Corpus-based Approach to Developing Teaching Materials for Aerospace English

Andrey Sergeevich Korzin [a]
korzin_as@pfur.ru
Department of Foreign Languages of the Academy of Engineering
Peoples' Friendship University of Russia (RUDN University), Russian Federation

Anna Sergeevna Zhandarova
annazhandarova@mail.ru
Faculty of Romanic and Germanic Languages
Moscow Region State University, Russian Federation

Yana Aleksandrovna Volkova
volkova-yaa@rudn.ru
Department of Foreign Languages in Theory and Practice
Peoples' Friendship University of Russia (RUDN University), Russian Federation

ABSTRACT

It is widely known that academic English is used for specific purposes in cross-cultural communication between scientists. Simultaneously, there is a shortage of teaching materials, leading to a demand for the development of such materials. A remote-sensing field was chosen for this study. This study describes the results of a corpus-based analysis of academic vocabulary in remote sensing articles. The research was conducted using corpus linguistics methods and distributive statistical analysis, and a corpus manager, Sketch Engine, was used as a tool to process a large amount of data. This study used a corpus compiled from academic papers published between 2020 and 2022. The frequency of lexical units was extracted to analyse the coverage of Academic Word List Sublist 1 in the corpus; keywords, multi-word units, and word formation were also analysed in this study. Units from two remote sensing glossaries were retrieved from the corpus to analyse how often they occurred in the corpus. Corpus linguistic methods and distributive statistical analysis proved effective in creating a discipline-specific shortlist that can be used by educators, ESP learners, and authors in the field of remote sensing. Despite the narrow field coverage of this study, the results obtained can be applied to general academic English vocabulary and to further research in the field of ESP.

Keywords: English for Specific Purposes; English for Academic Purposes; Academic Word List; Remote Sensing; Corpus; Terms

INTRODUCTION

The aerospace field has been developing over the past 20 years. For example, in 2016, the aerospace industry grew by 12.9% compared with 2015 (Najmon et al., 2019). Reaction Engines Limited launched a project aimed at developing a single-stage-to-orbit space plane (Petrescu et al., 2017). The development of this field has led to economic contracts such as the ESA-ISA

[a] *Corresponding author*

eISSN: 2550-2131
ISSN: 1675-8021

agreement, which illustrates cooperation in the aerospace field between Europe and Israel (Barok, 2013).

At the same time, there is a lack of vocabulary and teaching material in this area. This is due to the fact that the development of aerospace technology is several years ahead of the publication of terminological dictionaries. Moreover, in specialised dictionaries, no new low-frequency terms are accepted by specialists (Paltridge & Starfield, 2013).

ESP, or English for Specific Purposes, is a specialised branch of English language teaching that focuses on learners' language needs in a particular field or profession. Unlike General English courses, ESP courses are designed to meet learners'specific language demands in their professional or academic contexts (Hutchinson, 1987). ESP is particularly useful for learners who need English for work or academic purposes, as it helps them develop the language skills they need to succeed in their particular field (Johns & Dudley-Evans, 1991). Since learning objectives are highly specific, ESP teachers face various challenges when teaching specialised subjects.

Technical terminology can be complex and challenging for both teachers and learners, especially if they are unfamiliar with the subject matter. Regarding Earth remote sensing, teachers need to ensure that learners understand key terms and concepts, such as *satellite imagery, spectral bands, and image resolution* (Musikhin, 2016). As mentioned earlier, there is a shortage of teaching materials in many ESP fields, including Earth remote sensing. This presents a challenge for ESP teachers, who need to develop their own teaching materials or adapt existing materials to suit their learners' needs. For example, numerous attempts have been and are being made to create word lists for rather narrow professional fields using the corpus-based approach (Valipouri & Nassaji, 2013; Csomay & Petrović, 2012; Lei & Liu, 2016; Roesler, 2021; Muñoz, 2015). Such contributions are of great significance for ESP and EAP teachers and material developers and should be taken into consideration, especially when designing a job-specific course or a textbook. ESP teachers also struggle to find authentic materials that accurately reflect the language used in real-world contexts. In case of Aerospace English and Earth remote sensing, it may be difficult to find relevant and up-to-date academic articles and other resources that learners can use (Tevdovska, 2018; Vora, 2017). Finally, ESP teachers may struggle with time constraints while designing and delivering their lessons. They need to ensure that they cover the essential language skills and technical knowledge while allocating enough time for learners to practice and develop their language skills (Poedjiastutie, 2017; Stojković, 2018).

Overall, ESP teaching requires specialised knowledge, creativity, and adaptability from ESP teachers. They need to address these challenges effectively to ensure that their learners acquire the necessary language skills to communicate in their field (Laborda & Litzler, 2015). Richards (1974) pointed out that a word list is a list of words arranged according to the frequency of their occurrence in the text. Word lists can be useful for vocabulary learning, because they provide the most frequent lexical units. In particular, they are beneficial in providing vocabulary for aerospace English. Aerospace can be considered a specific field with its own terminology, and word lists serve as a representation of this terminology. Students and teachers can use word lists as material in ESP classes, and scientists can utilised them while writing papers (Richards, 1974).

This article focuses on developing word lists for a particular area of "Aerospace English", Earth Remote Sensing (RS). Aerospace English is a specialised form of English designed for use in the aerospace industry. It is an important communication tool used by pilots, air-traffic controllers, engineers, scientists, and other professionals in the field. Remote sensing is a rapidly growing field of aerospace technology that has a significant impact on environmental monitoring, natural resource management, and disaster response, and is one of the areas where the need for discipline-specific teaching materials is particularly urgent (Yakushev et al., 2019). The international nature of the aerospace industry contributes to a huge demand for learning materials in English, including those that provide insight into technical details and specific characteristics of the subject under study (Dvoryadkina & Mikheeva, 2018; Moraño-Fernandez et al, 2019; Lukianenko & Vadaska, 2020). Although the importance of English proficiency in the aerospace industry is widely recognised, teaching Aerospace English poses unique challenges for English for Specific Purposes (ESP) teachers (Netikšienė, 2006).

The aim of study is to present the results of a corpus-based analysis of academic vocabulary in remote sensing articles related to Earth observation, with the goal of identifying and creating a discipline-specific word list of academic vocabulary items that can be used by ESP teachers, learners, and authors in the field of remote sensing.

By completing these tasks, the authors aim to contribute to the development of teaching materials and resources for ESP teachers, learners, and authors in the field of remote sensing while also highlighting the importance of discipline-specific vocabulary in academic writing and communication.

LITERATURE REVIEW

Johns and Dudley-Evans discussed the history and importance of English for Specific Purposes (ESP). The authors point out that ESP pursues the goal of teaching English to adult learners for specific professional purposes, such as business and technology. According to Johns and Dudley-Evans, ESP requires careful research and design of pedagogical materials. This approach has rapidly gained popularity due to its effectiveness in meeting learners'specific needs, such as communication in professional domains (Johns & Dudley-Evans, 1991).

The needs of students are collected through analysis and observation of students during classes, which helps teachers to identify students' communication targets. Consequently, teachers can provide specific language instruction to help students succeed in their courses and future careers (Benesch, 1996; Belcher, 2006). Methods of corpus linguistics are commonly used for developing word lists for multiple fields. For example, Le and Miller identified frequently occurring medical morphemes to create a concise list for students (Le & Miller, 2020). The authors identified 344 frequently occurring morphemes in medical literature. Lexical units were identified using Sketch Engine. Moreover, the study provides a basis for designing vocabulary learning and teaching activities (ibid.).

Bi examines the vocabulary needs of Chinese computer science undergraduate students and builds a Computer Science Vocabulary List (CSVL) of 356 word families frequently used in computer science textbooks. Researcher suggests that targeted word lists are more effective for learners and that teachers should raise students' awareness of how words typically collocate in the context (Bi, 2020). Veenstra and Sato focused their study on the creation of the Science Textbook Word List (STWL) for undergraduate students studying science and engineering. The researchers attempted to prove the effectiveness of STWL against the Academic Word List and the Coxhead and Hirsh Science Word List. The study found that the STWL provided better coverage of the studied corpus than the AWL and Coxhead and Hirsh's science word list (Veenstra & Sato, 2018). Safari conducted an analysis of 3.6 million lexical units in the Equine Veterinary Corpus (EVC) in order to identify highly frequent words in the equine veterinary sub-discipline. The researcher aimed to develop a list of the most important words in the equine veterinary subdiscipline (Safari, 2019). Hsu provides an analysis of the vocabulary demands of compulsory engineering textbooks and proposes an Engineering English Word List. According to the author, engineering textbooks require a vocabulary within the range of the most frequent 5000-word families at 95% lexical coverage (Hsu, 2014). Another word list was created by Ward, who introduced a 299-word list called BEL for engineering students (Ward, 2009). Similar research was conducted by Ng et al. and Dang and Webb. The authors pointed out that the lexical threshold for successful reading comprehension is set at 95 percent, and the ideal coverage of vocabulary needed for dealing with any written text is 8,000-to 9,000-word families (Ng et al., 2020; Dang & Webb, 2014).

