http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

Frequency or Keyness?

Zorica Đurović, *Faculty of Maritime Studies Kotor,*
University of Montenegro, Kotor, Montenegro
(zoricag@ucg.ac.me)

Abstract: The possibility of compiling electronic corpora, as of the second half of the last century, has provided new opportunities for vocabulary research. This has also resulted in the development of a series of computer software solutions for the lexical analysis of texts and the building of vocabulary lists for language learners. In this article, the differences in building technical vocabulary lists according to their frequency and keyness in corpora of English for Specific Purposes (ESP) are discussed. Both criteria have been discussed in terms of their benefits and limitations, and the possibilities of the most convenient combination of both. Finally, the word frequency list has been upgraded with keywords to provide a more comprehensive, but still very attainable, word list suitable for building a bilingual glossary or to be extended into a dictionary.

Keywords: WORD LIST, FREQUENCY, KEYNESS, KEYWORDS, CORPUS, MARINE ENGINEERING, ENGLISH, LEXIS, VOCABULARY

Opsomming: Frekwensie of sleutelstatus? Die moontlikheid om elektroniese korpora sedert die tweede helfte van die laaste eeu saam te stel, het nuwe geleenthede vir woordeskatnavorsing geskep. Dit het ook gelei tot die ontwikkeling van 'n reeks rekenaarsagteware-oplossings vir die leksikale ontleding van tekste en die saamstel van woordeskatlyste vir taalaanleerders. In hierdie artikel word die verskille in die samestelling van tegniese woordeskatlyste volgens hul frekwensie en sleutelstatus in korpora van Engels vir Spesifieke Doelwitte (ESD) bespreek. Albei kriteria word in terme van hul voor- en nadele bespreek, asook die moontlikhede van die gerieflikste kombinasie van beide. Laastens is die woordfrekwensielys aangevul met sleutelwoorde om 'n omvattender, maar steeds heel haalbare woordelys te verskaf wat geskik is om 'n tweetalige glossarium saam te stel of wat uitgebrei kan word tot 'n woordeboek.

Sleutelwoorde: WOORDELYS, FREKWENSIE, SLEUTELSTATUS, SLEUTELWOORDE, KORPUS, SKEEPSINGENIEURSWESE, ENGELS, LEKSIS, WOORDESKAT

1. Introduction

The possibility of creating and analysing electronic corpora has provided course designers and lexicographers with new opportunities in designing vocabulary (teaching and learning) material. The first corpus-based dictionaries were General English (GE) ones, such as Monolingual Learner's Dictionaries and *Collins COBUILD English Language Dictionary*, providing for a justified selection of words and original corpus-based examples (cf. Hanks 2020). Today, corpora and corpus-based tools are considered almost a conventional approach

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

to building lexicographic materials (Sinclair 1992; Abdelzaher 2022). Therefore, it comes as a surprise that it has generally not been adequately acknowledged and precisely defined by technical dictionaries and glossaries, especially as the vocabulary volume in technical and scientific texts is not as large as in GE texts, thus having higher frequency (density) of core vocabulary (Chung and Nation 2004; Kovalev et al. 2019; Kruse and Heid 2021). This may be the case because of frequently and ad hoc developed technical (often bilingual) specialized glossaries (e.g. of some medical, business or nautical terms) which are often not compiled by language professionals, thus not receiving significant attention from lexicographers (cf. Nkomo and Madiba 2011). The aim of this article is therefore to tackle two possible methods of computer-based headwords selection for a technical English, to contrast them, but also to combine the advantages of both.

The first method taken into account is a frequency count, used to identify the most frequent target vocabulary and, in that way, build a frequency-based word list. Frequency has been a primary criterion for headword selection, initially with GE dictionaries, glossaries and word lists. Following the needs of language learners who are at the same time professionals in technical areas, software tools have been developed for producing frequency vocabulary lists, starting with GE ones, but also for more and more of those related to specialized areas and pertaining to specific professional corpora.

Keyness, on the other hand, aims to detect the key vocabulary for a specific area by comparing its vocabulary frequencies with those in a reference GE corpus. The two methods are tested and discussed with reference to a professional corpus of marine engineering instruction books, with English for Marine Engineering Purposes (EMEP) generally considered extremely demanding, vocabulary-wise (Hsu 2014; Đurović et al. 2021). Adding to this the globality of the seafaring profession and the fact that English is the official means of communication of this complex discourse community, as formally established after World War II, technical vocabulary has been a mandatory requirement, but also a major challenge, for non-native speakers of English, as well as for language instructors. Specifically, this article stems from previous research on a word frequency list for marine engineering purposes. Thus, this study is a continuation with a general overview of the previous research and findings and, beyond that, we are presenting further investigation regarding applicable methodology and the possibilities of improvement when it comes to building effective ESP word lists.

2. Theoretical background and previous research

Analysing the most practical language needs of the target language learners — in this case, future and active marine engineers attending English for Marine Engineering Purposes courses — we embarked on the ambitious project of seeking the most effective and practical vocabulary tool(s) for technical lan-

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

guage learning, but also for the overall marine engineering profession. Furthermore, software solutions were tested that could assist in determining a practical and successful methodology.

The project started with the collection and selection of the most technical and professional marine engineering corpus, i.e. ship instruction books. In targeting these research objectives, expert advice and extensive teaching experience in the area was followed, but, even more importantly, the official requirements and recommendations set out by the International Maritime Organization (IMO) as the global standard-setting authority for the international shipping industry was also adhered to. The IMO's International Convention on the Standards of Training, Certification and Watchkeeping for Seafarers (STCW, Part 2.2) and the IMO Model Course 3.17 — Maritime English, notably the part on Specialised Maritime English dedicated to marine engineering courses of English, were particularly used as guidance. Apart from general communication skills in terms of using internal communication systems, the majority of the language skills requirements (about 90% of the anticipated course and self-study hours) are dedicated to "adequate knowledge of the English language to use engineering publications" (IMO Model Course 3.17 2015: 153). Guided by these clear instructions, the area of interest has been reading comprehension of marine engineering publications, specifically instruction books (Đurović 2021).

