## Comparison of N-Gram[1] Rank Frequency Data from the Written Texts of the British National Corpus World Edition (BNC) and the author's Web Corpus

Both sets of texts were preprocessed to provide comparable data.

**British National Corpus.** For simplicity's sake, all text within <body> tags in the BNC data files was assumed to be written text. It was extracted from the files, then all SGML tags including grammatical markup were stripped from it and SMGL entities were expanded to the corresponding characters. Spaces around orthographic word-interior hyphen and apostrophe were removed. The resulting text data were amalgamated into nine large data files with 87,221,955 tokens total for further processing.

**Web Corpus.** The 4949 documents which survived the various "filters" and selection processes elaborated elsewhere were combined into a single file with 5,382,595 tokens total to facilitate processing (approximately 1/16 of the size of the BNC written corpus).

**Both Corpora.** With the author's software kfNgram[2] frequency lists of 1-, 2-, 3-, 4-, 5-, 25-, and 50-grams in the two corpora were produced. Relevant options chosen were: not case-sensitive, preserve word-interior hyphens and apostrophes, replace numerals with #, sort on the first 2048 characters of each n-gram, floor (i.e. minimum frequency for inclusion in the final list) 50. Standard kfNgram character remapping was chosen, so boundaries between sentences, paragraphs and even entire texts were ignored on the reasonable assumption that the random pseudo-n-grams resulting from this expediency would fall below the relatively high threshold chosen.

The 5000 most frequent alphabetic[3] n-grams for each value of *n* were then imported into a Microsoft Access database for further analysis. Three sets of queries yielded the record sets reproduced below (since these are frequency *ranks, lower* numbers reflect *higher* frequencies):

1. N-grams with a rank frequency of 1 to 250 in both corpora
2. N-grams with a rank frequency of 1 to 200 in one corpus and greater than 300 (i.e. relatively less frequent) in the other
3. N-grams with a rank frequency of 1 to 500 in one corpus not among the 5000 most frequent in the other.

**Future.** In this study no attempt was made to normalize spelling (e.g. *-ize / -ise, -er / -re, -or / -our,* as well as tendencies to write certain compound nouns with a hyphen, a space or without a space) between the predominantly British BNC texts and the American-biased Web documents, a planned refinement that will enhance comparability in the future.

---

[1] The term *n*-gram is understood here as a sequence of *n* **words**, also variously known as *cluster, lexical bundle,* or *chain*. In other fields this term is also used to mean sequence of *n* **characters**.

[2] Information on kfNgram and free download are available at http://kwicfinder.com/kfNgram/. Other tools used for this analysis either are already or will soon be available there as well.

[3] Since in a preliminary analysis the great variety of numeric data obscured the patterns of linguistic data, all n-grams containing numerals were excluded from this analysis.

**Similarities between British National Corpus and Web Corpus Data**

Dataset 1:

50 most frequent n-grams from those with a rank frequency of 1 to 250 in both corpora, ordered by rank frequency in BNC and in WC respectively

| Top 50 Shared Frequent 1-grams BNC Order | | |
|---|---|---|
| 1-Gram | BNC Rank | Web Rank |
| the | 1 | 1 |
| of | 2 | 2 |
| to | 3 | 4 |
| and | 4 | 3 |
| a | 5 | 5 |
| in | 6 | 6 |
| is | 7 | 7 |
| that | 8 | 9 |
| for | 9 | 8 |
| was | 10 | 24 |
| it | 11 | 16 |
| on | 12 | 10 |
| with | 13 | 11 |
| as | 14 | 17 |
| be | 15 | 12 |
| he | 16 | 39 |
| by | 17 | 19 |
| i | 18 | 20 |
| at | 19 | 21 |
| are | 20 | 14 |
| not | 21 | 25 |
| from | 22 | 22 |
| his | 23 | 44 |
| this | 24 | 13 |
| had | 25 | 68 |
| but | 26 | 33 |
| have | 27 | 23 |
| you | 28 | 15 |
| which | 29 | 35 |
| or | 30 | 18 |
| an | 31 | 26 |
| they | 32 | 36 |
| she | 33 | 155 |
| her | 34 | 130 |
| were | 35 | 53 |
| their | 36 | 41 |
| been | 37 | 55 |
| has | 38 | 34 |
| all | 39 | 30 |
| one | 40 | 38 |
| will | 41 | 27 |
| there | 42 | 47 |
| we | 43 | 28 |
| if | 44 | 32 |
| would | 45 | 62 |
| can | 46 | 31 |
| more | 47 | 37 |
| when | 48 | 49 |
| so | 49 | 61 |
| who | 50 | 54 |

| Top 50 Shared Frequent 1-grams Web Order | | |
|---|---|---|
| 1-Gram | BNC Rank | Web Rank |
| the | 1 | 1 |
| of | 2 | 2 |
| and | 4 | 3 |
| to | 3 | 4 |
| a | 5 | 5 |
| in | 6 | 6 |
| is | 7 | 7 |
| for | 9 | 8 |
| that | 8 | 9 |
| on | 12 | 10 |
| with | 13 | 11 |
| be | 15 | 12 |
| this | 24 | 13 |
| are | 20 | 14 |
| you | 28 | 15 |
| it | 11 | 16 |
| as | 14 | 17 |
| or | 30 | 18 |
| by | 17 | 19 |
| i | 18 | 20 |
| at | 19 | 21 |
| from | 22 | 22 |
| have | 27 | 23 |
| was | 10 | 24 |
| not | 21 | 25 |
| an | 31 | 26 |
| will | 41 | 27 |
| we | 43 | 28 |
| your | 81 | 29 |
| all | 39 | 30 |
| can | 46 | 31 |
| if | 44 | 32 |
| but | 26 | 33 |
| has | 38 | 34 |
| which | 29 | 35 |
| they | 32 | 36 |
| more | 47 | 37 |
| one | 40 | 38 |
| he | 16 | 39 |
| about | 57 | 40 |
| their | 36 | 41 |
| new | 74 | 42 |
| other | 61 | 43 |
| his | 23 | 44 |
| information | 206 | 45 |
| our | 100 | 46 |
| there | 42 | 47 |
| may | 69 | 48 |
| when | 48 | 49 |
| these | 76 | 50 |

