Making a Scientific Research Article Word List

Masaya Kanzaki

Abstract

This study created the Scientific Research Article Word List (SRAWL) out of the titles and abstracts of scientific research articles. The purpose of the list is to show scientists who are not native speakers of English what words they need to learn in order to handle scientific research articles in English. When scientists are determining whether or not a research article is worth reading, they look at the title first and then the abstract. Therefore, it is helpful if they learn words that are frequently used in these two parts. In order to make a list of such words, the titles and abstracts of 12,968 research articles and reports, published in *Science* between 2000 and 2016, were collected to create a corpus of 1.7 million words. Lexical coverage and word frequency were then investigated. The most frequent 7,850 lemmas appeared 13 times or more in the corpus and accounted for 94.75% of the total tokens. By excluding proper nouns, marginal words, and abbreviations from them, a list of 6,947 lemmas was created. The list was divided into four sublists according to the degrees of importance so that users will know which groups of words are more important than others.

Scientists need to read scientific research articles in English in order to keep up with the latest developments in their fields, even if their first language is not English, as the vast majority of such articles are published in English. At the same time, scientists are busy people, and so they have to be selective in choosing which articles to read. When they are deciding whether or not to read an article, they look at the title first. If they find it interesting and relevant to their needs, they look at the abstract next. This selection process would be easier, however, if they knew the words that most frequently appear in these two parts of articles. The purpose of this study was to make a list of such words.

Frequency of occurrence is the vital factor in choosing words for a word list. It has long

been known that frequency varies widely from word to word. Zipf (1949) described the mathematical relationship between the frequency of a word and its rank on a frequency table; the frequency is inversely proportional to its rank. Nation and Webb (2011) emphasized the relevance of this law, known as Zipf's Law, in vocabulary learning:

> What Zipf's Law describes is that there is a small number of very frequent items that cover a very large proportion of the text, and there is a very large number of infrequent items that cover only a small proportion of the text. Clearly, learning the frequent items first will be of great benefit to learners. (p. 132)

Following this line of thought, the study identified high-frequency words in the titles and abstracts of scientific research articles in order to make the Scientific Research Article Word List (SRAWL).

Method

Materials

The scientific journal used in this study was *Science*, which had the second highest impact factor after *Nature* among journals that cover the full range of scientific disciplines. *Science* was chosen over *Nature* because its online archive was well maintained and easy to navigate. *Science* is published weekly, and each issue has one to five research articles and 10 to 15 reports, which are shorter versions of research articles. Table 1 shows the numbers of research articles and reports published between 2000 and 2016, from which titles and abstracts were collected to create a scientific research article corpus (SRAC). The total number of research articles and reports published during those 17 years was 12,968.

Table 1

Research Articles and Reports Published in Science (2000-2016)

Year	Number
2000	772
2001	793
2002	797
2003	771
2004	767
2005	749
2006	771
2007	758
2008	758
2009	778
2010	761
2011	792
2012	737
2013	749
2014	750
2015	710
2016	755
Total	12,968

The SRAC was made by digitally copying the titles and abstracts from the online issues of the journal and pasting them one at a time into an MS Excel spreadsheet. The total number of tokens in the SRAC was 1,682,195.

Two vocabulary analysis software programs were used. One is AntWordProfiler, designed to perform vocabulary profiling. It compares a text against a set of vocabulary lists and generates vocabulary statistics and frequency information. The other is AntConc, which has multiple functions that support corpus linguistics research. The word list function of the program was used in this study to generate a lemmatized list.

Unit of Counting

There are different word counting units, and depending on the purpose of a word list, a suitable unit needs to be chosen. Bauer and Nation (1993) suggested seven levels of word families, each of which can be used as a unit of counting. Table 2 summarizes the criteria for the seven levels.