Following previous research findings and recommendations, the lexical profiling methodology and some of the most up-to-date software for the creation of a specialised marine engineering word list was applied (Đurović 2021). A frequency count was the starting point, which anticipates those words that appear most frequently across the corpora. Keyness, as a corpus linguistics method, on the other hand, refers to the frequency of the words in a special-purpose corpus compared to their frequency in a reference GE corpus. In addition to numerous other authors dealing with similar methodologies for building word lists (abundantly referred to in e.g. Archer et al. 2016 and Nation 2016), this article directly relies and builds upon the author's previous research with the same professional corpus, briefly presented below with reference to the marine engineering word frequency list.

2.1 Corpus

Following certain expert advice and experiences, primarily those of Chief Engineers, we sought to create a relevant selection of instructional engineering material of the utmost practical importance to marine engineers, ranging from familiarisation with a ship's systems and machinery, to regular maintenance, repairs and overhauling. The selection comprised technical manuals (most frequently referred to as instruction books) from a container ship, a tanker ship, a cruise ship and an offshore vessel. Additional material was added to enhance the diversity and bring the technologies up to date. The final corpus material comprises thousands of pages of electronic material related to ship machinery, devices and systems, converted, additionally "cleaned" and prepared (Nation 2016: 224) to accommodate the software requirements. The Corpus of Ship Instruction Books (CSIB) was, in this way, finalised with 1,769,821 running words (tokens). Bearing in mind the composition and size, we may say that our corpus is of representative importance to the discipline-specific genre so as to guarantee the validity of the results and conclusions produced (Đurović 2021). Further details on the corpus can be found in the author's referenced research article.

2.2 Word frequency lists

Modern research into "specialized or technical vocabulary has focused primarily on producing a word list of technical vocabulary in professional fields of expertise in English for Specific Purposes" (Coxhead and Demecheleer 2018; Đurović 2021). As both native speakers and language learners tend to acquire vocabulary according to its frequency, both in general language and in specialized areas of their interest or (business) activity (Nation 2006), frequency has been the main criterion for extracting core vocabulary.

The first corpus-based frequency word lists were, naturally, GE word lists (e.g. Fries and Traver 1950; West 1953). Since then, depending on the specific needs of the (English) language learners and non-native speaking language users, computer tools and methods have been developed to build specialized or technical word lists. Owing to the availability of electronic GE corpora and modern software possibilities, there has been a growing number of technical word lists aimed at early specialisation in the target professional and technical areas. The main presumption in the process is that the language learners have mastered at least 2,000–3,000 GE words, which are considered the most frequent GE words and expected to cover about 80% of texts (Nation 2006; Dang and Webb 2016; Van Zeeland and Schmitt 2013; Web and Rodgers 2009). Therefore, the designated software solutions, such as RANGE (Heatley et al. 2002) and, more recently, AntWordProfiler (https://www.laurenceanthony.net/software/antwordprofiler/), provide the possibility of eliminating the assigned word lists from being further counted. In the case of building a technical word list, the eliminated GE word lists would be those containing the 2,000–3,000 most frequent GE words (as those anticipated to have been already mastered). In determining the size of the list and the cut-off point in terms of the frequency count, the main purpose and evaluation criterion of the produced word lists is to reach the level of 95% (as adequate) and/or 98% (as the ideal threshold) of the text coverage, together with the 2,000–3,000 most frequent GE words (Dung and Web 2016; Nation 2016; Laufer 1992; Đurović 2021).

Although the notions of frequency and word lists can refer to other criteria, such as keyness, we will henceforth refer to word frequency lists as those built upon the frequency count only (usually accompanied by a determined cut-off point), and keyword lists will be the ones built upon the keyness metrics.

2.3 Keyness and keywords

The lexical units of a language are generally considered equal in status, but, when it comes to text, their significance and role vary (Bondi 2010). The new interest in "words" has been gaining in importance when it comes to lexical analysis of texts (genres) and related corpus linguistics areas of research. Generally, the notion of keywords has not been defined by official linguistics; keywords have rather earned by themselves their rising importance and attraction, especially in Englishes for Specific Purposes. First of all, authors, more or less intentionally or spontaneously, still use both written forms: *key word* as a collocation and/or *keyword* as a compound, as is the case in this article. Also, keywords have often been used as general, and more or less provisional, markers of "aboutness" and of the style of a text, e.g. papers and articles (Scott and Tribble 2006: 59). However, not many have been aware of the new possibilities of eliciting a statistically justified ranking of keywords, which is now available by using contemporary software solutions.

Unlike frequency counts, the keyness of a word does not necessarily anticipate a high, but rather an unusual, frequency of that word as compared to the general language — in our case General English. They are "key" because they capture the essence of particular types of discourses (Culpeper and Demmen 2015: 1). Their importance also signifies the cultural importance of lexical items (Li and Tarp 2022), as they relate in "culturally significant ways", and would "provide a representation of socially important concepts" (Gabrielatos 2018). The statistical software possibilities provide us with an insight into these particular words of "special status" (Stubbs 2010: 21). Based on their significantly increased frequency as compared to referent types of general texts, keywords point to the very nature of the text, i.e. the genre itself, and enable its easier comprehension (Baker 2004; Gabrielatos 2018). Keyness is one of the highlights in professional corpora such as, specifically, maritime genres. Generally, when there is a corpus of a very demanding lexical load — such as marine engineering publiccations — this demands special attention as regards adequate comprehension and mastering of specialised vocabulary. Therefore, our intention was to test both metrics criteria (frequency count and keyness) for eliciting specialised vocabulary to be focused on during language courses and professional work.

3. Methodology

The relevant methodology for a frequency count was based upon the use of the AntWordProfiler 1.4.0w software (https://www.laurenceanthony.net/software/antwordprofiler/) which is an upgraded version of the previously used RANGE programme (Heatley et al. 2002). Since the focus here is a highly technical branch of ESP, the general tendency and recommendation of e.g. Coxhead (2000), Hsu (2014), Yang (2015), Nation (2016), Kwary and Artha (2017) and Vuković Stamatović (2020), was followed to upgrade the first 2,000 or 3,000 GE words with specialised vocabulary lists, which together aim to reach the adequate

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

reading comprehension threshold in the most efficient way. The referent General English (GE) word lists used for the process were Nation's word lists produced from the British National Corpus and Corpus of Contemporary American English (BNC/COCA, https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists). The total of 25 lists contain about 1,000 word families each and, for this kind of research, they are usually accompanied by additional lists of the most frequent proper names, abbreviations, transparent compounds and marginal words developed by the same author (Nation 2004; 2006).

