Corpus-Based Vocabulary Learning in Technical English

Tamara Polić *tamarapolic5@gmail.com*
Traffic and Transport Department
Polytechnic of Rijeka
Vukovarska 58, 51000 Rijeka, Croatia

Elena Krelja Kurelović *elena@veleri.hr*
Business Department
Polytechnic of Rijeka
Vukovarska 58, 51000 Rijeka, Croatia

Abstract

One of the main challenges posed in front of English for Specific Purposes (ESP) teachers in developing syllabi at higher education level is the choice of vocabulary to be taught. This issue is particularly prominent in technical English, which apart from being abundant in nouns, requires students to learn the other highly frequent noun-based structures such as multi-word lexical units (MWLUs). Learning how to cope with these condensed structures both in reading and writing, will make the students competent and self-confident ESP users. The choice of lexical items is usually left to teachers' intuition. This paper intends to assist teachers in avoiding addressing the issue by making such intuitive decisions, offering the model of incorporating the corpus-based vocabulary findings into their ESP syllabi instead. Thus, the research questions addressed in this paper are: Can a computer software extract all the nouns, MWLUs and multi-noun lexical units (MNLUs) with 100% certainty? What is their precise number in the pedagogical corpus of English for Traffic and Transport Purposes (ETTP)? The paper approaches the issue from the analytical point of view, i.e. by instructing the teacher-researcher step by step in analysing the pedagogical specialized corpora, including the possible problems they might encounter using irreplaceable, yet not completely accurate computer software for the purpose. The paper proposes original solutions to overcome the encountered imperfections in order to get accurate and evidence-based lists of most frequent nouns, MWLUs and MNLUs, making some extra effort by manually complementing the computer-based analysis. By applying the methodology proposed by this paper teachers-researchers will no more have to wonder which nouns and MWLUs and MNLUs to teach, since they can create accurate lists by themselves. Furthermore, a specialized pedagogical corpus analysis provides a valuable basis for creating glossaries and specialized minimum dictionaries, serving as a source for creating syllabi and lexis oriented exercises, as well as for designing language tests – all of it with the ultimate scope of improving students' lexical competencies in a specific field of study.

Keywords: Lexical competencies, Corpus, Multi-word Lexical Units (MWLUs), Teaching, Technical English.

1. INTRODUCTION

One of the basic goals of teaching English for Specific Purposes (ESP) to non-native speakers at the tertiary education level, possibly secondary if they are taught vocational English, is to develop their lexical competencies. Technical English seems to be especially demanding, since multi-word lexical units (MWLUs) and multi-noun lexical units (MNLUs) are the lexical features occurring more frequently in technical English than in any other ESP [1]. More than General English (GE), ESP allows a sequence of premodifiers to be placed in front of the head noun, without functional words which could assist in establishing semantic relationship among them [2]. Moreover, the head noun premodifiers can be exclusively nouns [3], [4]. Not to mention that nouns are the most common category of words in the text and that on average every fourth word

in the text is a noun [5]. In academic prose, nouns are used twice as often as in everyday speech, which is explained by informativeness as a characteristic of the style of academic prose [5]. This is important for this paper, since the pedagogical specialised-genre corpus examined in this work is based on such texts. The other prominent authors confirm it by finding that the ratio of nouns in specialist discourse vs GE is 44%: 28% and that nouns together with adjectives cover 60% of the overall lexis [6].

The choice of nouns and lexis to be taught in the classroom is usually left to ESP teachers' own intuition, unless they resort to the analysis of the corpus they will compile of the texts used in teaching. It is not an easy task, but although labour intensive, it is worth the effort. A proper corpus compiled of authentic texts and a good concordancer may be very useful in ESP classes [7]. This paper is intended for ESP teachers, primarily those teaching technical English, as well as to their students, aiming at enhancing both learning and teaching of ESP lexis.

Computer corpus linguistics facilitates the analysis in a systematic way [8]. The work is based on the example of a detailed analysis of a pedagogical corpus comprising all the ESP texts (materials) that students of a particular study group have to deal with during their undergraduate professional (BSc) study. The aim of the work is to provide ESP teachers with a possible model of semi-automated method for creating an accurate list of the most frequent nouns in the corpus (appearing four times or more, $f \geq 4$), MWLUs and finally MNLUs, explaining in detail methodological procedures employed in gaining the results which can be then reproduced for any particular ESP. In a corpus analysis human aid is necessary, because no software is almighty and perfect. The semi-automated method helps to identify language patterns that might otherwise be inaccessible [9]. Approaches permitting mixed methods in a research are welcome, complementing the findings of other methods [10]. However, the aim of this paper is not to evaluate or criticize the irreplaceable computer programs, but to give ESP practitioners and researchers an insight into obstacles and imperfections they may encounter in a computer-aided corpus analysis, and possible solutions to overcome them.

Hence, our approach is of inductive analytical nature: the analysis of a specialized-genre pedagogical corpus with the aid of Sketch Engine and AntConc tools, assisted by manual analysis.

The article can be linked to the other ones in the International Journal of Computational Linguistics, regarding corpus-based approach in linguistic research [11], [12] and [13].