Table 2

Summary of the Bauer and Nation (1993) Levels

Level	Criteria
1	A different form is a different word. Capitalization is ignored.
2	Regularly inflected words are part of the same family. The inflectional categories are: plural; third person singular present tense; past tense; past participle; -ing; comparative; superlative; possessive.
3	-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-, all with restricted uses
4	-al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in-, all with restricted uses
5	-age (leakage), -al (arrival), -ally (idiotically), -an (American), -ance (clearance), -ant (consultant), -ary (revolutionary), -atory (confirmatory), -dom (kingdom; officialdom), -eer (black marketeer), -en (wooden), -en (widen), -ence (emergence), -ent (absorbent), -ery (bakery; trickery), -ese (Japanese; officialese), -esque (picturesque), -ette (usherette; roomette), -hood (childhood), -i (Israeli), -ian (phonetician; Johnsonian),

	-ite (Paisleyite; also chemical meaning), -let (coverlet), -ling (duckling), -ly (leisurely), -most (topmost), -ory (contradictory), -ship (studentship), -ward (homeward), -ways (crossways), -wise (endwise; discussion-wise), anti- (anti-inflation), ante- (anteroom), arch- (archbishop), bi- (biplane), circum- (circumnavigate), counter- (counter-attack), en- (encage; enslave), ex- (ex-president), fore- (forename), hyper- (hyperactive), inter- (inter-African, interweave), mid- (mid-week), mis- (misfit), neo- (neo-colonialism), post- (post-date), pro- (pro-British), semi- (semi-automatic), sub- (subclassify; subterranean), un- (untie; unburden)
6	-able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, re-
7	Classical roots and affixes

Note. Adapted from *Making and Using Word Lists for Language Learning and Testing* (p. 27) by I. S. P. Nation, 2016, Amsterdam: John Benjamins.

Note that the Level 1 word families are the same as "types," and the Level 2 word families are the same as "lemmas." Also, the term "word families" customarily refers to the Level 6 word families in the literature, and this study follows the convention.

Some of the influential word lists, such as West's (1953) General Service List (GSL), Coxhead's (2000) Academic Word List (AWL), and Nation's (2017) BNC/COCA Lists, use word families as the unit of counting. One could argue that the word family is an appropriate level of counting if a list is intended to assist receptive use. This is because once one member of a word family is learned, the learner can guess the meanings of the other members when they see them. However, some researchers have expressed doubts about this assumption, arguing that EFL learners do not have sufficient knowledge about affixes to do so (e.g., Ward, 2009; Ward & Chuenjundaeng, 2009). McLean (2018) found that Japanese EFL learners do not comprehend inflectional and derivational forms of words even when they know the base forms, suggesting that the word family is not an appropriate unit of counting for them.

Some newer lists, such as Brezina and Gablasova's (2013) New General Service List, Gardner and Davies's (2014) Academic Vocabulary List (AVL), Browne's (2014) New General Service List, and Lei and Liu's (2016) Medical Academic Vocabulary List (MAVL), use lemmas as the unit of counting. A lemma consists of the headword and its inflectional forms only, and therefore it is more suitable for low proficiency learners than the word family. Furthermore, English dictionaries are organized based on lemmas, and therefore learners are familiar with grouping words by lemmas. Since the target users of the SRAWL include scientists with low English proficiency who lack knowledge about affixes, the lemma was chosen as the unit of counting for the SRAWL.

Note that AntConc treats words with the same spelling as the same word. For example, the noun *work* and the verb *work* are counted as the same lemma, although strictly speaking they belong to different lemmas because they are different parts of speech. A distinctive term for a lemma without the part of speech restriction is "flemma," which was introduced in 2014 and has not yet been widely used outside of the vocabulary research literature. The word "lemma" is used to mean "flemma" in this study.

Another issue concerning the lemma is that lemma counting in this study was restricted by the lemma list used with AntConc. The edited version of the Someya English lemma list, available on Laurence Anthony's Website and containing 39,296 tokens in 14,189 lemma groups, was used in conjunction with AntConc to count lemmas. Although the lemma list is sufficiently large to lemmatize most of the words in the SRAC, there are a few words that are not in the list. Such words were not grouped together as lemmas in AntConc output. For example, *acceptor* and *acceptors* should be counted as the same lemma but were counted separately. That is to say, words that are not covered by the lemma list are counted by the type.