For formatting the lists into headwords only, lemmas, or expanding them into an all-family-members form, we used the Familizer + Lemmatizer program (https://www.lextutor.ca/familizer/). For the preparation of the corpus and converting it into "plain text" format, we used AntFileConverter (https://www.laurenceanthony.net/software/antfileconverter/). The detailed methodology and procedure for building a specialised word list (in this case a marine engineering one), as well as the final results, were meticulously presented in previous research (cf. Đurović 2021). Therefore, the specifics of the methodology and results will be briefly summarised here as is needed and as is relevant.

Considering the keyness metrics, this will be carried out using the AntConc software, version 3.5.8. (https://www.laurenceanthony.net/software/antconc/). Although there are more and more online programs offered for the selection of words, the methodology of frequency count lists and recognised scientific research based upon it was followed, including that of Laurence Anthony as the software designer. His software solutions were adopted since they are complementary and readily available. The advantage of these programs is that they are free-of-charge, they have been regularly updated, they provide comparable results, and can be used for any language.

Investigating the keyness of words is an important research method of corpus linguistics where comparative analysis is conducted between corpora, i.e. our target corpus and the reference one. As the comparison is usually done with a GE corpus, the FLOB corpus of contemporary British English was used. FLOB was created as a contemporary version of the Lancaster-Oslo/Bergen corpus (LOB) from 1961 and contains about a million words from various British genres. With the aim of creating an up-to-date British English corpus methodologically, resembling the Brown University Standard Corpus of Present-Day American English (Kučera and Francis 1967) in size, the Freiburg version of the LOB corpus was published by the University of Albert-Ludwig in Freiburg in 1999 under the acronym FLOB. Considering its up-to-dateness and the size being close to the target corpus, this GE corpus was opted for as the reference one. Comparing sufficiently large corpora of a similar size is convenient for ensuring similar frequency opportunities, thus providing for comparability of the results (Nation 2016).

The target methodology tested here is actually a "hybrid" one, aiming to combine the benefits of both ones mentioned above — the former based upon frequency and the latter focused on keyness. The intention is to provide a more comprehensive and effective word list that could still be attainable and practi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

cal for ESP (English for Marine Engineering Purposes — EMEP) classes and courses. The hope is to provide a solid recommendation for combined corpus linguistics methods applicable to other ESP areas and cases.

4. Marine engineering word frequency list

Aiming to provide our target learners of English for Marine Engineering Purposes (EMEP) with a practical vocabulary tool to help them reach an adequate reading comprehension text coverage of 95%, the methodology recommended by recognised authors from the area (e.g. those summarised in Nation 2016) were followed and applied for the purpose of comparison, evaluation and recognition. All the necessary decisions and interventions made on the way, as well as the specifics, limitations and further possibilities, were meticulously presented as a part of previous research (cf. Đurović 2021). Finally, a marine engineering word list of 337 word families was developed, accompanied by a list of 73 transparent compounds, which were derived from the corpus of marine engineering instruction books consisting of 1,769,821 running words. For practical reasons, the list is not provided in the addendum as it is readily available in the previous research cited. Nevertheless, it makes up an integral part of the final glossary list given in Addendum 3.

Since the produced word lists are evaluated by the adequate level of (cumulative) coverage in the target corpus, we are here briefly referring to the evaluation of our marine engineering word list (Table 1). Considering the specifics of the corpus and possibilities of extension, we refer to it here as the Word List of Ship Instruction Books (WLSIB).

Table 1: Coverage of Word List of Ship Instruction Books (WLSIB) in the corpus of marine engineering instruction books (Đurović 2021)

Word lists	Tokens	Coverage (%)
BNC/COCA 3,000 + proper nouns, abbreviations and marginal words	1,547,067	87.41
Transparent compounds	12,783	0.72
WLSIB without compounds	130,994	7.41
Outside of the lists	78,977	4.46
Total	1,769,821	100

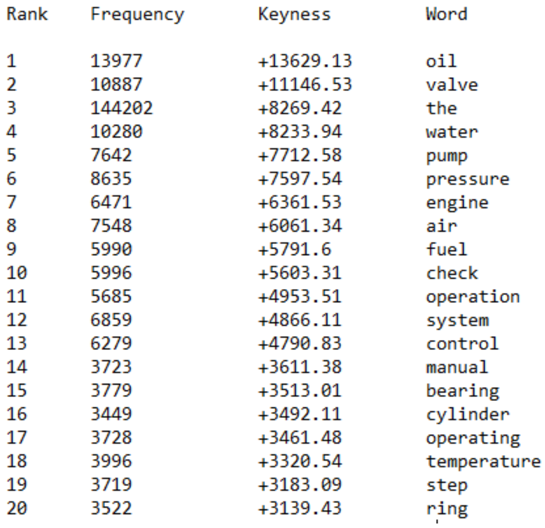
In total, together with the first 3,000 GE words, proper nouns, abbreviations and marginal words, the level of 95.54% (87.41% + 0.72% + 7.41%) is reached, thus reaching the goal of the adequate reading comprehension threshold, as recommended by Laufer (1989) and supported by e.g. Laufer and Ravenhorst-

Kalovski (2010) and Van Zeeland and Schmitt (2013). Taking into consideration that the desired level of coverage can be attained with no fewer than 12,000 general English words (only), as tested in Đurović et al. (2021), the final results perfectly fit the findings of Laufer and Ravenhorst-Kalovski (2010), by which the threshold of 95% is expected to be reached through the use of 4,000–5,000 word families (Đurović 2021).

Nevertheless, as human intervention and expertise are required and are indispensable throughout the process, we would not readily exclude the other valid criteria for vocabulary selection. We generally wanted to explore both criteria, compare them and possibly combine them to obtain more comprehensive results which would still be an attainable task both for students and trainees in marine engineering.

5. Keyword list

In reaching the positive evaluation of the list and the desired 95% coverage, we were further inspired by the possibilities of corpus linguistics in the selection of the most useful and most effective vocabulary for our target group of language learners. In the case of a specific technical corpus and genre of marine technical manuals, first we wanted to explore the keywords and the range of their keyness in terms of frequency when compared to General English (the FLOB corpus). For illustrative purposes, in Figure 1 the first 20 words are presented according to their keyness, i.e. the frequency ranking, resulting from comparison to their frequency in the reference GE corpus (FLOB) by means of the keyword metric.