| Top 50 Shared Frequent 2-grams BNC Order | | |
|---|---|---|
| 2-Gram | BNC Rank | Web Rank |
| of the | 1 | 1 |
| in the | 2 | 2 |
| to the | 3 | 3 |
| on the | 4 | 4 |
| and the | 5 | 6 |
| to be | 6 | 7 |
| for the | 7 | 5 |
| at the | 8 | 8 |
| it is | 9 | 12 |
| by the | 10 | 11 |
| that the | 11 | 14 |
| with the | 12 | 9 |
| of a | 13 | 15 |
| from the | 14 | 10 |
| it was | 15 | 39 |
| in a | 16 | 16 |
| as a | 17 | 18 |
| with a | 18 | 23 |
| is a | 19 | 13 |
| he was | 20 | 124 |
| for a | 21 | 22 |
| have been | 22 | 34 |
| will be | 23 | 17 |
| had been | 24 | 230 |
| as the | 25 | 26 |
| was a | 26 | 92 |
| is the | 27 | 19 |
| the first | 28 | 27 |
| one of | 29 | 31 |
| to a | 30 | 25 |
| has been | 31 | 30 |
| there is | 32 | 35 |
| and a | 33 | 37 |
| can be | 34 | 21 |
| the same | 35 | 32 |
| into the | 36 | 57 |
| would be | 37 | 60 |
| of his | 38 | 117 |
| is not | 40 | 33 |
| out of | 42 | 87 |
| on a | 43 | 41 |
| this is | 44 | 28 |
| all the | 45 | 54 |
| was the | 46 | 134 |
| to have | 47 | 96 |
| may be | 49 | 38 |
| over the | 50 | 78 |
| they are | 51 | 52 |
| there are | 52 | 42 |
| that it | 53 | 108 |

| Top 50 Shared Frequent 2-grams Web Order | | |
|---|---|---|
| 2-Gram | BNC Rank | Web Rank |
| of the | 1 | 1 |
| in the | 2 | 2 |
| to the | 3 | 3 |
| on the | 4 | 4 |
| for the | 7 | 5 |
| and the | 5 | 6 |
| to be | 6 | 7 |
| at the | 8 | 8 |
| with the | 12 | 9 |
| from the | 14 | 10 |
| by the | 10 | 11 |
| it is | 9 | 12 |
| is a | 19 | 13 |
| that the | 11 | 14 |
| of a | 13 | 15 |
| in a | 16 | 16 |
| will be | 23 | 17 |
| as a | 17 | 18 |
| is the | 27 | 19 |
| if you | 75 | 20 |
| can be | 34 | 21 |
| for a | 21 | 22 |
| with a | 18 | 23 |
| you can | 134 | 24 |
| to a | 30 | 25 |
| as the | 25 | 26 |
| the first | 28 | 27 |
| this is | 44 | 28 |
| of this | 63 | 29 |
| has been | 31 | 30 |
| one of | 29 | 31 |
| the same | 35 | 32 |
| is not | 40 | 33 |
| have been | 22 | 34 |
| there is | 32 | 35 |
| in this | 62 | 36 |
| and a | 33 | 37 |
| may be | 49 | 38 |
| it was | 15 | 39 |
| such as | 71 | 40 |
| on a | 43 | 41 |
| there are | 52 | 42 |
| number of | 65 | 43 |
| you are | 161 | 44 |
| about the | 64 | 45 |
| should be | 54 | 46 |
| the world | 118 | 47 |
| as well | 102 | 48 |
| the most | 66 | 49 |
| you have | 188 | 50 |

| Top 50 Shared Frequent 3-grams BNC Order | | |
|---|---|---|
| 3-Gram | BNC Rank | Web Rank |
| one of the | 1 | 1 |
| the end of | 2 | 14 |
| as well as | 3 | 2 |
| part of the | 4 | 5 |
| out of the | 5 | 32 |
| a number of | 6 | 8 |
| there is a | 7 | 11 |
| some of the | 8 | 7 |
| there was a | 9 | 164 |
| end of the | 10 | 34 |
| the fact that | 12 | 36 |
| it was a | 13 | 124 |
| in order to | 14 | 6 |
| there is no | 15 | 18 |
| be able to | 16 | 10 |
| to be a | 17 | 24 |
| it is not | 18 | 25 |
| it would be | 19 | 76 |
| the number of | 20 | 17 |
| at the end | 21 | 56 |
| it is a | 22 | 33 |
| a lot of | 23 | 9 |
| the use of | 24 | 13 |
| in terms of | 25 | 44 |
| that it is | 27 | 47 |
| most of the | 28 | 31 |
| on the other | 29 | 74 |
| the rest of | 30 | 65 |
| members of the | 31 | 42 |
| as a result | 32 | 39 |
| in the first | 33 | 50 |
| at the same | 34 | 45 |
| at the time | 35 | 53 |
| the first time | 37 | 118 |
| for the first | 40 | 84 |
| the same time | 41 | 59 |
| to be the | 42 | 68 |
| have to be | 43 | 91 |
| because of the | 44 | 67 |
| and in the | 45 | 102 |
| in the same | 46 | 96 |
| would have been | 47 | 230 |
| in which the | 48 | 85 |
| it is the | 49 | 66 |
| side of the | 50 | 137 |
| the united states | 51 | 4 |
| this is a | 53 | 19 |
| a series of | 54 | 107 |
| is to be | 55 | 163 |
| many of the | 56 | 70 |