Inclusion of a General Vocabulary

A general vocabulary is usually excluded from a specialized word list. For example, Coxhead (2000) excluded words in the GSL from the AWL, Coxhead and Hirsh (2007) chose words outside the GSL and the AWL for the Science-Specific Word List, and Hsu (2013) omitted the most frequent 3,000 words in the British National Corpus from the Medical Word List.

Some word list creators took a more nuanced approach in dealing with a general vocabulary. Gardner and Davies (2014), for example, included some general high-frequency words in the AVL based on relative frequencies. The criterion was that words must be at least 50% more frequent in the academic corpus than in the non-academic corpus. Moreover, Lei and Liu (2016) applied a special meaning criterion for general high-frequency words when choosing words for the MAVL; as well as being 1.5 times more frequent in the medical academic corpus than in the general corpus, the general high-frequency words in the MAVL had to be in two commonly used medical English dictionaries.

Contrary to these studies, this study took a simpler approach to a general vocabulary; the SRAWL included all the general high-frequency words that were frequent in the SRAC. The rationale behind this decision was that some of the prospective users would be low proficiency learners who have not mastered high-frequency words in general English. Therefore, it was thought to be helpful to include such words in the list. This approach was similar to the one taken by Ward (2009) in the creation of the Basic Engineering List, although he omitted function words from the list.

Results and Discussion

Vocabulary Profile of the SRAC

Lexical coverage of the SRAC was investigated with AntWordProfiler, first using the GSL and AWL family lists and then the BNC/COCA lists.

AntWordProfiler compares a text against vocabulary lists and shows what percentage of the items in the text are covered by each list. Table 3 shows the percentage, the cumulative coverage, and the number of word families of the first and second 1,000 words of the GSL and the AWL in the SRAC. The coverage of the GSL over the SRAC was 59.03%, comprising 53.53% over the first 1,000 and 5.50% over the second 1,000. The coverage of the AWL was 12.34%. Compared with the coverage figures reported in Coxhead and Hirsh (2007) for their pilot science corpus, the figures for the first and second 1,000 of the GSL were 12.16% and 0.33% lower, respectively. However, the AWL coverage over the SRAC was 3.38% higher than the coverage over Coxhead and Hirsh's science corpus.

The second 1,000 of the GSL has 419 more words than the AWL, but the coverage of the AWL was 6.84% higher.

Table 3

Percentage, Cumulative Coverage, and Number of Word Families of the GSL and AWL over the SRAC

Word list	Percentage (%)	Cumulative coverage (%)	Word families
GSL 1st 1,000	53.53	53.53	953
GSL 2nd 1,000	5.50	59.03	779
AWL 570	12.34	71.37	557
Not in the lists	28.63	100	NA

Note. The exact numbers of word families contained in the GSL 1st 1,000, GSL 2nd 1,000, and AWL 570 are 998, 988, and 569, respectively. The word *found* is in the GSL 1st 1,000, so it was removed from the AWL.

The SRAC was also examined against the BNC/COCA lists for lexical coverage. Among the 29 BNC/COCA lists, the first 25 contain 1,000 word families each are roughly based on

frequency; the first list contains the most frequent 1,000 word families, the second contains the next most frequent 1,000, and so on, although some adjustments were made to the first two lists to include some useful words for learners, irrespective of their frequency. The last four lists are of proper nouns, marginal words (e.g., letters of the alphabet), transparent compounds (= compound words whose meanings are clear from the meanings of the parts), and abbreviations, respectively.

Table 4 shows the percentage, the number of word families, and the cumulative coverage of the BNC/COCA lists over the SRAC. The coverage of the first three lists, which consist of general high-frequency words, is low compared to corpora consisting of other types of texts. For example, the coverage of the first three lists over Webb and Rodgers's (2009) movie corpus was 92.39, which is 17.01% higher than the coverage over the SRAC. The words in the SRAC spread across different frequency levels, but this does not mean that the vocabulary load of the SRAC is high. The total number of word families that appeared in the 25 lists is 9,245, which is 1,464 fewer than the figure for Webb and Rodgers's movie corpus. In this respect, the vocabulary load of the SRAC is lower than that of the movie corpus.