Word formation in academic English plays a pivotal role for learners as they can expand their vocabulary using or being familiar with key patterns. According to Abeyweera, word formation elements can be divided into 3 groups: prefixes, suffixes, neoclassical elements and phonologically neutral suffixes (Abeyweera, 2021).

METHODOLOGY

CORPUS-BASED APPROACH

This study used a corpus-based approach and a distributive statistical analysis. Zakharov pointed out that corpus linguistics includes applying linguistic corpora to test hypotheses or theories (Zakharov, 2015). This allowed obtaining the frequency of the use of lexical units in the corpus. The latter illustrates the distribution of words in a collection of documents and is associated with linguistic statistics. Distributive statistical analysis was used to assess the degree of semantic interrelation in the corpus (ibid.).

DATA DESCRIPTION

Sketch Engine is a corpus manager that was developed not only to generate concordances, but also to analyze metadata. This corpus manager can regroup documents according to extralinguistic factors and allows analysis to be performed based on the metadata attributes of each file. The size of the Remote Sensing Academic corpus (RSA) contains 999,812 words (1403398 tokens), and it was tagged with the Tree Tagger tool. This tool is related to machine learning and belongs to an unsupervised learning class with an inductive program, as it learns on untagged text and creates a tagset. Morphological tagging was performed as a basis for further analysis. In the process of tagging, lexical units were assigned not only a tag, but also grammatical categories, which enabled establishing which part of speech the lexical unit belongs to. According to the structural classification, the tagging is linear because it has a flexible structure. The corpus was also annotated by adding metadata, specifically, the year of publication (Schmid, 1994; Kilgarriff et al., 2004).

Figure 1 illustrates the process of creating the RSA corpus. A corpus consisting of academic articles published between 2020 and 2022 was used as the material for the study. The corpus was created using the Sketch Engine, which was built from articles published in journals such as the International Journal of Applied Earth Observation and Geoinformation, the ISPRS Open Journal of Photogrammetry and Remote Sensing, Remote Sensing Applications: Society and Environment, Remote Sensing of Environment, Remote Sensing. These articles belong to the topic of Earth remote sensing. Therefore, the corpus can be attributed to the second type according to Zakharov’s paradigmatic classification of corpora (Zakharov & Bogdanova ., 2020). The extracted words were compared to the Academic Word List, and the results are presented in Table 2.

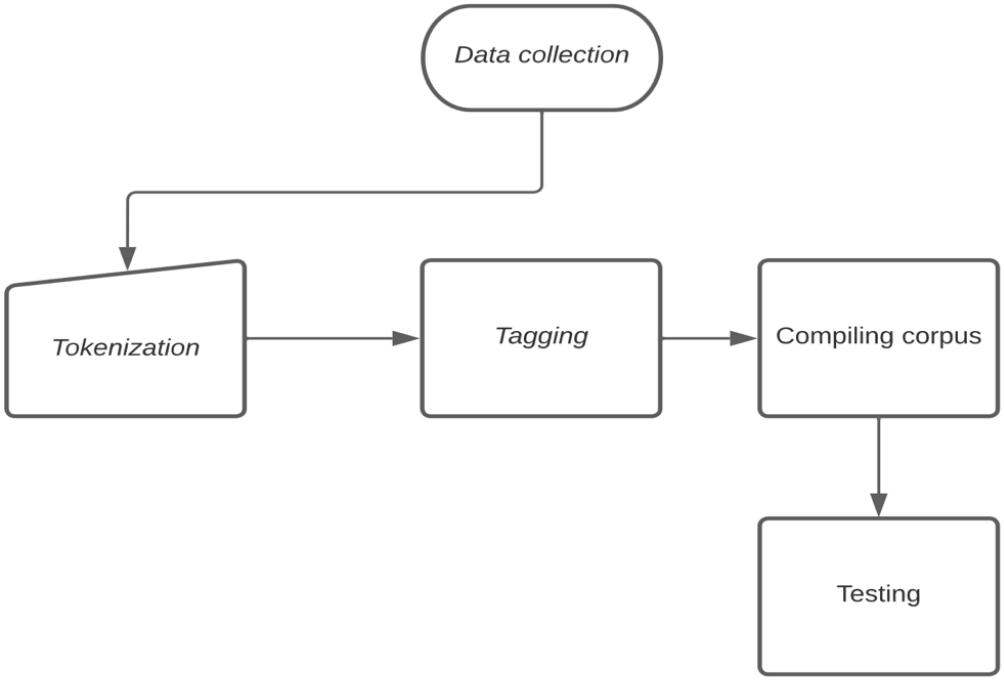

FIGURE 1. Data processing. Compiled by authors

BUILDING CORPUS WITH SKETCH ENGINE

Sketch Engine contains elements of distributive statistical analysis and allows the user to perform it automatically. Several tools are used in this study. The Key Words tool extracts terms from the corpus, which helps to define the topic of the corpus and the most common terms. The extraction process required a reference corpus. It is recommended to use a large universal corpus of the first type as it provides an extensive representation of language material. The Simple Maths method was used to calculate the keyness score, which requires finding the ratio of the normalised frequency of focus and reference corpora. Simple Math method was introduced in 2009 by Kilgarriff (Kilgarriff, 2009). According to Kilgarriff, Simple Math method can solve the problem that appears when there are no occurrences of the word in the reference corpus. It is also said that simple ratios provide a list of rarer lexical units, which makes this method more efficient than Log-likelihood. The keyness score of a word can be calculated using the following formula.

Formula 1

$$\frac{fpm_{rmfocus}+N}{fpm_{rmref}+N},$$

where $fpm_{rmfocus}$ is normalized frequency of the word in the focus corpus,
fpm_{rmref} is normalized frequency of the word in the reference corpus,

N is a smoothing parameter, and the default value is N = 1. However, this value can vary depending on the corpus size (Kilgarriff et al., 2014).

eISSN: 2550-2131
ISSN: 1675-8021

The Collocations tool extracts collocations from the corpus, and LogDice is used as a statistical measure, as it is more efficient than the standard Dice coefficient because it is compatible with small-sized samples. LogDice was introduced by Pavel Rychlý (Rychlý, 2008). This method is based on the Dice coefficient, which expresses the typicality of the collocations. LogDice can be calculated using the following formula:

Formula 2

$$LogDice = 14 + log_2 D = 14 + log_2 \frac{2f_{xy}}{f_x + f_y};$$

Formula 3

$$D = \frac{2f_{xy}}{f_x + f_y},$$

where f_x is the frequency of word X,
f_y is the frequency of word Y,
f_{xy} is the number of co-occurrences of words X and Y.

ANALYSING COLLOCATIONS WITH SKETCH ENGINE

Collocations can be obtained as a frequency list for the entire corpus or specific lexical units (Kilgarriff, 2009; Rychlý, 2008).

The Concordance tool extracts examples of the use of keywords in context, and the results can be grouped using metadata. The tool can find not only words, but also phrases and sentences. Additionally, Corpus Query Language (CQL) and Regular Expression are used to create complex queries, and CQL can search lemmas and wordforms. This language was also used to determine the frequency of the occurrence of affixes in the corpus. The Word List tool automatically generates frequency lists from the corpus. Advanced settings provide the selection of part of speech, as well as the minimum and maximum frequency indicators. These lists contain information regarding absolute and normalised frequencies and tags assigned to lexical units. The Word List tool was implemented to extract nouns with different affixes, which helped to analyse their semantic values. Advanced search was used to perform this operation, allowing the selection of parts of speech and possible affixes. The nouns were divided into groups according to G.H. Abeyweera, who attempted to analyse the use of affixes in academic English. G.H. Abeyweera distinguished neoclassical elements in word formation and phonologically neutral suffixes, prefixes, and suffixes, which were used to analyse affixes in the corpus (Abeyweera, 2021; Rychlý, 2008).

The Word Sketch tool extracts collocations from the corpus and creates semantic fields for keywords. The Word Sketch Difference tool allows the comparison of two words in terms of their semantics, as their collocates are compared. These tools also perform a comparison between the two semantic fields. Sketch Engine also allows visualization of the results, which simplifies the analysis process. Word Sketch tool was used to create a list of terms sorted by frequency. The list was composed of two glossaries: the Glossary of remote sensing by Canada Centre for Remote Sensing and the Glossary of remote sensing and image processing terms by Environmental Systems Research Institute (Esri, n.d.; Natural Resources Canada, 2015. Collocates for the terms and their grammatical categories were extracted using the Collocations tool, which also allows lexical units to be sorted by frequency (Kennedy, 2001; Mozaffari & Moini, 2014).