| Top 50 Shared Frequent 3-grams Web Order | | |
|---|---|---|
| 3-Gram | BNC Rank | Web Rank |
| one of the | 1 | 1 |
| as well as | 3 | 2 |
| the united states | 51 | 4 |
| part of the | 4 | 5 |
| in order to | 14 | 6 |
| some of the | 8 | 7 |
| a number of | 6 | 8 |
| a lot of | 23 | 9 |
| be able to | 16 | 10 |
| there is a | 7 | 11 |
| the use of | 24 | 13 |
| the end of | 2 | 14 |
| if you are | 130 | 16 |
| the number of | 20 | 17 |
| there is no | 15 | 18 |
| this is a | 53 | 19 |
| to be a | 17 | 24 |
| it is not | 18 | 25 |
| would like to | 196 | 28 |
| this is the | 70 | 29 |
| a variety of | 114 | 30 |
| most of the | 28 | 31 |
| out of the | 5 | 32 |
| it is a | 22 | 33 |
| end of the | 10 | 34 |
| in the world | 65 | 35 |
| the fact that | 12 | 36 |
| in addition to | 181 | 37 |
| back to the | 58 | 38 |
| as a result | 32 | 39 |
| of the most | 62 | 40 |
| the development of | 57 | 41 |
| members of the | 31 | 42 |
| in terms of | 25 | 44 |
| at the same | 34 | 45 |
| in the united | 190 | 46 |
| that it is | 27 | 47 |
| the department of | 249 | 49 |
| in the first | 33 | 50 |
| of the world | 103 | 51 |
| at the time | 35 | 53 |
| is one of | 91 | 54 |
| as part of | 101 | 55 |
| at the end | 21 | 56 |
| based on the | 205 | 57 |
| member of the | 98 | 58 |
| the same time | 41 | 59 |
| according to the | 102 | 61 |
| the rest of | 30 | 65 |
| it is the | 49 | 66 |

| Top 50 Frequent Shared 4-grams BNC Order | | |
|---|---|---|
| 4-Gram | BNC Rank | Web Rank |
| the end of the | 1 | 1 |
| at the end of | 2 | 4 |
| at the same time | 3 | 3 |
| for the first time | 4 | 21 |
| on the other hand | 5 | 14 |
| as a result of | 6 | 7 |
| the rest of the | 7 | 13 |
| in the case of | 9 | 19 |
| one of the most | 10 | 10 |
| by the end of | 12 | 53 |
| is one of the | 13 | 8 |
| on the basis of | 14 | 26 |
| at the time of | 15 | 18 |
| in the form of | 16 | 31 |
| as well as the | 17 | 9 |
| to be able to | 18 | 22 |
| in the middle of | 19 | 67 |
| the top of the | 20 | 32 |
| a member of the | 21 | 17 |
| the fact that the | 22 | 56 |
| a wide range of | 24 | 23 |
| the nature of the | 25 | 50 |
| was one of the | 26 | 60 |
| a great deal of | 27 | 48 |
| at the beginning of | 28 | 40 |
| will be able to | 29 | 16 |
| the middle of the | 30 | 74 |
| it is possible to | 31 | 90 |
| as part of the | 34 | 24 |
| in the context of | 35 | 61 |
| in the united states | 37 | 2 |
| in the course of | 38 | 166 |
| is likely to be | 39 | 198 |
| in terms of the | 40 | 128 |
| it is important to | 41 | 35 |
| a result of the | 42 | 62 |
| in front of the | 47 | 69 |
| in the first place | 48 | 47 |
| the beginning of the | 49 | 45 |
| the time of the | 51 | 70 |
| from time to time | 52 | 27 |
| in the absence of | 53 | 123 |
| it is difficult to | 54 | 184 |
| on the part of | 58 | 107 |
| at the top of | 60 | 75 |
| in the face of | 62 | 206 |
| that there is a | 63 | 112 |
| the size of the | 64 | 51 |
| for a long time | 68 | 116 |
| the role of the | 72 | 77 |

| Top 50 Frequent Shared 4-grams Web Order | | |
|---|---|---|
| 4-Gram | BNC Rank | Web Rank |
| the end of the | 1 | 1 |
| in the united states | 37 | 2 |
| at the same time | 3 | 3 |
| at the end of | 2 | 4 |
| at the university of | 220 | 5 |
| if you want to | 93 | 6 |
| as a result of | 6 | 7 |
| is one of the | 13 | 8 |
| as well as the | 17 | 9 |
| one of the most | 10 | 10 |
| the rest of the | 7 | 13 |
| on the other hand | 5 | 14 |
| will be able to | 29 | 16 |
| a member of the | 21 | 17 |
| at the time of | 15 | 18 |
| in the case of | 9 | 19 |
| for the first time | 4 | 21 |
| to be able to | 18 | 22 |
| a wide range of | 24 | 23 |
| as part of the | 34 | 24 |
| on the basis of | 14 | 26 |
| from time to time | 52 | 27 |
| the name of the | 101 | 28 |
| in addition to the | 94 | 30 |
| in the form of | 16 | 31 |
| the top of the | 20 | 32 |
| i would like to | 113 | 33 |
| can be used to | 127 | 34 |
| it is important to | 41 | 35 |
| in accordance with the | 142 | 38 |
| for the purpose of | 186 | 39 |
| at the beginning of | 28 | 40 |
| the bottom of the | 81 | 41 |
| a copy of the | 172 | 43 |
| the beginning of the | 49 | 45 |
| the use of the | 79 | 46 |
| in the first place | 48 | 47 |
| a great deal of | 27 | 48 |
| the nature of the | 25 | 50 |
| the size of the | 64 | 51 |
| by the end of | 12 | 53 |
| the fact that the | 22 | 56 |
| one of the best | 158 | 57 |
| as well as a | 108 | 58 |
| was one of the | 26 | 60 |
| in the context of | 35 | 61 |
| a result of the | 42 | 62 |
| a wide variety of | 192 | 64 |
| a large number of | 83 | 66 |
| in the middle of | 19 | 67 |