2. BACKGROUND

According to what was stated in the introduction part, this research is based on three principal intertwined backgrounds:

- nouns and their modification with emphasis on premodification resulting in
- MWLUs and MNLUs, and
- corpus studies and English language teaching.

2.1 Nouns and Their Premodification

The underlying grammatical category dealt with in this paper is a noun, defined by Huddleston and Pullum [14: 83] as "a grammatically distinct category of words which includes those denoting all kinds of physical objects, such as persons, animals and inanimate objects." In their grammar books, the majority of prominent authors begin the chapters on nouns either with the classification(s) of nouns [15], [16], their characteristics [17], or immediately embed them into a chapter on noun phrases [18], [19] and [20]. It can be justified by the fact that the noun in English, as the most frequent grammatical category, apart from the possibility of appearing alone, appears as a basic constituent of a noun phrase. A noun is a head of a noun phrase and it can be premodified, postmodified, or both. This paper deals with a corpus analysis the aim of which is the extraction of nouns, and nouns together with their premodifiers, not including

postmodification. According to Biber et al. [21], in academic prose almost 60% of all noun phrases are modified, 25% being premodified. They state that various parts of speech can be found in premodification: determiners, adjectives, nouns, participles and adverbs. The analysis conducted in this paper excludes determiners. It focuses on MWLUs and MNLUs. They are primarily the result of premodification. Noun premodification can be single or multiple [21].

The following are the examples of multiple premodification taken from the ETTP corpus created for the purpose of this paper:

two-word premodification: *light rail* train
three-word premodification: *internal combustion engine* vehicle
four-word premodification: *electric mass transit railway* system

The type and number of premodifiers vary through different genres and registers. More complex premodifications are generally more common in scientific texts and speech than in everyday conversation [1], [2], [3]. Although there are theoretically no limits regarding the number of premodifiers, it is not common to encounter more than four, as too many premodifiers could lead to interpretive overload or uncertainty about meaning. Nonetheless, in technical English multiple premodification is rather frequent, resulting in MWLUs and MNLUs.

2.2 Multi-Word (MWLU) and Multi-Noun Lexical Units (MNLU) Relationship

The study of MWLUs is a very complex field, since there is no uniformity in nomenclature. In her work Špiranec [2] provides a list of various terms used to denote MWLUs, such as *composites, compounds, multi-word lexical units, multi-word expressions, multi-word lexemes, collocations, phrasal verbs, idioms, fixed sintagms, phraseologisms, lexicalized phrases.* Besides, we have encountered *multi-noun compounds* [22], *premodified complex noun phrases* [23], *different-component compounds* [24], choosing *the multi-word lexical units (MWLUs)* which seems to be the most widespread term, and adopting Kereković's [1] definition, according to which MWLUs are multi-word syntagms consisting of two or more words (combination of words) conveying lexical meaning as a whole, and functioning as a single lexeme in a sentence.

The following is an example of a MWLU, precisely four-word lexical unit, belonging to the English railway terminology (example taken from [25]), where a head noun *resistance* is premodified by three premodifiers belonging to various parts of speech:

curved		track		rolling		resistance
↓		↓		↓		↓
-ed participle	+	noun	+	*-ing* participle	+	*head noun*

Still, multi-word premodification allows a head noun to be premodified by a sequence of nouns (i.e. just one part-of-speech type). Resulting MWLUs, consisting exclusively of nouns and functioning as single lexemes are called in this paper *multi-noun lexical units (MNLUs)*, being considered a hyponymous category in respect to MWLUs (cf. [3], [4]). The term is used to avoid heterogeneous terminology encountered throughout relevant philological/linguistic publications: Lauer [26] found *compound nominal, nominal compound, compound noun, complex nominal, nominalization, noun sequence, compound, noun compound, noun-noun compound, noun+noun compound, noun premodifier*. In addition, we found *complex nominal* [27], [23], *noun collocation* [28] and *noun-centered compound noun* [29].

The following is an example of a MNLU, precisely a five-noun lexical unit, extracted from the analyzed corpus, where a head noun *problem* is premodified by a sequence of nouns:

rush		hour		traffic		congestion		problem
↓		↓		↓		↓		↓
noun	+	noun	+	noun	+	noun	+	*head noun*

Our pedagogical specialized corpus analysis aims to extract precisely the most frequent nouns, MWLUs, and MNLUs that students must learn in order to become competent ETTP users.

2.3 Corpus Studies and English Language Teaching

As stated by Hardie [30], "in the twenty-first century, corpus linguistics has become a 'killer method' in linguistics, applied to a hugely diverse array of types of linguistic research".