This study focused on the following tasks:

1. Compiling a corpus of remote sensing articles published between 2020 and 2022.
2. Analysing the frequency of lexical units in the corpus to determine the coverage of the Academic Word List (AWL) Sublist 1, which is one of ten sublists comprising the AWL and consists of the most words, and identifying keywords, multi-word units, and word formation (Coxhead, 2017).
3. Retrieving units from remote sensing glossaries to investigate their distribution in the corpus.
4. Applying corpus linguistics methods and distributive statistical analysis to identify a discipline-specific shortlist of academic vocabulary items that are most relevant to the field of remote sensing of Earth.
5. Discussing the implications of the study's findings for ESP teaching and learning in remote sensing and related technical fields.

RESULTS

The main findings of this research can be found in Appendices A to F.

ACADEMIC WORD LIST

After the search was conducted in the RSA corpus, it was found that AWL Sublist 1 items in total cover 2,19% of the RSA corpus. Only one word (constitutional) from the AWL Sublist 1 was not found in the RSA corpus. The first 30 items are listed in Table 1.

For more details, see Appendix A.

TABLE 1. Top 10 units from AWL Sublist 1 in the RSA corpus by frequency

Rank	Lexical unit	Rank	Lexical unit	Rank	Lexical unit
1	data	11	distribution	21	assessment
2	area	12	structure	22	significant
3	method	13	indicate	23	factor
4	analysis	14	derive	24	occur
5	approach	15	similar	25	specific
6	estimate	16	variable	26	interpretation
7	process	17	function	27	create
8	environment	18	period	28	individual
9	research	19	section	29	identify
10	available	20	source	30	response

Table 1 illustrates the frequency list of lexical units from AWL Sublist 1 presented in the RSA corpus. Appendix A contains two indicators: absolute frequency and relative frequency. Absolute frequency represents the number of lexical units in a corpus. The ratio of the absolute frequency to the corpus size is represented as a result of the relative frequency. These two indicators allow for comparisons between lexical units. The word “data” is used 40,97% more than the word “area”. The absolute frequency of the noun “area” is higher by 30,14%. The least frequent lexical units are “constitutional,” “labor,” “legal,” “legislation,” “income,” “contract,” “authority,” “export,” “sector”. The reason for this low distribution could be that the RSA corpus consists of remote sensing articles, and the lexical units mentioned above belong to legal and business discourse.

When analysing the indicators of the absolute frequency, of lexical unit “analysis”, it can be assumed that it is used 79,21% less often than the word “data”; 64,12% common than the noun “area”; 49,51% less common than the word method. Thus, general words from AWL Sublist 1 like “data”, “analysis”, “methods”, “research”, “approach” are more represented in RSA corpus. At the same time, specialised lexis like “constitutional,” “labor,” “legal,” “legislation,” “income,” “contract,” “authority,” “export,” “sector” has low distribution in RSA corpus.

GLOSSARY AND COLLOCATIONS

Items from two glossaries cover 2,21% of the RSA corpus, with 112 out of 173 terms found. The full list is provided in Appendix B (Glossary of remote sensing and image processing terms; Glossary of remote sensing terms). Additionally, keywords were extracted from the corpus and will be discussed later. Both lists were compared, and the words present in both are listed in Table 2.

To provide more information on the glossary items found in the corpus, collocations were obtained for the top 20 items in order to illustrate the most frequent lexical units in the corpus. The rationale behind this is that the most frequent items in a corpus are those that are most likely to have a significant impact on overall language use in the field and are therefore the most important for language learners to acquire. By focusing on the top 20 items, we can identify key vocabulary items in our field and prioritise their inclusion in teaching materials.

In addition, analysing a smaller number of items in depth allows for a more detailed examination of their collocational patterns and use in context. This can provide insights into the specific ways in which the items are used in the field and help to identify any common collocation errors that learners may make. Fifteen collocates are provided for each base word in Appendix C. An example of raw collocation data is presented in Table 3.

TABLE 2**.** Keywords retrieved from the RSA corpus found in RS glossaries

Rank	Lexical unit	Rank	Lexical unit	Rank	Lexical unit
1	sensor	9	validation	17	georeferencing
2	satellite	10	calibration	18	anthropogenic
3	classification	11	footprint	19	multitemporal
4	pixel	12	scattering	20	backscatter
5	slope	13	sampling	21	phenology
6	cloud	14	topography	22	occlusion
7	resolution	15	amplitude	23	geoid
8	detection	16	histogram	24	dendrogram

Table 2 illustrates the coverage of the Glossary of remote sensing and image processing terms, and the Glossary of remote sensing terms of the RSA corpus. More detailed information is presented in Appendix B. Absolute frequency and relative frequency help to analyse the distribution of lexical units in the studied corpus. The least frequent lexical units are “spatial pattern analysis", "seamline", "resolving power", "orthorectification", "unit", "mensuration minimum mapping unit", "image statistics", "drone imagery", "discrete cosine transform", "digital data", "analogue", "solar insolation".

TABLE 3. Collocations with *'satellite'* retrieved from the RSA corpus using Sketch Engine by score

Keyword	**Grammatical Relation**	**Collocate**	**Freq**	**Score**
satellite	nouns modified by X	imagery	106	10,8
	nouns modified by X	image	146	9,86
	nouns modified by X	datum	118	9,2
	modifiers of X	geostationary	20	9,12
	nouns modified by X	constellation	18	8,99
	nouns modified by X	sensor	19	8,24
	verbs with X as subject	have	18	7,1
	adjective predicates of X	remote	16	6,6
	verbs with X as object	use	18	6,4
	verbs with X as subject	be	37	5,53

Table 3 contains information about collocates to the word satellite sorted by score. In terms of frequency, it is possible that the lexical unit "satellite" is often used as a noun modifier. There are also cases in which the word "satellite" is used as a subject for verbs, or it can be used with adjective predicates. LogDice is used as indicator of a score which shows the typicality of collocations, therefore collocation "imagery satellite" is the most typical for RSA corpus.

RETRIEVED KEYWORDS AND MULTIWORD UNITS

For corpus-based analysis, keywords were extracted from the RSA corpus using Sketch Engine tools to characterise the field of remote sensing in terms of vocabulary. Appendix D provides a list of 100 items by frequency. Apart from frequency, the keyness score is also a valuable indicator, as it can be used to distinguish terms prevailing in specific fields. In Appendix D, the keywords with high scores are in italics. The first 20 terms common to remote sensing by score are listed in Table 4.

TABLE 4. Top 20 keywords retrieved from the RSA corpus by score

Item	Keyword	Frequency (focus)	DOCF (focus)	Relative DOCF (focus)	Score
1	reflectance	736	57	52,77778	287,574
2	modis	597	45	41,66667	279,017
3	geoinformation	316	26	24,07407	215,076
4	multispectral	361	51	47,22222	211,375
5	hyperspectral	322	38	35,18519	178,408
6	convolutional	287	34	31,48148	145,485
7	spectral	1220	75	69,44444	143,518
8	photogrammetry	251	37	34,25926	140,313
9	photogramm	191	49	45,37037	136,93
10	vegetation	1716	77	71,2963	110,313
11	landslide	589	13	12,03704	99,49
12	spatial	2005	103	95,37037	96,463
13	mangrove	440	6	5,55556	91,17
14	crevasse	174	1	0,92593	87,561
15	cropland	187	22	20,37037	87,126
16	subsidence	193	8	7,40741	79,136
17	inundation	174	13	12,03704	78,306
18	spectrometer	286	12	11,11111	76,691
19	segmentation	511	38	35,18519	75,602
20	spatiotemporal	135	33	30,55556	74,527

Table 4 contains information about keywords extracted from the RSA corpus using the Keywords tool. The English Web Corpus 2020 (enTenTen20), which contains 36 billion words, was used as a reference corpus, and the texts were annotated and sorted by topic. Table 4 also shows the absolute frequency, which is the number of occurrences of lexical units in the corpus. Document frequency (DOCF) is the number of documents in which a lexical unit appears. There is also a relative DOCF, which is the ratio of documents with keywords to the number of documents in the corpus. Relative DOCF is similar to relative frequency, and these indicators can be used for comparative analysis of documents in corpora of different sizes. Detailed data are presented in Appendix D. The most frequent keywords are *reflectance, modis, geoinformation, multispectral, hyperspectral, convolutional, spectral, photogrammetry, photogramm,* and *vegetation.* Keywords can be used for the study of terminology, and more than they help students improve their competence is the target field.

Multiword terms were extracted from the corpus. As shown in Appendix E, there are two indicators: absolute and relative frequency. The English Web Corpus 2020 (enTenTen20) was used as a reference corpus. The most frequent lexical units are *remote sense, point cloud, study area, time series, spatial resolution, land cover, applied earth observation, neural network, water body, and satellite image*.

