj	7308	0,73%	23,82%	
you	Pron	6954	0,70%	24,52%	
he	Pron	6810	0,68%	25,20%	
be*	Verb	6644	0,66%	25,86%	
with	Prep	6575	0,66%	26,52%	
on	Prep	6475	0,65%	27,17%	
by	Prep	5096	0,51%	27,68%	
at	Prep	4790	0,48%	28,16%	
have*	Verb	4735	0,47%	28,63%	
are	Verb	4707	0,47%	29,10%	
not	Neg	4626	0,46%	29,56%	
this	DetP	4623	0,46%	30,02%	

BNC Sampler

tokens (running words)	2,127,880
types (different words)	44,001
hapaxes (words occurring only once)	15,736
examples	auctioneer, breezy, grandparent

Frequency facts

Vocabulary size	**Text coverage**
10	23,7%
100	49,0%
2,000	81,3
3,000	85,2%
5,000	89,4%
43,831	99%
86,741	100%

(Carrol, Davis and Richman 1971 after Nation 1990)

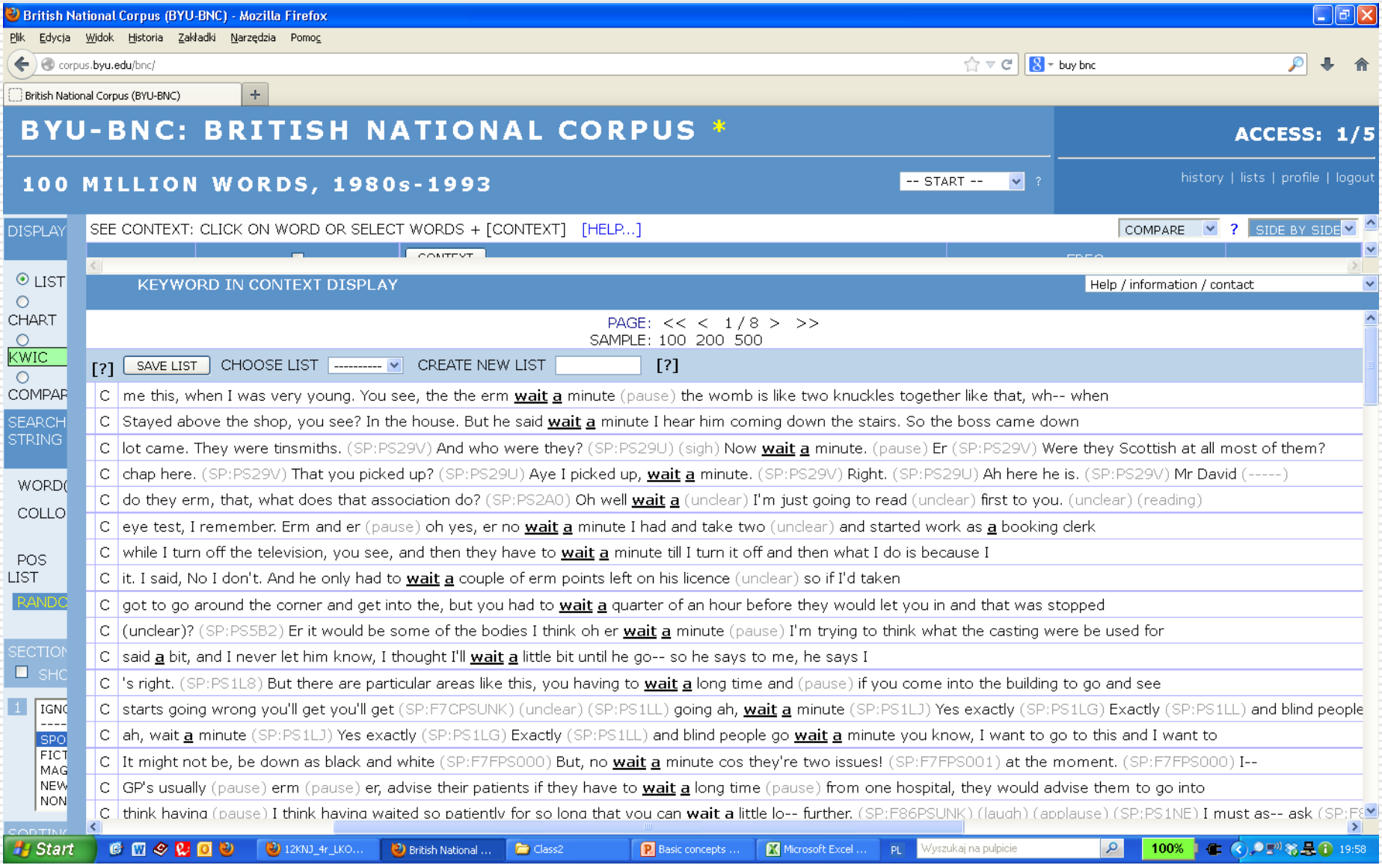
Minimum for language use

Skill	**Minimum vocabulary**
reading	3,000 word families (5,000 lexical items) for minimal comprehension 5,000 word families (8,000 lexical items) for reading for pleasure (Laufer 1992)
listening	half the words needed to understand written English (Nation 1990)
writing	2,000-3,000 (Nation 1990)
speaking	1,200 headwords (West 1960 after Nation 1990)