Rank	Frequency	Keyness	Word
1	13977	+13629.13	oil
2	10887	+11146.53	valve
3	144202	+8269.42	the
4	10280	+8233.94	water
5	7642	+7712.58	pump
6	8635	+7597.54	pressure
7	6471	+6361.53	engine
8	7548	+6061.34	air
9	5990	+5791.6	fuel
10	5996	+5603.31	check
11	5685	+4953.51	operation
12	6859	+4866.11	system
13	6279	+4790.83	control
14	3723	+3611.38	manual
15	3779	+3513.01	bearing
16	3449	+3492.11	cylinder
17	3728	+3461.48	operating
18	3996	+3320.54	temperature
19	3719	+3183.09	step
20	3522	+3139.43	ring

Figure 1: Frequency and keyness of keywords in CSIB as compared to the reference FLOB corpus

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

From the table and numerical presentation of the results above it is clear that the frequency and keyness of the vocabulary are similar at the beginning, with keyness, as expected, having a far more rapid decrease in values than the frequency. Also, as expected, the keywords reflect the extremely technical character of marine engineering. Here it must also be noted that the AntConc programme presents results in the form of word types (not word families), as seen in the example of *operation* and *operating*, which are here given as separate units. This is another notion that should be borne in mind when combining methodologies and comparing the results.

In order to overcome the limitations of the AntConc programme in terms of the (im)possibility of using various word lists in the analysis, the initial keyword list obtained from ship instruction books (2,437 word types/1,172,171 tokens) were subjected to further analysis through the AntWordProfiler software. The convenience of the AntWordProfiler comes from its option to eliminate designated words/word lists from further counting and analysis. In that way the first 3,000 GE (BNC/COCA) words could be eliminated, as well as the lists of the most frequent proper names, abbreviations and marginal words (Nation 2004; 2006). Additionally, in order to obtain vocabulary that would be distinctive in relation to the obtained frequency list of marine engineering vocabulary, one of the lists assigned to the AntConc software was also our initially produced frequency list (WLSIB), including the obtained list of the most frequent transparent compounds (Đurović 2021). The newly obtained (additional) keyword list, accompanied by the list of key transparent compounds, was further analysed and "purged" of abbreviations and typos, converted into word families and supplemented by "unclassified" words, i.e. those not recognised by the programs, thus not automatically classed into family or lemmatised categories (e.g. *arrester*, *retighten*, *feedwater*). Furthermore, the previously obtained frequency list (WLSIB) was also supplemented with additional members of WLSIB word families that had been detected in the new keyword list, and the same was done with the three GE word lists and the lists of the most frequent proper names, marginal words and abbreviations. In particular, the initial WLSIB was supplemented by some words that excelled in terms of their keyness but that were "missed" by the related word families in the initial word frequency list, such as *actuation*, *igniter* and *emulsify*. Finally, a list of 124 marine engineering keywords (Addendum 1) and an addition of 43 key compounds (Addendum 2) was obtained. Considering the size of the list, we opted to keep it in full.

Owing to the results newly obtained through the application of this combined method seeking the benefits of both software solutions, a joint list was created that can serve as a glossary of marine technical manuals. In addition, following practical procedures generally favoured by engineers, all the words were placed into an integrated list arranged alphabetically (Addendum 3). The units presented are word families (again, for practical reasons of presentation) although lemmas are preferred and recommended when it comes to glossaries,

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

especially dictionaries. This would provide for a separate presentation of word types within a word family, e.g. *alter*, *alternate*, *alternator*. In particular, the expanded glossary word list initially comprised 1,500 units, but, as is usually the case with word list presentation, they are condensed into a word family list (Addendum 3). This can be further expanded to lemmas or all family members with adequate programmes, such as Familizer + Lemmatizer (Cobb 2018), as used here.

Finally, by integrating the frequency and keyness lists obtained from the Corpus of Ship Instruction Books, a total list of 577 words was created.

Table 2: The frequency and keyness word lists from ship instruction books

Word lists	Number of word families
WLSIB	337
Frequency list of transparent compounds	73
List of keywords	124
List of key transparent compounds	43
Total	577

6. Pedagogical and lexicographical implications

Taking into consideration that the total number of word families obtained through both criteria — frequency and keyness — including transparent compounds, is "only" 577 (Table 2), i.e. still below 1,000 (Nation 2004) or below 800, as is deemed a realistically attainable task for a language learning period of two years (Dang and Web 2016: 174), a glossary obtained this way could have a very practical application in ESP classes and courses, especially throughout one's professional career. Another reason for adding keywords would be that keywords would further reflect the style and specificity of the genre of ship instruction books and manuals. Therefore, both criteria should be considered and included in the optimisation of the produced technical vocabulary tool. In this way, there can be provided for the inclusion of all the keywords, i.e. the words that are the most specific ones for the marine engineering genre in comparison with the GE genres.

A glossary based upon such a word list can be monolingual or bilingual. Considering the globality of the seafaring profession and English as its lingua franca, bilingual variations would be the most useful and practical ones for marine engineers. Another advantage would be that, once formed in English, the glossary can be used in combination with any other language.

Furthermore, and if needed, the glossary can also be expanded by the first 3,000 BNC/COCA (GE) words, thus comprising the total vocabulary required

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

for adequate reading comprehension of marine engineering technical manuals. In addition, it can also be supplemented with lower-frequency technical words by lowering the initial cut-off point of 50 (Đurović 2021) in order to obtain an expanded glossary or dictionary foundation.

7. Discussion and limitations

In order to test the significance and validity of the integrated vocabulary list, i.e. the glossary of ship instruction manuals, its coverage was tested in our corpus in the same way we did for our primary WLSIB word list.