Single-word terms play a key role in a specific area, as do multi-word units (MWUs), which facilitate both reading comprehension and writing. The list of such terms is provided in Appendix E. Moreover, the obtained array of MWUs was also compared with the glossaries chosen for the study. Despite a relatively low intersection, some matches were observed (Table 5).

TABLE 5. Multi-word units retrieved from the RSA corpus found in RS glossaries

Rank	Lexical unit	Rank	Lexical unit	Rank	Lexical unit
1	spatial resolution	5	temporal resolution	9	image analysis
2	earth observation	6	pixel value	10	composite image
3	overall accuracy	7	unmanned aerial vehicle	11	classification scheme
4	satellite imagery	8	spectral resolution	12	flow accumulation

WORD FORMATION

In modern linguistics, language is considered a complex and constantly changing system, in which the processes of development do not stop. Changes most often occur in the lexis, and word formation is a means of vocabulary extension. The current research is based on the work of Abeyweera (Abeyweera, 2021). Abeyweera mentioned neoclassical elements, affixes, and suffixes. Neoclassical elements are derived from the Greek and Latin languages. These elements were phonologically and morphologically assimilated. Neoclassical elements are typically used in academic discourse to develop terminology and create new terms (ibid.).

During the study, lexical units containing neoclassical elements were extracted from the RSA corpus using the Word List tool. The most frequent neoclassical element is *photo-* (1171 hits). It is used in the following terms: *photogrammetry, photogram, photosynthesis, photograph, photogrammetric, photosynthetic*. Another popular neoclassical element is *bio-* (733 hits). It is present in the following lexical units: *biomass, biodiversity, biome, biophysical, biological, biochemical, biogeoscience,* and others. The least frequent element is *logy-* (878 hits), which can be found in terms like *methodology, technology, ecology, phenology, climatology, geology, morphology, hydrology,* and others.

Phonologically neutral suffixes do not change the stress of a word when attached to a stem. The stress of the word is the same as before the addition of the phonologically neutral suffix was added. In 2021 Abeyweera distinguished the following elements: *propag-, adv-, art-, radi-* (Abeyweera, 2021). The most common phonologically neutral suffix in the RSA corpus is radi- (175 hits), which is used in words such as: *radiation, radiometric, radiometer*, *radioactive*, *radiodata*. Another element presented in the RSA corpus is *propag-* (99 hits), which is less common than the suffix *radi-*. The derivational element *propag-* is used in the following terms: *propagation* and *propagate*.

Prefixes were retrieved from the RSA corpus using the Word List tool and an advanced search was performed.

According to the results, ten most frequent prefixes are *in-* (5888 hits), *co-* (4065 hits), *pre-* (2023 hits), *multi-* (1797 hits), *dis-* (1535 hits), *inter-* (1366 hits), ex- (1281 hits), *un-* (1215 hits), *photo-* (1171 hits), and *sub-* (1126 hits). The Three least frequent prefixes included *de-* (13 hits), *anti-* (12 hits), *retro-* (12 hits), *eu-* (9 hits), and *des-* (8 hits).

In academic discourse, derivational elements, such as suffixes, are also popular, as they form the scientific vocabulary and terminology of the studied field. The suffix is a derivational unit that is attached to a word after a stem. The three most frequent suffixes are *-able* (2685 hits), *-ize/-yze* (1560 hits), and *-logy* (878 hits). The three least frequent suffixes were *-ise* (76 hits), *-fusion* (16 hits), and *-dom* (10 hits). Based on the results of the study, the suffix *-ize* is used 95,13% more than the suffix *-ise*, which means that American spelling is more common than British in the RSA corpus. Further research is needed for the final conclusions, because the RSA corpus covers only the field of Remote Sensing and was compiled with articles from certain academic journals.

For further details, see Appendix F.

DISCUSSION

With technology's evolving in almost all fields, the language as well is changing, and aerospace is not an exception. Advances in the field under consideration have caused shifts in lexical structures (Dmitrichenkova & Dolzhich, 2020). The glossaries mentioned in this article were published and are available online, although there still is a lack of ESP teaching materials for aerospace and remote sensing in particular. The above stated is a stimulus for further research into vocabulary of this field, which could focus on general academic vocabulary, d

The relatively low coverage for both AWL Sublist 1 and the glossaries (approximately 2% in each case) can be explained as follows. First, only 60 of 570 word families of the AWL were used for the study, with the complete list, the results are expected to be altered. Second, the relatively small size of the corpus may explain its low coverage. Alternatively, existing glossaries may require revision and update, as the articles that comprise the corpus date from 2020 to 2022.

Moreover, some AWL Sublist 1 words (*export, authority, contract, income, legislation, legal, labour*) ranked the lowest, which establishes the correlation (even the weak one) between the source corpus for AWL and the RSA corpus. Consequently, AWL Sublist 1 had low coverage in the RSA corpus because the latter was compiled from the articles on remote sensing and is not multidisciplinary.

The collocations identified were rather specific and characteristic of the field, despite the presence of some general structures (noun + *be/have/use*). The same has been observed in civil engineering texts (Otto, 2021).

As a fruitful approach to term extraction (Pérez & Rizzo, 2013), keyword search demonstrated positive results with 390 terms (or candidates) in total extracted automatically and 24 found in glossaries compiled by people (out of 112 found in the RSA corpus).

It has become evident that multi-word units play an important role in the academic language (Coxhead, 2017; Granger & Larsson, 2021), which underlines the significance of collocations and multi-word terms (or N-grams) for learners of English as a second language and, especially, for those who write academic articles in English. The number of extracted MWUs was 1000, which is noticeably higher than that of single-word terms, but only 80 were included in the final list because of their considerable frequency and score.

Word formation is a source of English vocabulary extension. Evidence from this study suggests that prefixes are used more frequently than suffixes to produce new words in academic discourse. According to the results, *in-* was the most prevalent prefix that formed the negative form of the lexical units. Prefixes pre-, multi-, dis-, inter-, ex-, un-, photo-, and sub- are also relatively common. These affixes are used in parasynthetic derivation, which produces word forms through two-word formation processes. For instance, an adjective *unavailable* was created with the prefix *-un* and suffix *-able*. Another example of parasynthetic derivation could be noun *ecohydrology*, which was formed with the prefix *eco-* and suffix *-logy*, which Abeyweera considers neoclassical elements of word formations, which are widely represented in the RSA corpus (Abeyweera, 2021).

Moreover, methods of corpus linguistics can be considered one of the branches of Data-Driven Learning (DDL). According to Chujo et al. DDL is an approach that motivates students to use authentic materials in ESP (Chujo et al., 2013). Authentic materials help to create intercultural competence, while the latter is considered a goal of language learning. Data-Driven Learning aims to create background knowledge to improve competence in a field. Simultaneously, this approach can only be successfully implemented at intermediate and advanced levels, and the complexity of DDL for beginner students is illustrated by two challenges. First, the target corpus may not correspond to students' levels. Second, corpus manager tools can be complicated for beginners. These challenges can be overcome through the proper design of ESP courses. Anthony proposes the "teacher as a student" approach, which requires linguistic corpora while designing ESP courses. Anthony pointed out that the lack of knowledge of the target field is more of an advantage than disadvantage, as it allows teachers to understand the needs of students and adapt materials to them (Anthony, 2007).

As mentioned earlier, linguistic corpora provide authentic material with concordances, and the winch illustrates the lexical units searched for in context. At the same time, there are different aspects of authenticity: sociocultural, lexical and functional. The sociocultural aspect is more significant in literary discourse as it represents the realities of the country of the studied language (Galskova & Gez, 2004). However, the linguistic and functional aspects can also be applied to academic discourse. The lexical aspect includes background lexical units that expand students' vocabulary. At the same time, lexical authenticity provides a wider representation of the terminology in the study area. Functionality is also an important parameter of authentic materials as it implies the natural selection of linguistic means. Many modern textbooks include text that teaches speech behaviour in the realities of the studied language and illustrates the generalised situations of communication. This helps students to learn common patterns in the target language (Ter-Minasova, 2000; Kitaygorodskaya, 2009).

Nevertheless, compiling corpora from authentic datasets may not always be effective for beginners due to the lexical and syntactic specifics of authentic materials, and because some lexical units can be beyond the understanding of students. Considering this, it is important to mention the issue of adapting authentic text to students' level of knowledge in accordance with their learning objectives. There are two methods for adapting authentic texts. The quantitative method consists of reducing the least significant lexical elements so that the main idea of the text becomes more understandable. As far as corpora are concerned, it is possible to preprocess data and delete secondary elements. Qualitive adaptation is the grammatical and lexical replacement of elements that students find difficult to perceive. Qualitative adaptation also includes an explanation of the concepts in the studied field and the introduction of new lexical units with the help of synonymy.