## Differences between British National Corpus and Web Corpus Data

Dataset 2:

N-grams with a rank frequency of 1 to 200 in one corpus and greater than 300 (i.e. relatively less frequent) in the other

| 200 most frequent 1-grams in BNC data Ranked 300 or lower in Web data | | |
|---|---|---|
| 1Gram | BNC Rank | Web Rank |
| mr | 123 | 371 |
| per | 139 | 304 |
| man | 146 | 414 |
| again | 151 | 346 |
| old | 153 | 319 |
| never | 155 | 331 |
| thought | 160 | 729 |
| away | 170 | 513 |
| left | 175 | 391 |
| although | 178 | 394 |
| going | 179 | 323 |
| put | 181 | 349 |
| social | 182 | 488 |
| came | 184 | 566 |
| though | 186 | 507 |
| rather | 189 | 499 |
| always | 190 | 358 |
| got | 191 | 508 |
| until | 195 | 357 |
| cent | 196 | 1648 |

| 200 most frequent 1-grams in Web data Ranked 300 or lower in BNC data | | |
|---|---|---|
| 1gram | BNC Rank | Web Rank |
| available | 301 | 123 |
| c | 451 | 192 |
| computer | 708 | 182 |
| contact | 933 | 194 |
| data | 490 | 120 |
| file | 1773 | 186 |
| free | 433 | 146 |
| including | 327 | 196 |
| links | 2274 | 187 |
| list | 808 | 167 |
| name | 363 | 160 |
| news | 688 | 139 |
| page | 1011 | 70 |
| please | 924 | 184 |
| program | 2380 | 113 |
| project | 638 | 199 |
| s | 755 | 89 |
| search | 1367 | 135 |
| services | 332 | 111 |
| site | 1054 | 67 |
| software | 1018 | 173 |
| students | 642 | 147 |
| systems | 521 | 193 |
| university | 586 | 114 |
| using | 344 | 150 |

| 200 most frequent 2-grams in BNC data Ranked 300 or lower in Web data | | |
|---|---|---|
| 2-gram | BNC Rank | Web Rank |
| against the | 195 | 501 |
| and he | 143 | 560 |
| as he | 170 | 945 |
| can not | 135 | 4008 |
| had to | 117 | 385 |
| he had | 39 | 446 |
| he said | 74 | 384 |
| i don't | 155 | 330 |
| i had | 181 | 389 |
| it would | 127 | 303 |
| like a | 171 | 465 |
| of her | 131 | 817 |
| per cent | 48 | 623 |
| she had | 83 | 2699 |
| she was | 55 | 739 |
| that there | 187 | 327 |
| the government | 159 | 305 |
| the house | 179 | 520 |
| the whole | 192 | 403 |
| there were | 148 | 416 |
| they had | 186 | 693 |
| this was | 191 | 515 |
| to her | 200 | 1752 |
| to his | 175 | 460 |
| was to | 184 | 707 |
| when he | 153 | 474 |
| who had | 193 | 1026 |

| 200 most frequent 2-grams in Web data Ranked 300 or lower in BNC data | | |
|---|---|---|
| 2-gram | BNC Rank | Web Rank |
| all of | 459 | 112 |
| based on | 326 | 133 |
| be used | 363 | 163 |
| department of | 987 | 152 |
| for more | 1317 | 162 |
| go to | 480 | 193 |
| have the | 312 | 158 |
| how to | 475 | 76 |
| in addition | 542 | 178 |
| information about | 2183 | 195 |
| information on | 2314 | 153 |
| like to | 424 | 197 |
| list of | 1248 | 190 |
| new york | 966 | 122 |
| of your | 432 | 137 |
| on this | 584 | 141 |
| return to | 833 | 157 |
| that are | 753 | 146 |
| the state | 468 | 167 |
| the united | 306 | 119 |
| the university | 1123 | 116 |
| to help | 342 | 184 |
| to provide | 394 | 168 |
| to this | 404 | 183 |
| to use | 321 | 103 |
| united states | 640 | 98 |
| university of | 2233 | 86 |
| use the | 796 | 175 |
| when you | 514 | 200 |
| you may | 898 | 181 |
| you to | 492 | 155 |
| you will | 337 | 89 |