Table 4

Percentage, Cumulative Coverage, and Number of Word Families of the BNC/COCA List over the SRAC

Word list	Percentage (%)	Cumulative coverage (%)	Word families
1,000	49.18	49.18	917
2,000	12.52	61.70	853
3,000	13.68	75.38	891
4,000	4.00	79.38	707
5,000	2.48	81.86	632
6,000	1.88	83.74	550

7,000	1.25	84.99	471
8,000	0.91	85.90	418
9,000	0.58	86.48	350
10,000	0.47	86.95	325
11,000	0.46	87.41	318
12,000	0.38	87.79	260
13,000	0.49	88.28	264
14,000	0.45	88.73	247
15,000	0.43	89.16	263
16,000	0.42	89.58	271
17,000	0.25	89.83	212
18,000	0.25	90.08	231
19,000	0.24	90.32	215
20,000	0.20	90.52	227
21,000	0.12	90.64	176
22,000	0.07	90.71	113
23,000	0.08	90.79	138
24,000	0.03	90.82	90
25,000	0.07	90.89	106
PNs	1.08		
MWs	0.47		
TCs	0.41		
ABs	0.47		
Not in the lists	6.71		

Note. PNs = proper nouns; MWs = marginal words; TCs = transparent compounds; ABs = abbreviations.

High-frequency Words in the SRAC

A list of all lemmas in the SRAC was generated by AntConc. In order to decide how many were included in the SRAWL, the coverage over the SRAC was considered. Research indicates that 95% coverage is required for adequate comprehension (i.e., learners need to know 95% of the words in a text to understand it). For example, Laufer (1989) suggested that 95% coverage is needed for reasonable reading comprehension of an academic text. Some studies have suggested an even higher coverage of 98% is required for comprehension (e.g., Hirsh & Nation, 1992; Hu & Nation, 2000). Therefore, the 95% coverage is considered to be the minimum threshold. In the SRAC, the most frequent 7,850 lemmas accounted for 94.75% of the total tokens, which was close to the 95% threshold.

These 7,850 lemmas included proper nouns, marginal words, and abbreviations, which are not as important as ordinary words. Some of the high-frequency proper nouns are *Drosophila* (a type of fly), *Fe* (the chemical symbol for iron), *HIV*, *Atlantic*, and *Pacific*. While a lay person may not be familiar with *Drosophila*, biologists are likely to know it. In all probability, the majority of proper nouns are known by scientists, and even when they are unknown, this does not hinder comprehension. Proper nouns were therefore excluded from the SRAWL.

All the marginal words in the SRAC were letters of the alphabet and thus excluded from the list. Abbreviations are combinations of letters and not proper words. Common abbreviations are known by scientists, and less common abbreviations are used after the full forms of the words are given, such as with *adenosine triphosphate (ATP)*. Consequently, including them in the list was deemed unnecessary. By removing the words, the SRAWL was formed with the 6,947 lemmas.

Dividing the SRAWL into Four Sublists

A list of 6,947 words was thought to be too long for learners to use effectively. Therefore, it was divided into four sublists to make it user friendly. The four sublists, in order of importance,

were named as BASIC1965, CORE1975, MIDDLE1768, and EXTRA1240.

BASIC1965 consists of 1,963 lemmas that belong to the first two BNC/COCA lists. The 1,963 lemmas account for 61.56% of the SRAC. Since all of these are general high-frequency words, scientists who have reached a certain level of English proficiency may know most of them. Tables 5 and 6 show the first and last 30 items in BASIC1965, respectively (see Supplementary Data for the link to the entire list).

Table 5

First 30 Items in BASIC1965

Rank	Frequency	Lemma	Rank	Frequency	Lemma
1	88,200	the	16	7,252	on
2	74,937	of	17	7,220	as
3	46,948	and	18	6,077	at
4	44,805	a	19	5,008	use
5	42,518	in	20	4,955	which
6	38,323	be	21	4,556	show
7	31,029	to	22	3,779	high
8	23,056	that	23	3,708	can
9	14,903	by	24	3,626	or
10	14,740	we	25	3,618	result
11	13,750	for	26	3,397	s
12	13,657	with	27	3,344	between
13	12,354	this	28	3,166	but
14	9,157	from	29	3,071	two
15	8,656	have	30	3,009	human

