The 'Schizandrafield' Corpus - an English 1-gram corpus 788,068,084 distinct-words strong

[Schizandrafield 1-gram Corpus, revision C, derives from next 44 corpora:]

Corpus Tag, Name	Corpus size (in bytes)	Total Words	Unique Words	Needed memory to rip in a single pass
AHD, American_Heritage_Dictionary_4_(En-En)_WHOLEWORDS.dsl	41,742,099	7,083,439	176,377	16,512KB
BNC, Machine-Learning_British-National-Corpus_XML-edition.tar	4,680,140,800	980,238,337	367,921	34,389KB
BRE, Britannica_Encyclopedia_2010_1.563_miled_(En-En)_ANSI.dsl	297,981,779	46,958,317	356,417	33,322KB
CAL, Cambridge_Advanced_Learner's_Dictionary_4th_Ed_(En-En).dsl	76,257,922	13,994,351	57,787	5,417KB
CCA, Collins_COBUILD_Advanced_Learner's_English_Dictionary_(5th_ed)_(En-En)_WHOLEWORDS.dsl	22,960,485	4,553,432	61,575	5,772KB
DMC, DeepMind_Q_and_A_Dataset_cnn_downloads_(92579_files).tar	7,270,204,928	988,686,300	499,244	46,623KB
DMD, DeepMind_Q_and_A_Dataset_dailymail_downloads_(219506_files).tar	59,019,643,392	7,533,039,068	886,887	82,582KB
EDR, encyclopediadramaticase-20150628-current.tar	302,959,104	41,642,066	456,251	42,635KB
EJD, Encyclopaedia_Judaica_(in_22_volumes)_TXT.tar	107,784,192	16,158,364	195,273	18,281KB
EJN, ENAMDICT_Japanese_names	26,392,511	1,645,705	343,460	32,106KB
FDU, For_Dummies_978-ebooks_Collection.tar	811,308,544	122,351,592	302,353	28,282KB
GGB, Google_Books_corpus_version_20130501_English_All_Nodes.txt	10,624,363,237	178,439,407	7,477,257	664,635KB
HCN, Hacker_News_2006_to_2017-jul.json	7,046,506,518	1,075,832,079	2,188,058	201,838KB
IST, INTERNET_SACRED_TEXT_ARCHIVE_DVD-ROM_9_(English_140479_htm_files).tar	2,037,880,832	304,410,076	1,333,036	123,688KB
LDC, Longman_Dictionary_of_Contemporary_English_5th_Ed_(En-En)_WHOLEWORDS.dsl	52,870,741	9,722,686	85,217	7,987KB
MCD, Macmillan_English_Dictionary_(En-En)_.dsl	79,686,074	11,750,813	67,340	6,312KB
MCT, Macmillan_English_Thesaurus_(En-En).dsl	29,580,755	4,650,089	39,528	3,708KB
NSO, New_Shorter_Oxford_English_Dictionary_fifth_edition.tar	132,728,832	25,920,769	259,990	24,321KB
OED, Oxford_English_Dictionary_2nd_Edition_Version_4_(En-En)_WHOLEWORDS.dsl.txt	564,235,251	101,798,550	1,089,240	101,214KB
OSH, OSHO.TXT	206,908,949	31,957,006	58,893	5,522KB
PGT, Project_Gutenberg_DVD-2010_(29180_files).tar	11,110,769,152	1,870,216,915	3,847,963	350,566KB
RDD, Reddit_Comments_(JSON_objects)_from_(2005-12_to_2018-01).json	2,277,975,364,152	333,829,940,270	689,949,388	54,703,959KB
RHW, Random_House_Webster's_Unabridged_Dictionary_(En-En)_.dsl	53,483,152	9,367,457	282,580	26,428KB
SNT, Machine-Learning_WestburyLab.NonRedundant.UsenetCorpus_(47860_English_language_non-binary-file_news_groups).tar	39,513,013,248	6,316,689,948	4,835,188	437,662KB
STX, archive.org_stackexchange_(346_corpora_2017-Oct-12).tar	274,935,801,856	38,077,068,727	29,194,792	2,344,214KB
TAL, the-anarchist-library-2016-01-18-en_html.tar	153,703,936	24,339,935	136,000	12,738KB
TXF, TEXTFILES.COM_(58096_files).tar	1,382,122,496	192,893,874	1,008,780	93,840KB
URB, Machine-Learning_Urban_Dictionary_Definitions_Corpus_(1999_-_May-2016).words.json	1,917,822,288	263,253,093	2,631,962	241,852KB