Activity 2

- What is the core meaning of the word *wait?*
- What are different meanings of this word?
- What are its most frequent collocations/colligations?
- Study the concordance lines in the next slide. Can you notice anything interesting about the meaning/fuction/collocations of this word that you did not think of before?
- Search the word in one of the online dictionaries (the link below). Does it list the particular meaning/function of *wait* you have just discovered?

  http://dictionary.reference.com/


British National Corpus (BYU-BNC) - Mozilla Firefox
Plik Edycja Widok Historia Zakładki Narzędzia Pomoc
corpus.byu.edu/bnc/
buy bnc
British National Corpus (BYU-BNC)
BYU-BNC: BRITISH NATIONAL CORPUS *
ACCESS: 1/5
100 MILLION WORDS, 1980s-1993
-- START --
history | lists | profile | logout
DISPLAY
SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]
COMPARE
SIDE BY SIDE
LIST
CHART
KWIC
COMPAR
SEARCH STRING
WORD(
COLLO
POS LIST
RANDC
SECTION
SHC
KEYWORD IN CONTEXT DISPLAY
Help / information / contact
PAGE: << < 1 / 8 > >>
SAMPLE: 100 200 500
[?] SAVE LIST CHOOSE LIST ---------- CREATE NEW LIST [?]
C me this, when I was very young. You see, the the erm wait a minute (pause) the womb is like two knuckles together like that, wh-- when
C Stayed above the shop, you see? In the house. But he said wait a minute I hear him coming down the stairs. So the boss came down
C lot came. They were tinsmiths. (SP:PS29V) And who were they? (SP:PS29U) (sigh) Now wait a minute. (pause) Er (SP:PS29V) Were they Scottish at all most of them?
C chap here. (SP:PS29V) That you picked up? (SP:PS29U) Aye I picked up, wait a minute. (SP:PS29V) Right. (SP:PS29U) Ah here he is. (SP:PS29V) Mr David (-----)
C do they erm, that, what does that association do? (SP:PS2A0) Oh well wait a (unclear) I'm just going to read (unclear) first to you. (unclear) (reading)
C eye test, I remember. Erm and er (pause) oh yes, er no wait a minute I had and take two (unclear) and started work as a booking clerk
C while I turn off the television, you see, and then they have to wait a minute till I turn it off and then what I do is because I
C it. I said, No I don't. And he only had to wait a couple of erm points left on his licence (unclear) so if I'd taken
C got to go around the corner and get into the, but you had to wait a quarter of an hour before they would let you in and that was stopped
C (unclear)? (SP:PS5B2) Er it would be some of the bodies I think oh er wait a minute (pause) I'm trying to think what the casting were be used for
C said a bit, and I never let him know, I thought I'll wait a little bit until he go-- so he says to me, he says I
C 's right. (SP:PS1L8) But there are particular areas like this, you having to wait a long time and (pause) if you come into the building to go and see
C starts going wrong you'll get you'll get (SP:F7CPSUNK) (unclear) (SP:PS1LL) going ah, wait a minute (SP:PS1LJ) Yes exactly (SP:PS1LG) Exactly (SP:PS1LL) and blind people
C ah, wait a minute (SP:PS1LJ) Yes exactly (SP:PS1LG) Exactly (SP:PS1LL) and blind people go wait a minute you know, I want to go to this and I want to
C It might not be, be down as black and white (SP:F7FPS000) But, no wait a minute cos they're two issues! (SP:F7FPS001) at the moment. (SP:F7FPS000) I--
C GP's usually (pause) erm (pause) er, advise their patients if they have to wait a long time (pause) from one hospital, they would advise them to go into
C think having (pause) I think having waited so patiently for so long that you can wait a little lo-- further. (SP:F86PSUNK) (laugh) (applause) (SP:PS1NE) I must as-- ask (SP:F
Start
12KNJ_4r_LKO...
British National ...
Class2
Basic concepts ...
Microsoft Excel ...
PL
Wyszukaj na pulpicie
100%
19:58