Table 6

Last 30 Items in BASIC1965

Rank	Frequency	Lemma	Rank	Frequency	Lemma
1,936	13	beat	1,951	13	obvious
1,937	13	breakage	1,952	13	perfectly
1,938	13	cast	1,953	13	rating
1,939	13	corner	1,954	13	remember
1,940	13	defensive	1,955	13	sight
1,941	13	economically	1,956	13	singly
1,942	13	empty	1,957	13	somewhat
1,943	13	exchanger	1,958	13	starve
1,944	13	expectancy	1,959	13	stay
1,945	13	faithful	1,960	13	strange
1,946	13	handedness	1,961	13	timely
1,947	13	helping	1,962	13	topic
1,948	13	labor	1,963	13	valuation
1,949	13	mismatches	1,964	13	weapon
1,950	13	neighborhoods	1,965	13	weed

CORE1975 consists of 1,975 lemmas that appeared 48 times or more in the SRAC and that are not included in the first two BNC/COCA lists. The 1,975 lemmas account for 25.58% of the SRAC. The minimum frequency of 48 was set in accordance with Coxhead's (2000) frequency criterion for the AWL. The words in the AWL had to occur at least 100 times in the Academic Corpus of 3.5 million words. The SRAC has 1,682,219 words, which is about 48% of 3.5 million words; 48 times in the SRAC is the same ratio as 100 times in Coxhead's

Academic Corpus. Note that Coxhead (2000) used the word family as the unit of counting, while in this study the lemma was used; 48 lemmas in 1.7 million words is a more stringent criterion than 100 word families in 3.5 million words. Tables 7 and 8 show the first and last 30 items in CORE1975, respectively (see Supplementary Data for the link to the entire list).

Table 7

First 30 Items in CORE1975

Rank	Frequency	Lemma	Rank	Frequency	Lemma
1	9,204	cell	16	1,813	quantum
2	6,008	protein	17	1,808	electron
3	4,321	gene	18	1,806	conjurer
4	3,761	structure	19	1,785	molecular
5	2,792	complex	20	1,737	molecule
6	2,654	mechanism	21	1,727	data
7	2,338	response	22	1,718	formation
8	2,324	reveal	23	1,678	receptor
9	2,318	function	24	1,633	carbon
10	2,165	dna	25	1,609	pathway
11	2,107	factor	26	1,583	mediate
12	2,069	induce	27	1,552	regulate
13	1,946	temperature	28	1,552	target
14	1,936	demonstrate	29	1,548	dynamic
15	1,872	rna	30	1,517	domain

Table 8

Last 30 Items in CORE1975

Rank	Frequency	Lemma	Rank	Frequency	Lemma
1,946	49	modifier	1,961	48	centromeres
1,947	49	moisture	1,962	48	colorectal
1,948	49	optimization	1,963	48	crossover
1,949	49	orchestrate	1,964	48	database
1,950	49	panel	1,965	48	decadal
1,951	49	pest	1,966	48	hotspots
1,952	49	photoelectron	1,967	48	nematic
1,953	49	plasmon	1,968	48	organelles
1,954	49	secretory	1,969	48	recessive
1,955	49	spliceosome	1,970	48	ribozyme
1,956	49	supermassive	1,971	48	robot
1,957	49	tau	1,972	48	salmon
1,958	49	unwind	1,973	48	stratosphere
1,959	48	acidification	1,974	48	transiently
1,960	48	attosecond	1,975	48	transparent

Unlike BASIC1965, which has general frequency words only, CORE1975 includes both high- and low-frequency words in general English, although general high-frequency words are more predominant than general low-frequency words. Table 9 shows how many items in CORE1975 are in each of the BNC/COCA lists.

Table 9

Number of CORE1975 Items in Each of the BNC/COCA Lists

Word list	# of items in CORE1975
3,000	630
4,000	281
5,000	191
6,000	123
7,000	92
8,000	78
9,000	52
10,000	47
11,000	43
12,000	35
13,000	34
14,000	44
15,000	41
16,000	34
17,000	21
18,000	23
19,000	13
20,000	13
21,000	6
22,000	3
23,000	6
24,000	0

25,000	5
TCs	22
Not in the lists	140

Note. TCs = transparent compounds.