| 200 most frequent 3-grams in BNC data Ranked 300 or lower in Web data | | |
|---|---|---|
| 3-gram | BNC Rank | Web Rank |
| a great deal | 154 | 371 |
| a matter of | 113 | 338 |
| a range of | 187 | 311 |
| and it was | 84 | 385 |
| and so on | 127 | 367 |
| and that the | 182 | 372 |
| as a whole | 165 | 632 |
| away from the | 121 | 657 |
| between the two | 185 | 569 |
| but it was | 80 | 635 |
| by the end | 179 | 745 |
| by the time | 192 | 827 |
| cent of the | 93 | 2631 |
| for a moment | 174 | 3848 |
| had to be | 69 | 532 |
| has to be | 119 | 457 |
| he did not | 169 | 1117 |
| he had been | 77 | 2034 |
| he was a | 106 | 869 |
| in favour of | 197 | 2865 |
| in relation to | 108 | 347 |
| in the middle | 116 | 348 |
| in the uk | 88 | 588 |
| is likely to | 115 | 411 |
| it had been | 117 | 1866 |
| it is possible | 184 | 320 |
| it was not | 71 | 555 |
| it was the | 38 | 343 |
| of the house | 161 | 678 |
| on to the | 76 | 836 |
| per cent of | 11 | 638 |
| secretary of state | 100 | 1880 |
| so that the | 157 | 363 |
| that he had | 82 | 1271 |
| that he was | 52 | 470 |
| that of the | 180 | 612 |
| that there was | 158 | 980 |
| the age of | 152 | 484 |
| the back of | 124 | 1505 |
| the centre of | 164 | 4926 |
| the effect of | 156 | 650 |
| the house of | 122 | 860 |
| the idea of | 120 | 595 |
| the level of | 188 | 321 |
| the middle of | 163 | 472 |
| the prime minister | 155 | 3177 |
| the secretary of | 134 | 1726 |
| the united kingdom | 159 | 1101 |
| the work of | 168 | 314 |
| to have been | 60 | 462 |
| to say that | 183 | 536 |
| was going to | 149 | 989 |
| was in the | 177 | 945 |
| was one of | 133 | 369 |

| 200 most frequent 3-grams in BNC data Ranked 300 or lower in Web data | | |
|---|---|---|
| 3-gram | BNC Rank | Web Rank |
| was to be | 128 | 2394 |
| would be a | 172 | 376 |
| would have to | 138 | 364 |
| would not be | 198 | 687 |

| 200 most frequent 3-grams in Web data Ranked 300 or lower in BNC data | | |
|---|---|---|
| 3-gram | BNC Rank | Web Rank |
| a copy of | 762 | 161 |
| a list of | 626 | 60 |
| a part of | 760 | 182 |
| a set of | 413 | 193 |
| all of the | 565 | 20 |
| around the world | 1725 | 125 |
| at the university | 1605 | 71 |
| can be found | 1252 | 108 |
| due to the | 312 | 126 |
| each of the | 373 | 99 |
| go to the | 389 | 112 |
| his or her | 623 | 148 |
| history of the | 743 | 175 |
| if you want | 546 | 48 |
| if you would | 4128 | 194 |
| in accordance with | 496 | 90 |
| in new york | 710 | 183 |
| in the future | 412 | 122 |
| information about the | 1457 | 113 |
| information on the | 2488 | 116 |
| more than one | 445 | 173 |
| name of the | 1099 | 184 |
| of the united | 661 | 81 |
| of the university | 1992 | 178 |
| return to the | 677 | 117 |
| that can be | 548 | 88 |
| the ability to | 586 | 100 |
| the history of | 316 | 101 |
| the office of | 2187 | 191 |
| the purpose of | 398 | 95 |
| the state of | 518 | 179 |
| the university of | 468 | 12 |
| there are many | 444 | 180 |
| to create a | 742 | 199 |
| to find out | 395 | 177 |
| to the top | 1281 | 200 |
| to use the | 354 | 110 |
| version of the | 666 | 156 |
| what is the | 724 | 136 |
| with respect to | 1127 | 130 |
| you do not | 1213 | 167 |
| you have a | 556 | 83 |
| you want to | 4843 | 22 |

| 200 most frequent 4-grams in BNC data Ranked 300 or lower in Web data | | |
|---|---|---|
| 4-gram | BNC Rank | Web Rank |
| the way in which | 23 | 650 |
| the back of the | 33 | 444 |
| in the light of | 45 | 411 |
| in the same way | 46 | 366 |
| at the age of | 67 | 398 |
| in an attempt to | 69 | 324 |
| on the one hand | 70 | 532 |
| at a time when | 71 | 727 |
| the other side of | 73 | 369 |
| the side of the | 76 | 419 |
| there has been a | 77 | 353 |
| on the other side | 78 | 327 |
| it is clear that | 80 | 466 |
| i don't want to | 82 | 999 |
| for the rest of | 85 | 306 |
| it would have been | 90 | 413 |
| nothing to do with | 97 | 884 |
| are likely to be | 104 | 304 |
| other side of the | 106 | 839 |
| the case of the | 109 | 417 |
| on the edge of | 110 | 1014 |
| by the fact that | 112 | 671 |
| in a number of | 115 | 307 |
| that it would be | 116 | 506 |
| it would be a | 118 | 832 |
| at the expense of | 120 | 811 |
| at the start of | 123 | 918 |
| in the hands of | 125 | 387 |
| turned out to be | 126 | 424 |
| as a means of | 129 | 565 |
| the heart of the | 145 | 330 |
| of a number of | 150 | 440 |
| as far as the | 152 | 982 |
| to deal with the | 156 | 423 |
| that it is not | 166 | 348 |
| one of the few | 167 | 503 |
| of the fact that | 171 | 501 |
| on top of the | 174 | 762 |
| the same time the | 179 | 352 |
| in spite of the | 185 | 435 |
| for the sake of | 188 | 584 |
| the creation of a | 189 | 474 |
| but it is not | 191 | 816 |