Table 3: Coverage of the glossary in the corpus of ship instruction books

Word lists	Tokens	Coverage (%)
BNC/COCA 3,000 + proper names, abbreviations and marginal words	1,547,071	87.41
Glossary list (with transparent compounds)	151,135	8.54
Outside of the lists	71,615	4.05
Total	1,769,821	100

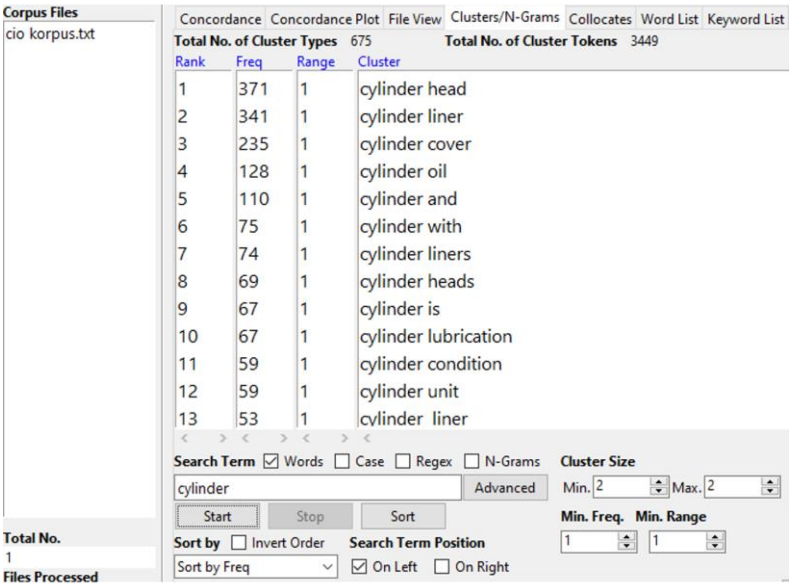
Based on the results presented in Table 3, a somewhat higher coverage of the glossary list (8.54%) was noticed as compared to the initial list with transparent compounds (8.13%). The total coverage of the glossary list with compounds is 95.95% (87.41% + 8.54%), which exceeds the coverage of WLSIB with transparent compounds by 0.41% (or about 7,500 corpus words) (see Table 1). It was expected not to make a drastic difference, based upon earlier research findings and conclusions related to the length of the word lists.

From the results, we can also confirm an earlier determined regularity whereby an additional extension to the list, i.e. inclusion of additional words with decreasing frequencies, is also accompanied by a rapid decrease in additional coverage in the corpus (Dang et al. 2017; Coxhead 2018; Nation 2016; Zipf 1935; 1949). Considering that in our specific case the difference does not significantly affect the final results, the initial methodology can be supplemented with this "hybrid" model. It would include additional key vocabulary as compared to the reference general one and would therewith enrich the frequency list and upgrade it to a more comprehensive and effective one.

In addition, we are well aware that words do not hold standalone meaning, but acquire their meaningfulness through combinations with other words. Therefore, collocations, n-grams and similar word combinations have been a recurrent topic of interest for ESP learners and instructors (Chen 2022). For this

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

purpose, the AntConc software can further be used for examining word relations such as collocations (and/or n-grams), which can be of use for additional development of glossary and dictionary items (Figure 2).



Corpus Files
cio korpus.txt

Total No.
1
Files Processed

Total No. of Cluster Types 675 Total No. of Cluster Tokens 3449

Rank	Freq	Range	Cluster
1	371	1	cylinder head
2	341	1	cylinder liner
3	235	1	cylinder cover
4	128	1	cylinder oil
5	110	1	cylinder and
6	75	1	cylinder with
7	74	1	cylinder liners
8	69	1	cylinder heads
9	67	1	cylinder is
10	67	1	cylinder lubrication
11	59	1	cylinder condition
12	59	1	cylinder unit
13	53	1	cylinder liner

Search Term ☑ Words □ Case □ Regex □ N-Grams
cylinder
Cluster Size Min. 2 Max. 2
Min. Freq. 1 Min. Range 1
Sort by □ Invert Order
Sort by Freq
Search Term Position ☑ On Left □ On Right

Figure 2: AntConc presentation of collocation and n-grams search

With the goal of building an effective and comprehensive, but still practical, vocabulary list for the target language learners (future and active marine engineers), we tested, contrasted, and combined the two methodologies presented above. On the way, the recommendations and previous findings and experiences we followed, aiming to contribute to the final product, but also to possible future methodologies. As metrics tools for the purpose, two software solutions developed by Laurence Anthony were mainly used: AntWordProfiler and AntConc. Lexical profiling using the AntWordProfiler software provides us with accurate information on the lexical characteristics and load of the target corpus. This software surpasses the AntConc software (and numerous others) in one very important aspect. It provides the opportunity to eliminate the available lists from further analysis, but it also measures the coverage of each list, as well as their cumulative coverage in the corpora. The keyness method (AntConc), however, provides us with the most specific vocabulary for the particular type of text by counting not the simple frequency, but rather the unusual frequency as compared to General English. As we can see in the example of our corpus, it provided us with additional specific technical vocabulary (e.g. *alloy*, *funnel*, *plumbing*, etc.) which do not belong to the most frequent English vocabulary (Nation 2004) but would certainly come in handy for marine engineers who are

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

non-native English speakers. One more reason for expanding the initial frequency list is the fact that the additional keyword list is not too long and the joint glossary list totals 577 (head)words (Table 2, Addendum 3), which is considered an attainable task for ESP courses (Dang and Webb 2016). On the other hand, when building a keyword list, the AntConc software does not have the option to eliminate any other list members from the count, thus by itself it cannot serve to upgrade the existing word lists. This is the reason why, for example, the AntWordProfiler was used with the obtained keyword list so that the members of the first GE word lists and the WLSIB frequency lists could be eliminated, which provided additional vocabulary only (with no repetition or overlapping).

Still, however statistically accurate the count, the methodology "does not work" without human expertise and intervention. In the software processing, especially in the case of technical corpora, there are always some "unrecognised" words which are presented as unclassified. They also need to be focused on and added to a certain word family or word list or eliminated in the case of a typo or similar error. Additional attention should also be paid to various spelling options coming from different publications (e.g. *manoeuvre* vs. *maneuver, minimise* vs. *minimize*, etc.) and putting them into the same families with cumulative frequency.

Further analysis of the list(s), as more or less statistical products, would open up some new possibilities and questions, such as those of a semantic nature. The software solutions do not recognise polysemous or cryptotechnical words, as referred to by Fraser (2009), which we should especially have in mind in the case of building a bilingual glossary. Another phenomenon has also been confirmed here, and that is the fact that the most frequent content words are also the most polysemic ones (Ravin and Leacock 2000). These words can formally belong to the first 3,000 GE words, but gain new meanings in marine engineering, either individually or in collocations. As recommended by previous research findings, they were added to the GE word families, although special attention should be paid to them in language courses. This also relativises the statistical results, as their frequency is added to the GE words only.