MIDDLE1768 consists of 1,768 lemmas with a frequency of between 20 and 47 in the SRAC, which are not included in the first two BNC/COCA lists. The 1,768 lemmas account for 3.21% of the SRAC, which is substantially lower than the 25.34% coverage of CORE1975. Tables 10 and 11 show the first and last 30 items in MIDDLE1768, respectively (see Supplementary Data for the link to the entire list).

Table 10

First 30 Items in MIDDLE1768

Rank	Frequency	Lemma	Rank	Frequency	Lemma
1	47	anions	16	47	equation
2	47	antagonize	17	47	finch
3	47	breakdown	18	47	fluctuate
4	47	coexistence	19	47	illumination
5	47	confidence	20	47	Jurassic
6	47	convective	21	47	larva
7	47	converge	22	47	metabolite
8	47	deacetylase	23	47	mild
9	47	dip	24	47	morphogenetic
10	47	dissociate	25	47	myeloid
11	47	dissolution	26	47	photoreceptors

12	47	eddy	27	47	postulate
13	47	effectiveness	28	47	proinflammatory
14	47	endosomal	29	47	proteolysis
15	47	entropy	30	47	responsiveness

Table 11

Last 30 Items in MIDDLE1768

Rank	Frequency	Lemma	Rank	Frequency	Lemma
1,739	20	photocurrent	1,754	20	shale
1,740	20	phytochrome	1,755	20	shortly
1,741	20	pilus	1,756	20	shrink
1,742	20	plumage	1,757	20	simplex
1,743	20	prolyl	1,758	20	speculate
1,744	20	proteasomes	1,759	20	sphingosine
1,745	20	pyrimidine	1,760	20	sponge
1,746	20	radioactive	1,761	20	syringae
1,747	20	receiver	1,762	20	tesla
1,748	20	recurrence	1,763	20	thrombin
1,749	20	resonators	1,764	20	toroidal
1,750	20	restructure	1,765	20	trailing
1,751	20	rotary	1,766	20	transduced
1,752	20	seafloor	1,767	20	vancomycin
1,753	20	seismogenic	1768	20	waveguides

Table 12 shows how many items in MIDDLE1768 are in each of the BNC/COCA lists.

Compared to the results from CORE1975, the numbers in the higher frequency lists decreased, whereas the numbers in the lower frequency lists increased. In addition, the number of words not in any lists increased substantially, from 140 to 367.

Table 12

Number of MIDDLE1768 Items in Each of the BNC/COCA Lists

Word list	# of items in MIDDLE1768
3,000	258
4,000	158
5,000	132
6,000	92
7,000	79
8,000	76
9,000	55
10,000	47
11,000	44
12,000	42
13,000	51
14,000	41
15,000	56
16,000	48
17,000	30
18,000	35
19,000	32
20,000	36

21,000	15
22,000	13
23,000	5
24,000	5
25,000	9
TCs	42
Not in the lists	367

Note. TCs = transparent compounds.

EXTRA1240 consists of 1,240 lemmas with a frequency of between 13 and 19 in the SRAC, which are not included in the first two BNC/COCA lists. The 1,240 lemmas account for 1.15% of the SRAC, which is negligibly low. Tables 13 and 14 show the first and last 30 items in EXTRA1240, respectively (see Supplementary Data for the link to the entire list).

Table 13

First 30 Items in EXTRA1240

Rank	Frequency	Lemma	Rank	Frequency	Lemma
1	19	adhere	16	19	chemosensory
2	19	aggressive	17	19	coarse
3	19	alga	18	19	colitis
4	19	apetala	19	19	colocalized
5	19	arrestins	20	19	confirmation
6	19	arthropods	21	19	corroborate
7	19	behaviorally	22	19	deduce
8	19	biogeography	23	19	deer

9	19	birefringence	24	19	defend
10	19	bowel	25	19	delocalized
11	19	brood	26	19	demethylase
12	19	brucei	27	19	dependency
13	19	buildup	28	19	diagnosis
14	19	caveolae	29	19	dichotomy
15	19	chemokines	30	19	disequilibrium