| 200 most frequent 4-grams in Web data Ranked 300 or lower in BNC data | | |
|---|---|---|
| 4-gram | BNC Rank | Web Rank |
| in an effort to | 314 | 183 |
| all over the world | 316 | 83 |
| of the united states | 322 | 11 |
| to the top of | 324 | 82 |
| can be found in | 334 | 65 |
| is part of the | 336 | 105 |
| some of the most | 341 | 150 |
| in the field of | 357 | 190 |
| to make sure that | 359 | 172 |
| is based on the | 396 | 149 |
| the purpose of the | 401 | 152 |
| the results of the | 411 | 154 |
| if you have a | 421 | 36 |
| the best way to | 423 | 81 |
| in the area of | 491 | 159 |
| to return to the | 505 | 79 |
| the president of the | 534 | 193 |
| to the united states | 596 | 173 |
| the story of the | 612 | 163 |
| have the right to | 615 | 176 |
| a few of the | 712 | 125 |
| in the near future | 716 | 197 |
| if you wish to | 934 | 85 |
| of the university of | 985 | 49 |
| if you do not | 1069 | 55 |
| of the department of | 1153 | 169 |
| the purpose of this | 1205 | 162 |
| the united states the | 1310 | 153 |
| if you are a | 1689 | 106 |
| you don't have to | 1623 | 99 |
| take a look at | 2040 | 44 |
| the office of the | 2088 | 179 |
| a look at the | 2099 | 141 |
| if you would like | 2264 | 25 |
| having regard to the | 2372 | 148 |
| if you have any | 2518 | 29 |
| if you are not | 2313 | 127 |
| for a period of | 2438 | 165 |
| is a member of | 2455 | 160 |
| on a regular basis | 2775 | 110 |

## Differences between British National Corpus and Web Corpus Data

Dataset 3:

N-grams with a rank frequency of 1 to 500 in one corpus not among the 5000 most frequent in the other.

**500 most frequent 1-grams in BNC data**
**Missing among 5000 most frequent in Web data**

**[None]**

| 500 most frequent 1-grams in Web data<br>Missing among 5000 most frequent in BNC data | | |
|---|---|---|
| **Rank** | **1-Gram** | **Freq** |
| 72 | web | 6360 |
| 117 | internet | 3740 |
| 158 | http | 2918 |
| 178 | click | 2637 |
| 231 | center | 2138 |
| 234 | online | 2121 |
| 236 | copyright | 2111 |
| 237 | email | 2110 |
| 252 | e-mail | 1980 |
| 283 | server | 1792 |
| 314 | u.s | 1669 |
| 318 | © | 1640 |

| 500 most frequent 2-grams in BNC data Missing among 5000 most frequent in Web data | | |
|---|---|---|
| BNC Rank | 2-Gram | Freq |
| 297 | as she | 11719 |
| 439 | she could | 8859 |
| 449 | at her | 8803 |
| 463 | the police | 8653 |
| 485 | see p | 8346 |

| 500 most frequent 2-grams in Web data Missing among 5000 most frequent in BNC data | | |
|---|---|---|
| Web Rank | 2-gram | Freq |
| 64 | web site | 1716 |
| 84 | the internet | 1487 |
| 100 | click here | 1354 |
| 110 | this site | 1281 |
| 113 | the web | 1264 |
| 140 | home page | 1119 |
| 151 | all rights | 1016 |
| 154 | rights reserved | 1005 |
| 160 | this page | 976 |
| 232 | more information | 782 |
| 243 | copyright © | 745 |
| 282 | contact us | 661 |
| 299 | the u.s | 635 |
| 302 | the program | 631 |
| 307 | links to | 629 |
| 318 | click on | 615 |
| 395 | here for | 529 |
| 399 | real estate | 525 |
| 419 | welcome to | 502 |
| 426 | web page | 495 |
| 440 | space ghost | 485 |
| 463 | to top | 467 |
| 477 | site is | 455 |
| 493 | check out | 445 |

| 500 most frequent 3-grams in Web data Missing among 5000 most frequent in BNC data | | |
|---|---|---|
| BNC Rank | 3-Gram | Freq |
| 39 | can not be | 7109 |
| 83 | read in studio | 5019 |
| 141 | video-taped report follows | 3765 |
| 213 | she had been | 3084 |
| 297 | the labour party | 2610 |
| 336 | that she had | 2443 |
| 359 | of state for | 2360 |
| 383 | my hon friend | 2291 |
| 394 | so far as | 2260 |
| 417 | but there was | 2197 |
| 461 | on the floor | 2030 |
| 488 | but he was | 1964 |
| 494 | the sound of | 1957 |
| 495 | they had been | 1957 |
| 497 | he had to | 1953 |

| 500 most frequent 3-grams in Web data Missing among 5000 most frequent in BNC data | | |
|---|---|---|
| Web Rank | 3-Gram | Freq |
| 3 | all rights reserved | 1002 |
| 21 | click here to | 534 |
| 23 | for more information | 524 |
| 26 | on the web | 480 |
| 27 | on the internet | 479 |
| 43 | table of contents | 371 |
| 52 | click here for | 352 |
| 62 | click on the | 330 |
| 64 | you need to | 325 |
| | | |
| 73 | back to top | 294 |
| 75 | this web site | 286 |
| 77 | world wide web | 285 |
| 87 | top of page | 273 |
| 97 | how do i | 263 |
| 119 | you will be | 241 |
| 128 | you would like | 229 |
| 145 | inc all rights | 218 |
| 155 | on this site | 209 |
| 165 | this site is | 201 |
| 171 | the world wide | 195 |
| 185 | of the corporation | 187 |
| 190 | welcome to the | 182 |
| 201 | the mutual fund | 176 |
| 210 | frequently asked questions | 171 |
| 215 | you have any | 171 |
| 228 | percent of the | 167 |
| 232 | more information on | 166 |
| 240 | you will find | 164 |
| 241 | allows you to | 163 |
| 242 | be sure to | 163 |
| 253 | you can find | 161 |
| 254 | you can use | 161 |