There is always a possibility of including, at least partially, the most frequent GE words that the ESP list has been built upon. In the example of the target corpus, words such as *actuator*, *mess*, *skirt*, *pin* and similar have been added to the first 3,000 BNC/COCA word families, as suggested by the established methodology. However, in English for Marine Engineering, the terms relate to specific parts of a ship or propulsion machinery and can have various translations in different languages. This goes a step further with collocations. For example, *arm* and *rock* also belong to the most frequent GE words. However, *rocker arms* are very important parts of the valve opening/closing mechanisms that have different translations in different languages. Thus, here again, the importance of expert intervention and attentiveness must be emphasised, regardless of the detailed methodology and previous recommendations and findings.

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

As concerns the corpus selection, another highlight should be put forward here. When building a word list, it is always related and should refer to the specific corpus. In this case, it was the professional corpus of marine engineers. However, in other possible research that could be, for example, dedicated to English language learners undertaking marine engineering studies in English, the corpus could comprise marine engineering course books in English, or they could be an upgrade to the corpus of ship instruction books, which may, again, be of different compositions and sizes.

8. Conclusion

Aiming to provide the target language learners with a concentrated and specialised word list elicited from their professional corpus of marine engineering publications, we initially followed the established methodology for building a word list from ship instruction books and manuals. This specialised word list comprised 348 words with 75 transparent compounds. Tempted to try out other criteria, specifically the keyness of words in the texts, we explored the possibilities of combining the two methodologies, building on the benefits of one over the other, while at the same time overcoming the limitations of both. Ultimately, we finalised our recommended glossary list at a total of 577 words (word families), which is still sufficiently practical in size to be used in EMEP courses or in building effective bilingual glossaries for members of this challenging discourse community with English as a non-native language.

In building technical vocabulary lists, we have been able to attest that the AntWordProfiler software especially comes in handy since it provides us with the opportunity to exclude lists of the most frequent GE words (or any other type of word list) from further processing, thus focusing the frequency count on the technical vocabulary to be mastered in order to reach the adequate (reading) comprehension of professional genre(s). On the other hand, it can also be used (and we wholeheartedly recommend it) with the keywords list obtained through AntConc in order to obtain additional vocabulary of key significance to ESP learners and users. In this way, the methodologies build upon each other and, if attentively conducted, the resulting word list can serve as a good basis for a glossary that can be made bilingual in combination with any other language. We hope that the presented methodology and exemplar results can inspire other researchers and ESP teachers to use them either individually, or in combination, as presented here. Furthermore, the methodologies can also be further extended to build frequency and/or keyness dictionaries.

References

Abdelzaher, E.M. 2022. An Investigation of Corpus Contributions to Lexicographic Challenges over the Past Ten Years. *Lexikos* 32: 162-179. https://doi.org/10.5788/32-1-1714.

Anthony, L. 2014. *AntWordProfiler* (Version 1.4.1.0) [Computer Software]. Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software/antwordprofiler.

Anthony, L. 2017. *AntFileConverter* (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

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

Archer, D. (Ed.). 2016. *What's in a Word-list? Investigating Word Frequency and Keyword Extraction.* London: Routledge. https://doi.org/10.4324/9781315547411.

Baker, P. 2004. Querying Keywords: Questions of Difference, Frequency and Sense in Keywords Analysis. *Journal of English Linguistics* 32(4): 346-359.

Bondi, M. 2010. Perspectives on Keywords and Keyness. Bondi, M. and M. Scott (Eds.). 2010. *Keyness in Texts*: 1-18. Amsterdam/Philadelphia: John Benjamins.

Chen, Y. 2022. The Effect of Learning Conditions on Collocation Gains: A Case Study of Task-based Dictionary Use Instruction. *Lexikos* 32: 1-30. https://doi.org/10.5788/32-1-1679.

Chung, T.M. and I.S.P. Nation. 2004. Identifying Technical Vocabulary. *System* 32(2): 251-263. https://doi.org/10.1016/j.system.2003.11.008.

Cobb, T. 2018. From Corpus to CALL: The Use of Technology in Teaching and Learning Formulaic Language. Siyanova-Chanturia, A. and A. Pellicer-Sanchez (Eds.). 2018. *Understanding Formulaic Language: A Second Language Acquisition Perspective:* 192-211. New York: Taylor & Francis.

Coxhead, A. 2000. A New Academic Word List. *TESOL Quarterly* 34(2): 213-238.

Coxhead, A. 2018. *Vocabulary and English for Specific Purposes Research: Quantitative and Qualitative Perspectives.* London/New York: Routledge.

Coxhead, A. and M. Demecheleer. 2018. Investigating the Technical Vocabulary of Plumbing. *English for Specific Purposes* 51: 84-97.

Culpeper, J. and J. Demmen. 2015. Keywords. Biber, D. and R. Reppen (Eds.). 2015. *The Cambridge Handbook of English Corpus Linguistics*: 90-105. Cambridge: Cambridge University Press.

Dang, T.N.Y. and S. Webb. 2016. Making an Essential Word List for Beginners. Nation, I.S.P. (Ed.). 2016. M*aking and Using Word Lists for Language Learning and Testing*: 153-167. Amsterdam: John Benjamins.

Dang, T.N.Y., A. Coxhead and S. Webb. 2017. The Academic Spoken Word List. *Language Learning* 67(4): 959-997.

Đurović, Z. 2021. Corpus Linguistics Methods for Building ESP Word Lists, Glossaries and Dictionaries on the Example of a Marine Engineering Word List. *Lexikos* 31: 259-282. https://doi.org/10.5788/31-1-1647.

Đurović, Z., Vuković Stamatović, M. and M. Vukičević. 2021. How Much and What Kind of Vocabulary do Marine Engineers Need for Adequate Comprehension of Ship Instruction Books and Manuals? *Círculo de lingüística aplicada a la comunicación* 88: 123-133. https://dx.doi.org/10.5209/clac.78300.

Fraser, S. 2009. Breaking Down the Divisions Between General, Academic, and Technical Vocabulary: The Establishment of a Single, Discipline-based Word List for ESP Learners. *Hiroshima Studies in Language and Language Education* 12: 151-167.

Fries, C.C. and A.A. Traver. 1950. *English Word Lists*. Ann Arbor: George Wahr.

Gabrielatos, C. 2018. Keyness Analysis: Nature, Metrics and Techniques. Taylor, C. and A. Marchi (Eds.). 2018. *Corpus Approaches To Discourse: A Critical Review*: 225-258. Oxford: Routledge.

Hanks, P. 2020. English Dictionaries and Corpus Linguistics. Ogilvie, S. (Ed.). 2020. *The Cambridge Companion to English Dictionaries*: 219-239. Cambridge: Cambridge University Press.