Table 14

Last 30 Items in EXTRA1240

Rank	Frequency	Lemma	Rank	Frequency	Lemma
1,211	13	succeed	1,226	13	transcriptomic
1,212	13	supercapacitors	1,227	13	transducers
1,213	13	supercooled	1,228	13	transmissibility
1,214	13	superelastic	1,229	13	transposase
1,215	13	symmetrical	1,230	13	tributary
1,216	13	synchronously	1,231	13	truncatula
1,217	13	tb	1,232	13	trypanosomes
1,218	13	technical	1,233	13	unidirectionally
1,219	13	tetragonal	1,234	13	urgent
1,220	13	therapeutically	1,235	13	utilize
1,221	13	thiolate	1,236	13	vigorous
1,222	13	thrive	1,237	13	viscoelastic
1,223	13	thuringiensis	1,238	13	vol
1,224	13	thyroid	1,239	13	volumetric

1,225	13	trail	1,240	13	zona

Table 15 shows how many of the items in EXTRA1240 are in each of the BNC/COCA lists. The items are spread more evenly across the levels than in the other sublists.

Table 15

Number of EXTRA1240 Items in Each of the BNC/COCA Lists

Word list	# of items in EXTRA1240
3,000	137
4,000	95
5,000	72
6,000	66
7,000	46
8,000	30
9,000	34
10,000	26
11,000	35
12,000	26
13,000	26
14,000	22
15,000	26
16,000	33
17,000	23
18,000	21
19,000	20

20,000	24
21,000	13
22,000	9
23,000	17
24,000	2
25,000	10
Not in the lists	397

How to Use the SRAWL

The SRAWL is intended to be used as a checklist that users can first go over to determine if they know the items. BASIC1965 contains the most basic 1,965 lemmas, which cover 61.56% of the SRAC. Therefore, it is important to learn them first. If there are unknown words in BASIC1965, users should look them up and learn them. However, most scientists with a decent level of English proficiency will likely know most of the words. Only those with low English proficiency will need to spend much time on BASIC1965.

For most users, the real learning will start with CORE1975, which contains advanced level vocabulary. The 1,975 lemmas cover 25.58% of the SRAC, which means they regularly appear in scientific research articles. It would therefore be advantageous to learn them, especially the high-frequency items. On the contrary, the coverages of MIDDLE1768 and EXTRA1240 are low, at 3.21% and 1.15%, respectively, so it may not be worthwhile to spend a lot of time on them, since even if you were to learn them, you would not see them often.

Supplementary Data

The four sublists of the SRAWL as well as the entire list with 6,947 lemmas are available at http://bit.ly/SRAWL20191125.

References

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2014). AntWordProfiler (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253-279. doi:10.1093/ijl/6.4.253

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, *36*(1), 1-22. https://doi.org/10.1093/applin/amt018

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, *3*, 1-10.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée*, *12*(2), 65-78. https://doi.org/10.3917/rfla.122.0065

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305-327. https://doi.org/10.1093/applin/amt015

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*(2), 689-696. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl82hirsh.pdf

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403-430. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research*, *17*(4), 454-484. http://dx.doi.org/10.1177/1362168813494121

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-323). Clevedon, UK: Multilingual Matters.

Lei, L. & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, *22*, 42-53. https://doi.org/10.1016/j.jeap.2016.01.008

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823-845. https://doi.org/10.1093/applin/amw050

Nation, I. S. P. (2017). The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]. Available from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle Cengage Learning.

Webb, S., & Rodgers, M. P. H. (2009). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407-427. https://doi.org/10.1093/applin/amp010

Ward, J. (2009). A basic engineering English word list for less proficiency foundation engineering undergraduates. *English for Specific Purposes*, *28*, 170-182. https://doi.org/10.1016/j.esp.2009.04.001

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. System, *37*, 461-469. https://doi.org/10.1016/j.system.2009.01.004

West, M. (1953). *A General Service List of English Words*. London: Longman, Green & Co.

Zipf, G. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Hafner.