| 500 most frequent 3-grams in Web data Missing among 5000 most frequent in BNC data | | |
|---|---|---|
| Web Rank | 3-Gram | Freq |
| 256 | available on the | 160 |
| 257 | of the internet | 160 |
| 285 | you can also | 153 |
| 288 | check out the | 151 |
| 303 | this page is | 148 |
| 315 | you will need | 145 |
| 316 | more information about | 144 |
| 318 | to the internet | 144 |
| 323 | your web site | 143 |
| 327 | cost to advertiser | 141 |
| 333 | feel free to | 140 |
| 336 | terms and conditions | 140 |
| 352 | in the u.s | 136 |
| 353 | on the net | 136 |
| 360 | new york city | 135 |
| 388 | learn more about | 130 |
| 389 | return to top | 130 |
| 402 | you wish to | 128 |
| 407 | the web site | 127 |
| 415 | portion of the | 125 |
| 416 | a web site | 124 |
| 435 | university of texas | 122 |
| 450 | are interested in | 119 |
| 468 | how to use | 117 |
| 480 | in the state | 116 |
| 481 | links to other | 116 |
| 483 | take a look | 116 |
| 485 | the internet and | 116 |
| 497 | microsoft internet explorer | 114 |
| 499 | where you can | 114 |

| 500 most frequent 3-grams in BNC data<br>Missing among 5000 most frequent in Web data | | |
|---|---|---|
| BNC Rank | 5-gram | Freq |
| 12 | ask the secretary of state | 590 |
| 13 | at the back of the | 574 |
| 14 | at the end of a | 459 |
| 15 | at the start of the | 458 |
| 16 | award type research grant project | 430 |
| 17 | but at the same time | 426 |
| 18 | at the other end of | 365 |
| 19 | as in the case of | 291 |
| 20 | at the foot of the | 268 |
| 21 | for now our next bulletin | 262 |
| 22 | department of trade and industry | 235 |
| 23 | award type research development substantive | 231 |
| 24 | development substantive award ref no | 231 |
| 25 | all from us for now | 229 |
| 26 | evening anne dawson wesley smith | 226 |
| 27 | does my right hon friend | 224 |
| 28 | by the secretary of state | 218 |
| 30 | bbc summary of world broadcasts | 186 |
| 31 | and up at two marks | 182 |
| 33 | ask the prime minister if | 180 |
| 34 | as a matter of fact | 177 |
| 35 | and down at two marks | 171 |
| 36 | before the end of the | 163 |
| 37 | at the time of writing | 152 |
| 38 | barnes reports video-taped report follows | 148 |
| 40 | at the end of this | 145 |
| 41 | at the heart of the | 141 |
| 42 | at the expense of the | 129 |
| 43 | bed and breakfast beach charges | 125 |
| 44 | and summarised in appendix a | 123 |
| 45 | between fields is fully described | 123 |
| 46 | fields is fully described in | 123 |
| 47 | description the use of keys | 122 |
| 48 | detailed description the use of | 122 |
| 49 | economy economic indicators gdp growth | 119 |
| 50 | a you must provide the | 117 |
| 51 | appendix a you must provide | 117 |
| 52 | as a result of a | 115 |
| 53 | by the court of appeal | 115 |
| 54 | on the other side of | 115 |
| 55 | and in the case of | 112 |
| 56 | award type research grant award | 112 |
| 57 | by the hon member for | 110 |
| 58 | economic and social history discipline | 109 |
| 60 | barnett reports video-taped report follows | 102 |
| 61 | the end of the year | 100 |
| 62 | at a time when the | 99 |
| 63 | as a member of the | 98 |
| 64 | follows read in studio voice | 98 |
| 65 | by the house of lords | 97 |
| 66 | as the case may be | 95 |
| 67 | does my hon friend agree | 94 |

| 500 most frequent 3-grams in BNC data<br>Missing among 5000 most frequent in Web data | | |
|---|---|---|
| BNC Rank | 5-gram | Freq |
| 68 | conference on security and co-operation | 93 |
| 69 | a point of order mr | 92 |
| 70 | clare lafferty reports video-taped report | 91 |
| 71 | during the course of the | 90 |
| 72 | for foreign and commonwealth affairs | 90 |
| 74 | anne still to come on | 89 |
| 76 | an example of this option | 88 |
| 77 | example of this option is | 88 |
| 78 | far eastern economic review of | 88 |
| 79 | and business studies primary subject | 87 |
| 80 | business studies primary subject area | 87 |
| 81 | for the first time since | 87 |
| 82 | towards the end of the | 87 |
| 83 | evening anne dawson harriet ryley | 85 |
| 84 | agree with my hon friend | 84 |
| 85 | award type research contract award | 84 |
| 86 | be in a position to | 83 |
| 87 | an integral part of the | 82 |
| 88 | does the secretary of state | 82 |
| 89 | a breach of the peace | 80 |
| 90 | dinner bed and breakfast beach | 80 |
| 91 | discipline unknown subject area unknown | 79 |
| 92 | am grateful to my hon | 77 |
| 93 | and if he will make | 77 |
| 94 | and social history primary subject | 77 |
| 95 | economic and social history primary | 77 |
| 96 | at the centre of the | 76 |
| 97 | and co-operation in europe csce | 75 |
| 98 | anne dawson wesley smith video-taped | 75 |
| 99 | by my right hon friend | 75 |
| 100 | dawson wesley smith video-taped report | 75 |
| 101 | elected for a five-year term | 74 |
| 102 | first quarter net profit up | 73 |
| 103 | on the edge of the | 73 |
| 104 | contact sema software technology technical | 72 |
| 105 | award type unknown award ref | 71 |
| 106 | between the government and the | 71 |
| 107 | in the same way as | 71 |
| 109 | can be found in appendix | 70 |
| 110 | cases are referred to in | 70 |
| 111 | a bottle of sparkling wine | 69 |
| 112 | erika barnes reports video-taped report | 69 |
| 113 | the end of the century | 69 |
| 114 | the other end of the | 69 |
| 115 | agreement on tariffs and trade | 68 |
| 116 | are referred to in the | 68 |
| 117 | for the rest of the | 68 |
| 118 | the secretary of state for | 68 |
| 120 | there is no doubt that | 68 |
| 121 | both sides of the house | 67 |
| 122 | the department of the environment | 67 |
| 123 | and limitations no special lifespan | 66 |
| 124 | department unknown institution london | 66 |