Heatley, A., I.S.P. Nation and A. Coxhead. 2002. Vocabulary Analysis Programs. The Range Program (online) https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs.

Hsu, W. 2014. Measuring the Vocabulary Load of Engineering Textbooks for EFL Undergraduates. *English for Specific Purposes* 33: 54-65.

IMO. 2015. *Maritime English*. Model Course 3.17. London: International Maritime Organization.

IMO. 2017. *STCW Convention and STCW Code*. London: International Maritime Organization.

Kovalev, I.V., S. Yu Piskorskaya, M.V. Karaseva and A.A. Voroshilova. 2019. Computer-aided Approach to Synthesis of the Frequency Dictionary on System Analysis in Electronic Machinery, Aviation and Space Industry. *IOP Conference Series: Materials Science and Engineering* 537(4): 1-7. https://doi.org/10.1088/1757-899X/537/4/042085.

Kruse, T. and U. Heid. 2021. Lemma Selection and Microstructure: Definitions and Semantic Relations of a Domain-Specific e-Dictionary of the Mathematical Field of Graph Theory. Gavriilidou, Z., M. Mitsiaki and A. Fliatouras (Eds.). 2021. *Proceedings of the XIX Euralex International Congress, Alexandroupolis, Greece, 7–11 September 2021. Volume 1:* 227-233. Komotini, Greece: European Association for Lexicography.

Kučera, H. and W.N. Francis. 1967. *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Kwary, D.A. and A.F. Artha. 2017. The Academic Article Word List for Social Sciences. *MEXTESOL* 41(4): 1-11.

Laufer, B. 1989. What Percentage of Text-Lexis is Essential for Comprehension? Lauren, C. and M. Nordman (Eds.). 1989. *Special Language: From Human Thinking to Thinking Machines:* 316-323. Clevedon: Multilingual Matters.

Laufer, B. 1992. How Much Lexis Is Necessary for Reading Comprehension? Arnaud, P.J.L. and H. Bejoing (Eds.). 1992. V*ocabulary and Applied Linguistics:* 129-132. London: Macmillan.

Laufer, B. and G.C. Ravenhorst-Kalovski. 2010. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language* 22(1): 15-30.

Li, Q. and S. Tarp. 2022. Differentiated Treatment of Cultural Items in Lexicographical Products: A Necessary Adaptation to the Digital Environment. *Lexikos* 32: 90-117. https://doi.org/10.5788/32-1-1706.

Nation, I.S.P. 2004. A Study of the Most Frequent Word Families in the British National Corpus. Bogaards, P. and B. Laufer (Eds.). 2004. *Vocabulary in a Second Language: Selection, Acquisition and Testing*: 3-13. Amsterdam: John Benjamins.

Nation, I.S.P. 2006. How Large a Vocabulary is Needed for Reading and Listening? *Canadian Modern Language Review* 63(1): 59-82.

Nation, I.S.P. 2016. *Making and Using Word Lists for Language Learning and Testing.* Amsterdam: John Benjamins.

Nkomo, D. and M. Madiba. 2011. The Compilation of Multilingual Concept Literacy Glossaries at the University of Cape Town: A Lexicographical Function Theoretical Approach. *Lexikos* 21: 144-168. https://doi.org/10.5788/21-1-41.

Ravin, Y. and C. Leacock (Eds.). 2000. *Polysemy: Theoretical and Computational Approaches*. Oxford: Oxford University Press.

Scott, M. and C. Tribble. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education.* Philadelphia: John Benjamins.

Sinclair, J. 1992. *Collins COBUILD English Language Dictionary*. London: HarperCollins.

http://lexikos.journals.ac.za; https://doi.org/10.5788/33-1-1807 (Article)

Stubbs, M. 2010. Three Concepts of Keyness. Bondi, M. and M. Scott (Eds.). 2010. *Keyness in Texts*: 21-42. Amsterdam/Philadelphia: John Benjamins.

Van Zeeland, H. and N. Schmitt. 2013. Lexical Coverage in L1 and L2 Listening Comprehension: The Same or Different from Reading Comprehension? *Applied Linguistics* 34(4): 457-479. http://doi.org/10.1093/applin/ams074.

Vuković Stamatović, M. 2020. Vocabulary Complexity and Reading and Listening Comprehension of Various Physics Genres. *Corpus Linguistics and Linguistic Theory* 16(3): 487-514. https://doi.org/10.1515/cllt–2019–0022.

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

Yang, M.-N. 2015. A Nursing Academic Word List. *English for Specific Purposes* 37: 27-38. https://doi.org/10.1016/j.esp.2014.05.003.

Zipf, G.K. 1935. *The Psycho-biology of Language*. Boston: Houghton Mifflin.

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley Press.

Addendum 1: List of additional keywords from ship instruction books (without transparent compounds)

acetylene
adhere
affixed
alloy
armature
arrester
ascertain
baffle
beverage
breather
buffet
bulb
burrs
buzzer
chassis
chock
circlip
citric
clicking
coalescer
compensators
conformity
contactor
coupler
deficiencies
descaling
dew
diaper
diarrhea
diffuser
dimensioned
diode
disengaged
disulfide
dongle
dowel
duplex
elapsed
elysator
emery
encoder
erection
expel
extern
finned
fountains
funnel
galvanize
garbage
goggles
grams
graphite
grease
hydrazine
hydroxide
hysteresis
illuminate
increment
ingress
inhalation
insoluble
ion
kerosene
lamellar
lapping
lateral
locker
lowing
magnifying
malfunctions
mandrel
micron
modulating
molluscan
mop
nitrite
notification
ohm
opacity
osmosis
pallet
pantries
pentane
pertaining
pest
plumbing
portable
poultry
prerequisites
radiator
readout
receptacle
recess
reciprocating
rectify
refrigerant
reportable
retention
ridges
scuffing
scum
serum
setter
shim
sling
slotted
slushing
sterilizer
strap
swabs
swirler
synopsis
tapered
tappet
Teflon
template
turbidity
unitor
vanadium
vapor
Vaseline
vomit
wedges
wobb