WKD, dumps.wikimedia.org_Germany_dewiki-20180220-pages-articles.xml	18,954,897,343	2,362,729,484	17,415,343	1,467,593KB
WKE, dumps.wikimedia.org_English_enwiki-20180220-pages-articles.xml	65,865,333,874	8,739,196,084	39,440,894	3,071,920KB
WKF, dumps.wikimedia.org_France_frwiki-20180220-pages-articles.xml	17,802,386,071	2,429,769,009	12,192,025	1,055,599KB
WKI, dumps.wikimedia.org_Italy_itwiki-20180220-pages-articles.xml	10,887,321,918	1,372,430,005	8,960,466	790,544KB
WKN, dumps.wikimedia.org_Netherlands_nlwiki-20180220-pages-articles.xml	6,808,875,477	800,283,886	8,596,100	760,225KB
WKP, dumps.wikimedia.org_Portugal_ptwiki-20180220-pages-articles.xml	6,891,588,341	940,571,722	6,736,349	602,633KB
WKS, dumps.wikimedia.org_Spain_eswiki-20180220-pages-articles.xml	12,200,295,384	1,682,091,803	9,780,910	858,723KB
WMB, dumps.wikimedia.org_English_enwikibooks-20180220-pages-articles.xml	641,413,774	94,801,223	971,592	90,438KB
WMN, dumps.wikimedia.org_English_enwikinews-20180220-pages-articles.xml	201,872,863	27,328,349	404,890	37,839KB
WMP, dumps.wikimedia.org_English_specieswiki-20180220-pages-articles.xml	1,009,303,358	107,856,282	2,765,079	254,016KB
WMQ, dumps.wikimedia.org_English_enwikiquote-20180220-pages-articles.xml	410,396,147	64,809,361	561,894	52,455KB
WMS, dumps.wikimedia.org_English_enwikisource-20180220-pages-articles.xml	8,352,677,820	1,282,920,260	8,547,509	755,716KB
WMU, dumps.wikimedia.org_English_enwikiversity-20180220-pages-articles.xml	369,987,893	52,322,592	640,606	59,767KB
WMV, dumps.wikimedia.org_English_enwikivoyage-20180220-pages-articles.xml	354,919,743	49,701,293	734,403	68,474KB
WMW, dumps.wikimedia.org_English_enwiktionary-20180220-pages-articles.xml	5,379,842,566	597,315,425	14,743,543	1,259,179KB
WUD, Webster's_Unabridged_3_(En-En)_WHOLEWORDS_ANSI.dsl	134,706,719	24,014,478	364,352	34,052KB

[Schizandrafield 1-gram Corpus, revision C, holds within next 44 tagged-counted-wordlists:]

```
237,859,509 dumps.wikimedia.org_Spain_eswiki-20180220-pages-articles.1gram
161,745,186 dumps.wikimedia.org_Portugal_ptwiki-20180220-pages-articles.1gram
211,751,899 dumps.wikimedia.org_Netherlands_nlwiki-20180220-pages-articles.1gram
216,667,608 dumps.wikimedia.org_Italy_itwiki-20180220-pages-articles.1gram
969,819,544 dumps.wikimedia.org_English_enwiki-20180220-pages-articles.1gram
452,471,230 dumps.wikimedia.org_Germany_dewiki-20180220-pages-articles.1gram
294,822,021 dumps.wikimedia.org_France_frwiki-20180220-pages-articles.1gram
 23,100,108 dumps.wikimedia.org_English_enwikibooks-20180220-pages-articles.1gram
204,113,124 dumps.wikimedia.org_English_enwikisource-20180220-pages-articles.1gram
  9,347,743 dumps.wikimedia.org_English_enwikinews-20180220-pages-articles.1gram
 13,259,587 dumps.wikimedia.org_English_enwikiquote-20180220-pages-articles.1gram
 15,131,207 dumps.wikimedia.org_English_enwikiversity-20180220-pages-articles.1gram
 18,031,161 dumps.wikimedia.org_English_enwikivoyage-20180220-pages-articles.1gram
355,881,851 dumps.wikimedia.org_English_enwiktionary-20180220-pages-articles.1gram
 65,846,403 dumps.wikimedia.org_English_specieswiki-20180220-pages-articles.1gram