The glossary created as part of the study has an impact on two areas: education and science. For instance, teachers can use glossaries in ESP classes to help students develop their academic writing skills. These materials can assist in acquiring the necessary vocabulary related to aerospace. Similarly, scientists can utilise glossaries to write academic articles and communicate professionally at an international level. Furthermore, this glossary can be seen as an addition to the previous research conducted by Valipouri and Nassaji, Roeseler, and Muños. However, the peculiarity of this glossary lies in its domain specificity (Valipouri & Nassaji, 2013; Roesler, 2021; Muñoz, 2015).

In conclusion, it is important to mention that pre-processing data and adaptation can be effective for beginners, but it is better to use unprocessed data for corpora for advanced-level students. Further research is needed to prove the effectiveness of implementing linguistic corpora as authentic materials.

CONCLUSION

An attempt was made to create a discipline-specific word list that might be of use for educators, ESP learners, and authors in the field of remote sensing. Such work is highly needed in other fields as well (Mozaffari & Moini, 2014; Valipouri & Nassaji, 2013). Undoubtedly, more research should be conducted on this topic, along with its linguistic aspects, and not be solely limited by vocabulary studies. In addition, an expert-judged approach can be employed to further improve the obtained keywords and MWU lists (Ackermann & Chen, 2013).

Although the current study focused mostly on narrow-field vocabulary, another vital factor for successful communication is the vast general vocabulary, which can be applied in various scenarios such as delivering presentations (Dang, 2022). This fact should not be overlooked by teachers and material developers when paying attention to discipline-specific lexis, as it is likely to occur not only in a limited set of contexts.

The evidence presented thus far supports the idea that learners may greatly benefit from using word lists to boost their vocabulary by looking up terms, defining them, and studying term usage in context (Smith, 2020). The present study may be used as a basis for such type of independent learning, along with guided discovery. The findings presented in this study may also be employed for planning course curricula or developing ESP materials aimed at vocabulary expansion.

Corpus linguistic methods and distributive statistical analysis can effectively process large amounts of data. The corpus manager enables the extraction of collocations and keywords. Moreover, it includes distributive analysis methods in its system, which increase the effectiveness of the study. However, corpus linguistic methods have some limitations such as low-quality data, incorrect tagging, and false queries. The quality of the data affects the results of the search and can be solved while creating a corpus by pre-processing the data. For example, if a corpus consists of academic papers, it is recommended to delete references and information about authors; otherwise, these data will appear in concordances and will cause false results. In the case of unrepressed texts, the results can be sorted manually. Incorrect tagging can also appear during the compilation of the corpus. To avoid this problem, texts can be tagged with unsupervised taggers, such as Tree Tagger. False queries can lead to inaccurate results, and the use of CQL as an advanced search tool could be a solution to this problem (Schmid, 1994).

The current study could instigate novel investigations into the linguistic features characteristic of academic language in the field of aerospace, which is considerably broader than remote sensing. Future work might refine the findings obtained and make them more relevant for ESP learners, professionals, and authors, facilitating discipline-specific language acquisition.

REFERENCES

Abeyweera, G. H. (2020). The use of affixation in academic English: A lexical explanation on affixation, root and meaning. *Journal of Social Sciences and Humanities Review*, *5*(4).

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic purposes*, *12*(4), 235-247.

Anthony, L. (2007, October). The teacher as student in ESP course design. In *The Proceedings of 2007 International Symposium on ESP & Its Applications in Nursing and Medical English Education* (pp. 70-79).

Barok, D. (2013). Cooperation in Space between Europe and Israel in light of the recent ESA-ISA agreement. *Yearbook on Space Policy 2010/2011: The Forward Look*, 191-206.

Belcher, D. D. (2006). English for specific purposes: Teaching to perceived needs and imagined futures in worlds of work, study, and everyday life. *TESOL quarterly*, *40*(1), 133-156.

Benesch, S. (1996). Needs analysis and curriculum development in EAP: An example of a critical approach. *Tesol Quarterly*, *30*(4), 723-738.

Bi, J. (2020). How large a vocabulary do Chinese computer science undergraduates need to read English-medium specialist textbooks?. *English for Specific Purposes*, *58*, 77-89.

Chujo, K., Anthony, L., Oghigian, K., & Yokota, K. (2013). Teaching remedial grammar through data-driven learning using AntPConc. *Taiwan International ESP Journal*, *5*(2), 65-90.

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

Coxhead, A. (2017). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives*. routledge.

Csomay, E., & Petrović, M. (2012). "Yes, your honor!": A corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System*, *40*(2), 305-315.

Dang, T. N. Y. (2022). Vocabulary in academic lectures. *Journal of English for Academic Purposes*, *58*, 101123. https://doi.org/10.1016/j.jeap.2022.101123

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66-76.

Dmitrichenkova, S. V., & Dolzhich, E. A. (2020). Space technologies to form lexical structure of academic discourse. In *Advances in the Astronautical Sciences* (pp. 913-918).

Dvoryadkina, N., & Mikheeva, N. (2018). Tackling tasks of professionally oriented English language training for cosmonauts and specialists of aerospace industry using computer-assisted teaching materials based on project activities. In *EDULEARN18 Proceedings* (pp. 4850-4854). IATED.

Esri. (n.d.). *Glossary of remote sensing and image processing terms.* Version 1.0. Retrieved June 25, 2022, from http://downloads.esri.com/resources/imageryandraster/ImageryGlossaryIA_v01.pdf

Galskova, N. D., & Gez, N. I. (2004). Theory of teaching foreign languages. Linguodidactics and methodology. *Moscow: Academiya*.

Granger, S., & Larsson, T. (2021). Is core vocabulary a friend or foe of academic writing? Single-word vs multi-word uses of THING. *Journal of English for Academic Purposes*, *52*, 100999.

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes*, *33*, 54-65.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes*. Cambridge university press.

Johns, A. M., & Dudley-Evans, T. (1991). English for specific purposes: International in scope, specific in purpose. *TESOL quarterly*, *25*(2), 297-314.

Johns, A. M., & Dudley-Evans, T. (1991). English for specific purposes: International in scope, specific in purpose. *TESOL quarterly*, *25*(2), 297-314.

Kennedy, H. (Ed.). (2001). *Dictionary of GIS terminology*. Environmental Systems research institute.

Kilgarriff, A. (2009). Simple Maths for Keywords. In *Proceedings of the Corpus Linguistics Conference 2009* (CL2009), (p. 171).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36.

Kilgarriff, A., Reddy, S., Pomikálek, J., & Avinesh, P. V. S. (2010, May). A Corpus Factory for Many Languages. In *LREC*.

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, 105–115.

Kitaygorodskaya G.A. (2009) Intensive foreign language teaching. Theory and practice. Publishing house "Higher School"

Kohnke, L., Zou, D., & Zhang, R. (2021). Exploring discipline-specific vocabulary retention in L2 through app design: Implications for higher education students. *RELC Journal*, *52*(3), 539-556.

Laborda, J. G., & Litzler, M. F. (2015). Current perspectives in teaching English for specific purposes. *Onomázein*, (31), 38-51.

Le, C. N. N., & Miller, J. (2020). A corpus-based list of commonly used English medical morphemes for students learning English for specific purposes. *English for Specific Purposes*, *58*, 102-121.

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

Li, Y., & Qian, D. D. (2010). Profiling the Academic Word List (AWL) in a financial corpus. *System*, *38*(3), 402-411.

Lukianenko, V., & Vadaska, S. (2020). Evaluating the efficiency of online English course for first-year engineering students. *Revista Romaneasca Pentru Educatie Multidimensionala*, *12*(2Sup1), 62-69.

Moraño-Fernandez, J. A., Moll-Lopez, S., Sanchez-Ruiz, L. M., Vega-Fleitas, E., López-Alfonso, S., & Puchalt-López, M. (2019, October). Micro-Flip Teaching with e-learning Resources in Aerospace Engineering Mathematics: A Case Study. In *Proceedings of the World Congress on Engineering and Computer Science (WCECS), San Francisco, CA, USA* (pp. 22-24).

Mozaffari, A., & Moini, R. (2014). Academic words in education research articles: A corpus study. *Procedia-Social and Behavioral Sciences*, *98*, 1290-1296.

Muñoz, V. L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *English for Specific Purposes*, *39*, 26-44.

Musikhin, I. A. (2016). ENGLISH FOR SPECIFIC PURPOSES: TEACHING ENGLISH FOR SCIENCE AND TECHNOLOGY. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, *3*(6).

Najmon, J. C., Raeisi, S., & Tovar, A. (2019). Review of additive manufacturing technologies and applications in the aerospace industry. *Additive manufacturing for the aerospace industry*, 7-31.