Addendum 2: List of additional key transparent compounds from ship instruction books

aftmost
backpressure
backup
backwash
burnertype
carryout
checkbag
convertbox
deckhead
download
downtime
feedwater
flowmeter
foodborne
foreword
framebox
gearbox
gearwheel
hereby
inline
logout
lowermost
manhole
microorganisms
overpressure
pushbutton
salinometer
shellstock
shipyard
startstop
startup
staybolts
testbed
tubesheet
undercooked
underside
undersize
upwards
usefor
wastewater
website
workcard
worktable

Addendum 3: Glossary list of headwords from ship instruction books

aboard
abrasive
accessory
accord
acetylene
acid
actuate
acute
adhere
adhesive
adjacent
adsorb
affixed
aft
aftmost
align
alkaline
alloy
alternate
aluminum
ambience
amplify
analog
annex
annular
anode
anti
appendix
appliance
armature
arrester
arrow
ascertain
ash
assemble
astern
automate
automobile
auxiliary
axis
backflow
backpressure
backup
backwash
baffle
barge
barrel
batch
bedplate
bellow
beverage
bilge
blade
blink
bolt
bonnet
bracket
brass
breakdown
breather
bronze
buffer
buffet
bulb
bulkhead
bunker
burnertype
burrs
buzzer
bypass
cabin
calibrate
calorific
cam
camshaft
carrieout
cartridge
casing
caterpillar
caution
cavity
centrifuge
centripetal
certify
changeover
chassis
checkbag
chlorine
chock
circlip
citric
clamp
classification
clicking
clockwise
clog
clutch
coalescer
cock
coil
collar
combustion
communicable
compartment
compatible
compensators
comply
compress
con
condense
cone
configure
conformity
console
contactor
contaminate
contouch
convertbox
copper
copyright
corrode
countdown
coupler
crane
crank
crankcase
crankpin
crankshaft
crankthrow
crosshead
crosswise
cylinder
datalogger
debris
decant
deckhead
default
defect
deficiencies
deflect
deform
descaling
detergent
deteriorate
deviate
dew
diagnosis
diagram
dial
diameter
diaper
diaphragm
diarrhea
diesel
differential
diffuser
digit
dimensioned
din
diode
dip
dipstick
discard
discrete
disengaged
disinfect
dismantle
dispense

displace
dissolve
distillate
disulfide
dongle
dowel
download
downstream
downtime
drip
droop
dual
duct
duplex
durable
duration
dynamic
effluent
eject
elapsed
electrode
elysator
emery
emulsion
enclose
encoder
erection
erosion
evaporate
ex
expel
expire
extern
eyebolts
fasten
fax
fecal
feedback
feedwater
finned
fixture
flange
flap
flotation
flue
fluid
flush
flywheel
foodborne
fore
foreword
foul
fountains
framebox
freshwater
friction
funnel
furnace
fuse
galvanize
garbage
gasket
gastroenteritis
gastrointestinal
gauge
gearbox
gearwheel
generator
geometry
gland
globe
glove
goggles
grams
flue
fluid
flush
flywheel
foodborne
fore
foreword
foul
fountains
framebox
freshwater
friction
funnel
furnace
galvanize
garbage
gasket
gastroenteritis
gastrointestinal
gauge
gearbox
gearwheel
generator
geometry
gland
globe
glove
goggles
grams
graphite
grease
grease
grind
groove
gudgeon
halogen
hammer
handwashing
handwheel
harness
hereby
hexagon
hoist
hood
horizontal
hose
hub
hull
humid
hydraulic
hydrazine
hydroxide
hysteresis
icon
identical
idle
ignite
illuminate
impulse
incinerate
increment
inert
ingress
inhalation
inhibit
inlet
inline
insoluble
insulate
intact
intake
integral
intercept
interface
interlock
intermediate
interval
ion
jacked
kerosene
keyboard
kit
knob
lamellar
lance
lapping
lateral
layout
lever
linear
linen
liner
liter
locker
login
logout
loop
lowermost
lowing
lube
lubricate
magnifying

flowmeter
mandrel
manhole
manifold
maneuver
manometer
membrane
mesh
micro
micron
microorganisms
millimeters
mineral
minimize
mist
modulating
moisture
molluscan
molybdenum
mop
mount
multi
nameplate
needle
nipple
nitrite
node
nominal
notification
nozzle
offset
ohm
onboard
opacity
optimum
opus
orifice
osmosis
outbreak
outlet
overboard
overflow
overhaul
overheating
fuse
overpressure
override
overspeed
overview
oxidation
oxygen
pallet
pantries
parameter
particle
password
paste
pentane
permissible
pertaining
pest
pinion
pipelines
piston
pliers
plumbing
plunge
pneumatic
polyamide
polymer
portable
potable
potentiometer
poultry
precaution
preface
preliminary
prerequisites
prescribe
preset
prolong
propel
propulsion
proximity
pulley
pulse
puncture
purge
impel
rack
radial
radiator
ram
ramp
readout
receptacle
recess
reciprocating
recreation
rectify
refract
refrigerant
relay
reportable
residue
resilience
retention
ridges
rim
rinse
rod
rotate
rubber
rudder
rust
saline
salinometer
sanitize
satisfactory
sauer
scavenge
scrape
screwdriver
scrubber
scuffing
scum
seawater
seizure
selfjector
sensor
serial
serum
malfunctions
setpoint
setter
setup
shaft
shellfish
shellstock
shim
shipbuilding
shipyard
shutdown
silicon
sketch
sleeve
sling
slotted
sludge
slushing
socket
sodium
soiled
solenoid
solvent
soot
sootblower
spa
span
spanner
spark
specimen
spindle
spiral
splash
spool
stack
standby
standstill
startstop
startup
static
staybolts
steer
sterilizer
stool

overlay
overload
stuffed
suction
sulphur
sump
surge
surveillance
swabs
swirler
synopsis
synthetic
tab
tag
tapered
tappet
Teflon
telescope
template
terminal
testbed
thermal
pushbutton
quarantine
thermometer
thermostat
thread
threshold
throttle
throughput
thrust
tiller
tilt
tolerance
torch
torque
torsion
touchscreen
toxic
transducer
troubleshooting
tubesheet
turbidity
turbine
servo
servomotor
turbocharger
tween
undercooked
underside
undersize
unitor
upstream
uptake
upward
upwards
usage
usefor
utensil
vacuum
valve
vanadium
vane
vapor
Vaseline
velocity
strap
stud
ventilate
verify
vertical
vibrate
viscous
volt
vomit
warewashing
warranty
wastewater
website
wedges
weld
whirlpool
wobble
workcard
worktable
wrench
yoke