  4,653,581 Encyclopaedia_Judaica_(in_22_volumes)_TXT.1gram
  8,889,960 Webster's_Unabridged_3_(En-En)_WHOLEWORDS_ANSI.dsl.1gram
  3,288,498 the-anarchist-library-2016-01-18-en_html.tar.1gram
  6,973,687 Random_House_Webster's_Unabridged_Dictionary_(En-En)_.dsl.1gram
 26,501,920 Oxford_English_Dictionary_2nd_Edition_Version_4_(En-En)_WHOLEWORDS.dsl.txt.1gram
  1,404,700 OSHO.TXT.1gram
  6,280,919 New_Shorter_Oxford_English_Dictionary_fifth_edition.tar.1gram
    941,039 Macmillan_English_Thesaurus_(En-En).dsl.1gram
  1,609,423 Macmillan_English_Dictionary_(En-En)_.dsl.1gram
  2,044,933 Longman_Dictionary_of_Contemporary_English_5th_Ed_(En-En)_WHOLEWORDS.dsl.1gram
  7,354,174 For_Dummies_978-ebooks_Collection.tar.1gram
 11,464,911 encyclopediadramaticase-20150628-current.tar.1gram
  8,840,362 ENAMDICT_Japanese_names.1gram
  1,466,506 Collins_COBUILD_Advanced_Learner's_English_Dictionary_(5th_ed)_(En-En)_WHOLEWORDS.dsl.1gram
  1,380,759 Cambridge_Advanced_Learner's_Dictionary_4th_Ed_(En-En).dsl.1gram
  8,621,181 Britannica_Encyclopedia_2010_1.563_miled_(En-En)_ANSI.dsl.1gram
  4,268,314 American_Heritage_Dictionary_4_(En-En)_WHOLEWORDS.dsl.1gram
186,006,261 Google_Books_corpus_version_20130501_English_All_Nodes.1gram
 32,262,044 INTERNET_SACRED_TEXT_ARCHIVE_DVD-ROM_9_(English_140479_htm_files).1gram
 96,997,413 Project_Gutenberg_DVD-2010_(29180_files).1gram
 24,179,167 TEXTFILES.COM_(58096_files).1gram
797,764,729 archive.org_stackexchange_(346_corpora_2017-Oct-12).1gram
  8,823,193 Machine-Learning_British-National-Corpus_XML-edition.1gram
 65,616,610 Machine-Learning_Urban_Dictionary_Definitions_Corpus_(1999_-_May-2016).words.1gram
120,939,472 Machine-Learning_WestburyLab.NonRedundant.UsenetCorpus_(47860_English_language_non-binary-file_news_groups).1gram
 12,375,594 DeepMind_Q_and_A_Dataset_cnn_downloads_(92579_files).1gram
 20,878,468 DeepMind_Q_and_A_Dataset_dailymail_downloads_(219506_files).1gram
16,261,229,827 reddit.1gram
 54,744,506 Hacker_News_2006_to_2017-jul.json.1gram
```

```
s - surname (138,500)
p - place-name (99,500)
u - person name, either given or surname, as-yet unclassified (139,000)
g - given name, as-yet not classified by sex (64,600)
f - female given name (106,300)
m - male given name (14,500)
h - full (usually family plus given) name of a particular person (30,500)
pr - product name (55)
c - company name (34)
st - stations (8,254)
```

[The way they were ripped/concatenated/sorted:]

```
E:\Schizandrafield_workshop>dir dumps.wikimedia.org_English_enwiki-20180220-pages-articles.xml/b  >dumps.wikimedia.org_English_enwiki-20180220-pages-articles.lst
E:\Schizandrafield_workshop>echo WKE>Leprechaun.tag
E:\Schizandrafield_workshop>Leprechaun_x-leton_32bit_Intel_01_008p.exe dumps.wikimedia.org_English_enwiki-20180220-pages-articles.lst dumps.wikimedia.org_English_enwiki-20180220-pages-articles.1gram 1399888 Y
E:\Schizandrafield_workshop>type dumps.wikimedia.org_English_enwiki-20180220-pages-articles.1gram|more
WKE_0,000,003_byjnf
WKE_0,000,001_richardmullaney
...
C:\>copy/b *.1gram unsorted
C:\>sort.exe /+15 /M 1048576 /T d: "unsorted" /O "Schizandrafield_Corpus_revision_C_(44-corpora_-unique-words).sorted"
```