| 500 most frequent 3-grams in BNC data<br>Missing among 5000 most frequent in Web data | | |
|---|---|---|
| BNC Rank | 5-gram | Freq |
| | university | |
| 125 | for the benefit of the | 66 |
| 126 | in the wake of the | 66 |
| 127 | at a meeting of the | 65 |
| 128 | at the far end of | 65 |
| 129 | in the centre of the | 65 |
| 130 | in the house of commons | 65 |
| 131 | the way in which the | 65 |
| 132 | commonwealth of independent states cis | 64 |
| 133 | department department of economics institution | 64 |
| 134 | following cases are referred to | 64 |
| 135 | at the same time it | 63 |
| 136 | committee of the red cross | 62 |
| 137 | facts are stated in the | 62 |
| 139 | are required to use this | 61 |
| 140 | cases were cited in argument | 61 |
| 141 | clark reports video-taped report follows | 61 |
| 142 | commission of the european communities | 61 |
| 143 | as my hon friend the | 60 |
| 144 | court of appeal held that | 60 |
| 145 | dinner or lunch bed and | 60 |
| 146 | additional cases were cited in | 59 |
| 147 | and methodology primary subject area | 59 |
| 148 | computing and methodology primary subject | 59 |
| 149 | contract for the sale of | 59 |
| 150 | debt service ratio as of | 59 |
| 151 | effect of an accounting change | 59 |
| 152 | at the head of the | 58 |
| 153 | by my hon friend the | 58 |
| 154 | cumulative effect of an accounting | 58 |
| 155 | in the direction of the | 58 |
| 156 | a loss last time of | 57 |
| 157 | at the request of the | 57 |
| 158 | cent of gross domestic product | 57 |
| 159 | does the minister agree that | 57 |
| 160 | first lateral arm plate the | 57 |
| 161 | and for the first time | 56 |
| 162 | and i am sure that | 56 |
| 163 | compagnie des machines bull sa | 56 |
| 165 | per cent of the vote | 56 |
| 166 | the turn of the century | 56 |
| 167 | after the second world war | 55 |
| 168 | am grateful to the hon | 55 |
| 169 | as well as in the | 55 |
| 170 | department department of psychology institution | 55 |
| 172 | at a press conference on | 54 |
| 173 | at half-past two o'clock prayers | 54 |
| 174 | on the side of the | 54 |
| 176 | the department of trade and | 54 |
| 177 | the end of the day | 54 |
| 178 | cent of the vote and | 53 |

| 500 most frequent 3-grams in BNC data<br>Missing among 5000 most frequent in Web data | | |
|---|---|---|
| BNC Rank | 5-gram | Freq |
| 179 | child in the family of | 53 |
| 180 | a contract for the sale | 52 |
| 181 | are stated in the judgment | 52 |
| 182 | arm plate the oral shield | 52 |
| 183 | breakfast supplements per person per | 52 |
| 184 | it is not surprising that | 52 |
| 186 | the end of the war | 52 |
| 187 | and breakfast supplements per person | 51 |
| 188 | and international relations primary subject | 51 |
| 189 | and social history subject area | 51 |
| 190 | asset no mouse no others | 51 |
| 191 | bed and breakfast supplements per | 51 |
| 192 | economic and social history subject | 51 |
| 193 | following additional cases were cited | 51 |
| 194 | and breakfast beach charges are | 50 |
| 195 | at an additional cost of | 50 |
| 196 | decision of the court of | 50 |
| 197 | in the first half of | 50 |
| 198 | in the hands of the | 50 |
| 199 | of the second world war | 50 |
| 200 | to be one of the | 50 |

| 500 most frequent 3-grams in Web data<br>Missing among 5000 most frequent in BNC data | | |
|---|---|---|
| Web Rank | 5-Gram | Freq |
| 2 | if you would like to | 109 |
| 4 | messages sorted by date thread | 74 |
| 5 | on the world wide web | 74 |
| 6 | sorted by date thread subject | 74 |
| 7 | by date thread subject author | 72 |
| 9 | university of texas at austin | 62 |
| 10 | if you have any questions | 61 |
| 11 | the university of texas at | 57 |
| 15 | img src http broccoli.mfn.ki.se aj | 51 |
| 16 | src http broccoli.mfn.ki.se aj gifs | 51 |
| 17 | user interfaces for information visualization | 51 |