Natural Resources Canada. (2015, November 25). Glossary of remote sensing terms. Retrieved June 25, 2022, from https://natural-resources.canada.ca/maps-tools-and-publications/satellite-imagery-and-air-photos/satellite-imagery-products/educational-resources/glossary-remote-sensing-terms/9483

Netikšienė, N. (2006). Teaching English for specific purposes. *Santalka: Filologija, Edukologija*, *14*(4), 80-82.

Ng, Y. J., Chong, S. T., Thiruchelvam, S., Chow, M. F., & Karthikeyan, J. (2020). Vocabulary threshold for the comprehension of Malaysian secondary engineering texts as compared to the non-engineering genres. *International Journal of Innovation, Creativity and Change*, *14*(1), 488-504.

Otto, P. (2021). Choosing specialized vocabulary to teach with data-driven learning: An example from civil engineering. *English for Specific Purposes*, *61*, 32-46.

Paltridge, B., & Starfield, S. (Eds.). (2013). *The handbook of English for specific purposes* (Vol. 592). Boston: Wiley-blackwell.

Pérez, María José Marín, and Camino Rea Rizzo. "Automatic Access to Legal Terminology Applying two Different Automatic Term Recognition Methods." *Procedia-Social and Behavioral Sciences* 95 (2013): 455-463.

Petrescu, R. V., Aversa, R., Akash, B., Bucinell, R., Corchado, J., Apicella, A., & Petrescu, F. I. (2017). Modern propulsions for aerospace-a review. *Journal of Aircraft and Spacecraft Technology*, *1*(1).

Poedjiastutie, D. (2017). The Pedagogical Challenges of English for Specific Purposes (ESP) Teaching at the University of Muhammadiyah Malang, Indonesia. *Educational Research and Reviews*, *12*(6), 338-349.

Richards, J. C. (1974). Word lists: Problems and prospects. *RELC journal*, *5*(2), 69-84.

Roesler, D. (2021). When a bug is not a bug: An introduction to the computer science academic vocabulary list. *Journal of English for Academic Purposes*, *54*, 101044.

Rychlý, P. (2008, December). A Lexicographer-Friendly Association Score. In *RASLAN* (pp. 6-9).

Safari, M. (2019). English vocabulary for equine veterans: How different from GSL and AWL words. *Iranian Journal of English for Academic Purposes*, *8*(2), 51-65.

Schmid, H. (1994). Part-of-Speech Tagging With Neural Networks. *International Conference on Computational Linguistics.*

Schmid, H., & Laws, F. (2008, August). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* (pp. 777-784).

Shabani, M. B., & Tazik, K. (2014). Coxhead's AWL across ESP and Asian EFL journal research articles (RAs): A corpus-based lexical study. *Procedia-Social and Behavioral Sciences*, *98*, 1722-1728.

Smith, S. (2020). DIY corpora for Accounting & Finance vocabulary learning. *English for Specific Purposes*, 57, 1–12. https://doi.org/10.1016/j.esp.2019.08.002Simon Smith, (2020) DIY corpora for Accounting & Finance vocabulary learning, English for Specific Purposes. 57, 1-12. https://doi.org/10.1016/j.esp.2019.08.002.

Stojković, N. (Ed.). (2018). *Positioning English for specific purposes in an English language teaching context*. Vernon Press.

Ter-Minasova, S. G. (2000). Language and intercultural communication. *Moscow: Slovo*, *634*.

Tevdovska, E. S. (2018). Authentic materials vs textbooks in ESP (English for Specific Purposes). *The Journal of Languages for Specific Purposes (JLSP)*

Thompson, S. A. (1989). *A discourse approach to the cross-linguistic category 'adjective'* (Vol. 61, p. 245). John Benjamins Publishing.

Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes*, *12*(4), 248-263.

Veenstra, J., & Sato, Y. (2018). Creating an institution-specific science and engineering academic word list for university students. *Journal of Asia TEFL*, *15*(1), 148.

Vongpumivitch, V., Huang, J. Y., & Chang, Y. C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, *28*(1), 33-41.

Vora, R. (2017). Integrating authentic materials and language skills in teaching english for specific purposes. In *Noi tendinţe în predarea limbajelor de specialitate în contextul racordării învăţământului* (pp. 140-144).

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for specific purposes*, *28*(3), 170-182.

Yakushev, V. P., Dubenok, N. N., & Loupian, E. A. (2019). Earth remote sensing technologies for agriculture: application experience and development prospects. *Current problems in remote sensing of the Earth from space*, *16*(3), 11-23.

Zakharov, V. P. (2015). Corpus-based approach to thesaurus and ontology construction, Strukt. *Prikl. Lingvist*, (11), 123-141.

Zakharov, V. P., & Bogdanova, S. Y. (2020). Corpus linguistics. St. Petersburg: Publishing house of St. Petersburg University.

Zhang, X., Chen, Y., & Hu, J. (2018). Recent advances in the development of aerospace materials. *Progress in Aerospace Sciences*, *97*, 22-34.

APPENDIX A

AWL SUBLIST 1 BY FREQUENCY IN THE RSA CORPUS

Rank	Word	Frequency	Relative Frequency
1	data	6744	0,481%
2	area	3913	0,279%
3	method	2781	0,198%
4	analysis	1404	0,100%
5	approach	1034	0,074%
6	estimate	1003	0,071%
7	process	948	0,068%
8	environment	865	0,062%
9	research	744	0,053%
10	available	743	0,053%
11	distribution	709	0,051%
12	structure	668	0,048%
13	indicate	650	0,046%
14	derived	607	0,043%
15	similar	598	0,043%
16	variables	596	0,042%
17	function	548	0,039%
18	period	544	0,039%
19	section	534	0,038%
20	source	519	0,037%
21	assessment	519	0,037%
22	significant	505	0,036%
23	factors	497	0,035%
24	occur	391	0,028%
25	specific	335	0,024%
26	interpretation	289	0,021%
27	create	238	0,017%
28	individual	235	0,017%
29	identified	233	0,017%
30	response	229	0,016%
31	context	215	0,015%
32	required	192	0,014%
33	consistent	180	0,013%
34	assume	167	0,012%
35	economic	155	0,011%
36	procedure	152	0,011%

eISSN: 2550-2131
ISSN: 1675-8021

37	major	143	0,010%
38	role	143	0,010%
39	established	132	0,009%
40	policy	111	0,008%
41	financial	96	0,007%
42	benefit	93	0,007%
43	concept	86	0,006%
44	issues	81	0,006%
45	principle	79	0,006%
46	definition	74	0,005%
47	formula	72	0,005%
48	theory	71	0,005%
49	evidence	58	0,004%
50	percent	56	0,004%
51	involved	55	0,004%
52	sector	26	0,002%
53	export	20	0,001%
54	authority	14	0,001%
55	contract	10	0,001%
56	income	5	0,000%
57	legislation	4	0,000%
58	legal	2	0,000%
59	labour	1	0,000%
60	constitutional	0	0,000%

eISSN: 2550-2131
ISSN: 1675-8021

APPENDIX B

GLOSSARY TERMS IN THE RSA CORPUS BY FREQUENCY

Rank	Term	Frequency	Relative Frequency
1	datum	5498	0,392%
2	image	4075	0,290%
3	satellite	1734	0,124%
4	cloud	1627	0,116%
5	classification	1625	0,116%
6	resolution	1352	0,096%
7	pixel	1286	0,092%
8	lidar	966	0,069%
9	sensor	965	0,069%
10	index	954	0,068%
11	remote sensing	851	0,061%
12	detection	844	0,060%
13	landsat	731	0,052%
14	monitoring	674	0,048%
15	scale	629	0,045%
16	application	578	0,041%
17	slope	485	0,035%
18	spatial resolution	471	0,034%
19	earth observation	430	0,031%
20	footprint	417	0,030%
21	target	372	0,027%
22	UAV	342	0,024%
23	radar	283	0,020%
24	sampling	281	0,020%
25	platform	248	0,018%
26	key	190	0,014%
27	overall accuracy	182	0,013%
28	satellite imagery	171	0,012%
29	aspect	157	0,011%
30	topography	156	0,011%
31	orbit	139	0,010%
32	scanner	129	0,009%
33	enhancement	124	0,009%
34	transform	124	0,009%
35	spectrum	116	0,008%
36	temporal resolution	109	0,008%
37	histogram	106	0,008%
38	texture	106	0,008%
39	georeferencing	90	0,006%

40	anthropogenic	78	0,006%
41	reflection	73	0,005%
42	raster	69	0,005%
43	UAS	68	0,005%
44	multitemporal	67	0,005%
45	stability	64	0,005%
46	GPS	56	0,004%
47	pixel value	55	0,004%
48	Unmanned aerial vehicle	55	0,004%
49	reference data	52	0,004%
50	spectral resolution	52	0,004%
51	image analysis	50	0,004%
52	phenology	44	0,003%
53	occlusion	43	0,003%
54	geoid	36	0,003%
55	mosaic	36	0,003%
56	nadir	35	0,002%
57	transmit	32	0,002%
58	composite image	25	0,002%
59	classification scheme	24	0,002%
60	dendrogram	24	0,002%
61	near infrared	24	0,002%
62	bathymetry	22	0,002%
63	emit	21	0,001%
64	flow accumulation	21	0,001%
65	block adjustment	17	0,001%
66	ground station	17	0,001%
67	orthogonal	17	0,001%
68	tone	17	0,001%
69	principal component analysis	16	0,001%
70	backscattering	15	0,001%
71	error matrix	13	0,001%
72	inertial measurement unit	13	0,001%
73	pyramid	13	0,001%
74	Unmanned aerial system	13	0,001%
75	wavelet transform	12	0,001%
76	Global Positioning System	9	0,001%
77	parallelepiped	9	0,001%
78	pan sharpening	8	0,001%
79	float	7	0,000%

eISSN: 2550-2131
ISSN: 1675-8021

80	Gamma	7	0,000%
81	nonparametric	7	0,000%
82	insolation	6	0,000%
83	line-of-sight	6	0,000%
84	map accuracy	6	0,000%
85	radiometric resolution	6	0,000%
86	web service	6	0,000%
87	image elements	5	0,000%
88	thematic accuracy	5	0,000%
89	reprojection	4	0,000%
90	electromagnetic spectrum	3	0,000%
91	GeoTIFF	3	0,000%
92	orthophotography	3	0,000%
93	tiling	3	0,000%
94	basemap	2	0,000%
95	compression	2	0,000%
96	ground truthing	2	0,000%
97	ortho	2	0,000%
98	orthoimage	2	0,000%
99	positional accuracy	2	0,000%
100	radarsat	2	0,000%
101	solar insolation	2	0,000%
102	analogue	1	0,000%
103	digital data	1	0,000%
104	discrete cosine transform	1	0,000%
105	drone imagery	1	0,000%
106	image statistics	1	0,000%
107	mensuration	1	0,000%
108	minimum mapping unit	1	0,000%
109	orthorectification	1	0,000%
110	resolving power	1	0,000%
111	seamline	1	0,000%
112	spatial pattern analysis	1	0,000%

APPENDIX C

COLLOCATIONS WITH GLOSSARY ITEMS IN THE RSA CORPUS

Item	Base (*X*)	Collocate
1	datum	use *X*, sense *X*, satellite *X*, collect *X*, LiDAR *X*, SAR *X*, *X* be, training *X*, base on *X*, Landsat *X*, airborne *X*, acquire *X*, provide *X*, MODIS *X*, Sentinel-2 *X*
2	image	satellite *X,* sense *X,* Landsat *X,* use *X,* SAR *X, X* classification, cloudy *X*, know *X*, acquire *X*, *X* segmentation, MODIS *X*, multispectral *X*, Sentinel-2 *X*, *X* be, optical *X*
3	remote sensing	science *X, X* applications, *X* environment, photogrammetry X, geoscience *X, X* image, *X* data, *X* symposium, *X* imagery, optical *X*, satellite *X,* using *X,* multispectral *X,* hyperspectral *X, X* dataset
4	satellite	*X* imagery, *X* image, *X* datum, geostationary *X, X* constellation, *X* sensor, *X* have, remote *X,* use *X, X* be
5	cloud	point *X, X* removal, *X* cover, 3D *X, X* coverage, *X* size, LiDAR *X, X* shadow, *X* mask, *X* registration, *X* and/or shadow, thick *X,* dense *X,* remove *X,* thin *X*, photon *X*
6	classification	cover *X*, land *X*, *X* accuracy, supervised *X*, *X* result, LULC *X,* urban *X,* image *X, X* scheme, *X* method, land-use *X, X* algorithm, hyperspectral *X,* base *X,* accurate *X, X* performance
7	resolution	spatial *X,* temporal *X,* high *X,* m *X,* fine *X*, *X* imagery, *X* of m, spectral *X,* coarse *X, X* of km, km *X, X* image, have *X,* low *X,* increase *X*
8	pixel	*X* size, MODIS *X,* target *X,* number of *X,* AF *X,* mixed *X, X* candidate, *X* value, *X* belong, similar *X, X* and/or pixel, select *X,* training *X,* value of *X, X* have, *X* in image
9	lidar	airborne *X, X* datum
10	sensor	ERS *X,* AVHRR *X,* inspection *X,* LiDAR *X,* satellite *X,* different *X,* use *X, X* be, *X* datum
11	index	refractive *X,* normalize *X,* vegetation *X,* leaf *X,* difference *X,* clump *X,* area *X, X* calculation, semantic *X,* water *X,* NDVI *X, X* value, *X* be, be *X*
12	detection	change *X,* object *X,* crack *X,* active *X,* fire *X, X* algorithm, *X* rate, *X* method, *X* error, *X* accuracy, cloud *X, X* and/or classification, point *X, X* use, method for *X*

13	landsat	*X* OLI, *X* ETM, *X* TM, *X* imagery, *X* images, *X* series, *X* Sentinel-2, using *X,* OLI *X, X* Thematic, *X* MODIS, MODIS *X, X* data, *X* time, *X* Sentinel
14	monitoring	forest *X*
15	scale	regional *X,* large *X,* global *X,* landscape *X,* spatial *X,* temporal *X,* different *X,* small *X, X* use, *X* be
16	application	agricultural *X,* its *X,* sense *X, X* be
17	slope	*X* filter, bank *X, X* and/or aspect, *X* and/or elevation, *X* be
18	spatial resolution	coarse *X,* high *X,* fine *X,* higher *X,* m *X,* km *X,* medium *X, X* temporal, at *X,* low *X,* very *X,* finer *X, X* multispectral, *X* satellite
19	earth observation	*X* geoinformation, applied *X, X* EO, *X* group, *X* satellites, *X* and, of *X, X* cubes, *X* big, *X* centre, *X* committee, *X* launches, *X* data, learning *X, X* satellite
20	footprint	*X* location, GEDI *X, X* extraction, *X* be

APPENDIX D

KEYWORDS EXTRACTED FROM RSA CORPUS BY FREQUENCY

Item	Keyword	Frequency (focus)	DOCF (focus)	Relative DOCF (focus)	Score
1	*remote*	4301	108	100	59,874
2	*spatial*	2005	103	95,37037	96,463
3	forest	1995	73	67,59259	17,898
4	*satellite*	1734	91	84,25926	36,561
5	*vegetation*	1716	77	71,2963	110,31
6	*accuracy*	1716	96	88,88889	51,128
7	cloud	1627	83	76,85185	17,955
8	*classification*	1625	81	75	57,097
9	resolution	1352	103	95,37037	16,189
10	algorithm	1317	95	87,96296	31,42
11	*pixel*	1286	89	82,40741	46,57
12	*spectral*	1220	75	69,44444	143,52
13	*dataset*	1103	95	87,96296	74,35
14	observation	1070	98	90,74074	18,069
15	*mapping*	990	86	79,62963	48,868
16	sensor	965	83	76,85185	15,272
17	parameter	958	89	82,40741	14,51
18	detection	844	85	78,7037	22,946
19	measurement	843	85	78,7037	14,245
20	*imagery*	778	82	75,92593	52,891
21	respectively	753	101	93,51852	14,658
22	*reflectance*	736	57	52,77778	287,57
23	derive	709	85	78,7037	14,073
24	*estimation*	676	87	80,55556	63,081
25	*temporal*	605	72	66,66667	49,866
26	*modis*	597	45	41,66667	279,02
27	*landslide*	589	13	12,03704	99,49
28	ecosystem	568	58	53,7037	18,273
29	validation	551	84	77,77778	34,219
30	extraction	515	54	50	34,134
31	*segmentation*	511	38	35,18519	75,602
32	prediction	498	56	51,85185	17,081
33	slope	485	50	46,2963	17,807
34	density	482	60	55,55556	13,586
35	*indices*	456	46	42,59259	61,885
36	*int*	443	92	85,18519	37,289
37	*mangrove*	440	6	5,55556	91,17
38	*calibration*	420	38	35,18519	40,996

39	footprint	417	27	25	26,848
40	elevation	417	61	56,48148	21,169
41	*precipitation*	412	41	37,96296	44,897
42	wetland	404	24	22,22222	29,236
43	correlation	398	78	72,22222	19,983
44	regression	396	56	51,85185	35,219
45	*retrieval*	393	31	28,7037	46,226
46	*photon*	385	5	4,62963	47,1
47	coefficient	369	61	56,48148	31,019
48	*sentinel*	367	30	27,77778	49,638
49	atmospheric	367	51	47,22222	22,614
50	intensity	364	50	46,2963	13,605
51	*multispectral*	361	51	47,22222	211,38
52	neural	361	50	46,2963	28,258
53	applied	342	44	40,74074	19,975
54	classify	335	63	58,33333	13,631
55	*hyperspectral*	322	38	35,18519	178,41
56	*geoinformation*	316	26	24,07407	215,08
57	*deforestation*	316	14	12,96296	68,178
58	airborne	313	40	37,03704	36,186
59	semantic	299	23	21,2963	35,485
60	fusion	299	41	37,96296	16,407
61	*convolutional*	287	34	31,48148	145,49
62	*spectrometer*	286	12	11,11111	76,691
63	deviation	285	64	59,25926	26,232
64	*als*	283	13	12,03704	62,644
65	*trans*	282	74	68,51852	47,787
66	*glacial*	275	8	7,40741	57,2
67	applications	265	47	43,51852	19,573
68	variability	261	59	54,62963	26,569
69	aerial	252	46	42,59259	15,745
70	*photogrammetry*	251	37	34,25926	140,31
71	*spectra*	251	20	18,51852	38,536
72	*normalize*	242	67	62,03704	36,997
73	fuse	226	30	27,77778	16,385
74	infrared	209	57	52,77778	18,884
75	*classifier*	207	36	33,33333	63,867
76	proceedings	205	62	57,40741	13,809
77	*situ*	200	25	23,14815	38,422
78	*high-resolution*	197	65	60,18519	37,663
79	*subsidence*	193	8	7,40741	79,136
80	*photogramm*	191	49	45,37037	136,93
81	*scattering*	190	20	18,51852	38,07
82	remotely	190	48	44,44444	14,525

83	coarse	189	29	26,85185	25,032
84	quantify	189	59	54,62963	18,566
85	*cropland*	187	22	20,37037	87,126
86	sampling	186	50	46,2963	13,568
87	*impervious*	179	9	8,33333	64,165
88	deformation	178	9	8,33333	32,196
89	gradient	175	44	40,74074	17,067
90	*crevasse*	174	1	0,92593	87,561
91	*inundation*	174	13	12,03704	78,306
92	baltic	173	3	2,77778	26,348
93	denote	170	40	37,03704	13,705
94	stockpile	169	1	0,92593	34,105
95	grassland	168	26	24,07407	24,166
96	*chlorophyll*	166	23	21,2963	58,37
97	susceptibility	166	10	9,25926	31,992
98	*topographic*	165	36	33,33333	53,157
99	polygon	163	24	22,22222	28,033
100	simulated	163	28	25,92593	22,747

eISSN: 2550-2131
ISSN: 1675-8021

APPENDIX E

MULTI-WORD TERMS FROM THE RSA CORPUS BY FREQUENCY

Item	Multi-word term	Frequency	Relative frequency
1	remote sense	2047	0,146%
2	point cloud	794	0,057%
3	study area	578	0,041%
4	time series	496	0,035%
5	spatial resolution	460	0,033%
6	land cover	343	0,024%
7	applied earth observation	314	0,022%
8	neural network	302	0,022%
9	water body	270	0,019%
10	satellite image	261	0,019%
11	ieee trans	259	0,018%
12	glacial lake	253	0,018%
13	ecosystem service	214	0,015%
14	deep learning	204	0,015%
15	vegetation index	187	0,013%
16	white mica	180	0,013%
17	overall accuracy	179	0,013%
18	random forest	173	0,012%
19	satellite datum	165	0,012%
20	satellite imagery	164	0,012%
21	semantic segmentation	160	0,011%
22	impervious surface	158	0,011%
23	proposed method	157	0,011%
24	lidar datum	155	0,011%
25	sensing datum	149	0,011%
26	canopy height	148	0,011%
27	spatial distribution	147	0,010%
28	earth obs	140	0,010%
29	remote sensing image	139	0,010%
30	sensing image	139	0,010%
31	land surface	137	0,010%
32	version of this article	133	0,009%
33	vegetation indices	132	0,009%
34	figure legend	132	0,009%
35	convolutional neural network	132	0,009%
36	remote sensing datum	131	0,009%
37	web version	131	0,009%
38	spectral band	128	0,009%

eISSN: 2550-2131
ISSN: 1675-8021

39	crop yield	128	0,009%
40	forest structure	127	0,009%
41	change detection	127	0,009%
42	landslide susceptibility	125	0,009%
43	spectral reflectance	123	0,009%
44	interpretation of the references	122	0,009%
45	learning model	122	0,009%
46	training sample	117	0,008%
47	vegetation type	114	0,008%
48	earth observation	112	0,008%
49	x for peer	111	0,008%
50	dl model	109	0,008%
51	combination feature	107	0,008%
52	spatial pattern	106	0,008%
53	access article	105	0,007%
54	temporal resolution	105	0,007%
55	open access article	103	0,007%
56	burned area	99	0,007%
57	training datum	98	0,007%
58	nm combination	97	0,007%
59	stream boundary	97	0,007%
60	surface reflectance	97	0,007%
61	forest canopy	97	0,007%
62	leaf area	97	0,007%
63	normalized difference	96	0,007%
64	nm combination feature	95	0,007%
65	structural type	94	0,007%
66	qinghai lake	94	0,007%
67	atmospheric correction	92	0,007%
68	image classification	92	0,007%
69	ground truth	91	0,006%
70	m resolution	90	0,006%
71	area index	89	0,006%
72	laser scan	88	0,006%
73	cotton field	87	0,006%
74	spectral information	85	0,006%
75	leaf area index	85	0,006%
76	ground photon	84	0,006%
77	classification accuracy	84	0,006%
78	urban village	84	0,006%
79	nighttime light	79	0,006%
80	snow depth	78	0,006%

eISSN: 2550-2131
ISSN: 1675-8021

APPENDIX F

AFFIXES IN THE RSA CORPUS BY FREQUENCY

Item	Affix	Number of hits for affix	Relative frequency	Type
1	in	5888	0,4196%	Prefix
2	co	4065	0,2897%	Prefix
3	able	2685	0,1913%	Suffix
4	pre	2023	0,1442%	Prefix
5	multi	1797	0,1280%	Prefix
6	ize/yze	1560	0,1112%	Suffix
7	dis	1535	0,1094%	Prefix
8	inter	1366	0,0973%	Prefix
9	ex	1281	0,0913%	Prefix
10	un	1215	0,0866%	Prefix
11	photo	1171	0,0834%	Prefix
12	sub	1126	0,0802%	Prefix
13	ab	951	0,0678%	Prefix
14	logy	878	0,0626%	Suffix
15	eco	825	0,0588%	Prefix
16	pro	805	0,0574%	Prefix
17	bio	733	0,0522%	Prefix
18	non	671	0,0478%	Prefix
19	out	544	0,0388%	Prefix
20	max	524	0,0373%	Prefix
21	auto	456	0,0325%	Prefix
22	mini	385	0,0274%	Prefix
23	dynam	353	0,0252%	Prefix
24	bi	269	0,0192%	Prefix
25	contr	260	0,0185%	Prefix
26	ism	181	0,0129%	Suffix
27	uni	177	0,0126%	Prefix
28	radio	175	0,0125%	Prefix
29	post	171	0,0122%	Prefix
30	under	160	0,0114%	Prefix
31	hood	155	0,0110%	Suffix
32	arti	110	0,0078%	Prefix
33	propag	99	0,0071%	Prefix
34	alter	96	0,0068%	Prefix
35	an	91	0,0065%	Prefix

36	a	77	0,0055%	Prefix
37	ise	76	0,0054%	Suffix
38	demo	23	0,0016%	Prefix
39	hypo	23	0,0016%	Prefix
40	fusion	16	0,0011%	Suffix
41	de	13	0,0009%	Prefix
42	anti	12	0,0009%	Prefix
43	retro	12	0,0009%	Prefix
44	dom	10	0,0007%	Suffix
45	eu	9	0,0006%	Prefix
46	des	8	0,0006%	Prefix

ABOUT THE AUTHORS

Andrey S. Korzin, Assistant Lecturer at the Department of Foreign Languages of the Academy of Engineering, RUDN University. The author of several research articles, textbooks, and teaching materials. His research focuses on corpus and computational linguistics, lexicology, translation studies as well as English for specific purposes (ESP) and Academic English.

Anna S. Zhandarova, Master's student at the Faculty of Romanic and Germanic Languages, Moscow Region State University. Her research focuses on the usage of linguistic corpora for analysing public opinion in media discourse, as well as for the development of teaching materials focused on English for specific purposes (ESP).

Yana A. Volkova, Doctor of Philology, Professor at the Department of Foreign Languages in Theory and Practice, RUDN University. Her research focuses on linguistics of emotions, in particular disruptive communication. Her other research interests include ESL and psycholinguistic studies of emotions. She is the author of more than 90 papers in this field.