Practicing language teachers engage in corpus research for a variety of reasons, finding motivation for their research in the immediate teaching environment, or being motivated by their personal or academic interest, but crucial is their professional curiosity [31]. Our paper encompasses all the three sources of motivation, with the emphasis on enhancing ESP teaching. One of the main areas in which corpora can benefit language teaching is by incorporating corpus-based findings into language syllabi and teaching materials [32], and that is exactly the idea behind this research. Many authors have found (cf. [33], [34], [21], [23], [27], [35] etc.) that not only non-native English speakers experience difficulties in receiving/interpreting and producing MWLUs and MNLUs, but it is also challenging for native speakers as well, especially when they are not familiar with the subject field. Therefore, every ESP teacher should conduct a pedagogically oriented research by creating and analyzing the corpus of teaching materials (texts) aimed at preparing students for their future careers, extracting the discipline-specific vocabulary lists, primarily nouns, MWLUs and MNLUs, in order to avoid relying merely on his/her intuition as far as the choice is concerned. Even small corpora are useful in a sense that they help making decisions for teaching particular linguistic features [36], having a greater concentration of vocabulary [37].

3. RESEARCH

This paragraph includes general data about the corpus and research methodology.

3.1 About the Corpus – General Data

The research dataset is a small pedagogical specialized genre corpus. It is compiled of written texts (units) used for the compulsory courses of English for Traffic and Transport Purposes (ETTP) as a part of the curriculum of the undergraduate professional study of Traffic and Transport at an institution of higher education, where English is taught as the first foreign language.

ETTP is studied here during the first two academic years and organised in four courses (terms), *ETTP 1, ETTP 2, ETTP 3* and *ETTP 4*. There are three classes per week for 90 weeks, which makes a total of 270 classes. When freshmen/freshwomen start attending the ETTPS classes, they are considered independent users according to the *Common European Framework of Reference for Languages (CEFRL)* [38], at least B2 level.

The teaching/learning materials (texts) used for compiling the corpus cover just one, but the most important ETTP learning outcome that students are supposed to master during their studies, which is *the development of lexical competencies*.

The corpus containing 14,428 words is compiled of 27 texts taken from various sources: 16 from the textbook designed for studying ETTP, one from a specialized monograph, 8 texts taken from the Internet sources, and two specialized journal articles, shortened and adapted by their teachers.

3.2 Methodology

Research design, data collection and data analyses follow the generally accepted procedures of linguistic corpus research, naming in the paper the authors from whom the ideas were taken at the proper places. Yet, the offered inductive semi-automated methodology is original, arising from the need of identifying precisely the most frequent nouns, MWLUs and MNLUs to be learnt by the students of ETTP, which cannot be done by a software alone. We introduced extensive concordancing and hand-and-eye analyses in order to:

– distinguish with certainty nouns from other parts of speech, i.e. *–ing* forms, verbs, adverbs, adjectives and other parts of speech (section 4.2),
– deal with capitalized and non-capitalized words, British and American variants of the English language, orthography and lemmatization in terms of singular and plural forms (section 4.2),
– provide the list of premodifiers of the most frequent nouns resulting in the accurate list of multi-word lexical units (MWLUs) (section 4.3),
– provide the list of noun premodifiers of the most frequent nouns resulting in the accurate list of multi-noun lexical units (MNLUs) (section 4.4).

Besides, adopting hand-and-eye methodology and a calculator, we present the exact number and percentage of two, three, four, five and six-word lexical units in the analysed corpus, as well as two, three, four and five-noun lexical units, which as far as we know, has never been done before, particularly not for ETTP.

Statistically, the corpus comprises 14,428 tokens (defined by Hardie, [30:510] as "single instances of any word at a particular place in a corpus") and 3,028 types ("individual word-forms which can occur many times in the corpus" (ibid.).

The following paragraphs will cover the corpus creating procedure and steps undertaken in the corpus analysis.

3.2.1 Corpus Creating Procedure

After collecting all the texts to be included in the corpus, their format had to be established in order to convert them in Word format and in .txt. format subsequently, since corpus analysing tools require texts to be in .txt. format. The original texts were mainly in electronic format, either Word or .pdf. Those already available in Word were copied and pasted, while those in .pdf. format had to be converted by optical scanning (OCR = Optical Character Recognition), which is not entirely reliable, leaving the quality of the original text rather degenerated. Therefore, mistakes in the texts were further corrected using spell checker. Two of the texts were available only as printed materials and had to be copied by hand. All the non-lexical material such as graphs, photographs, tables, diagrams and drawings were removed, yet retaining the text describing them. The lexical items presented in the texts as lists (for instance car's interior and exterior parts, elements of periodic road maintenance, and the like) had to be separated by hand-inserted points at the end of each item, because otherwise the corpus analysing tool would consider them multi-word lexical units. The same was done with titles and subtitles, after which a period was added. The material needed further adaptation by stripping off all formatting coding such as for instance unequal word spacing and huge white-space areas occurring in the texts which were copied and pasted from web resources, as well as superscript circles / non-breaking spaces (the Internet source [39] was found very useful for the purpose). Since computers lack intelligence, the texts to be included in a corpus must not contain anything unexpected [40], so our texts were manually cleaned up from e.g. opening double quote which is coded 93, closing double quote coded as 94, em dash (long dash) coded 96, or apostrophe coded as 92. Thus, the contracted forms of auxiliary verbs *be, have* and *do* were transformed by hand in uncontracted forms (e.g. 'll → will/shall, 's → is, 've → have, don't → do not, etc.). Diacritical marks such as *š, ž, ô, é* were replaced by hand by their nearest orthographic counterparts (*s, z, o, e*). Upon completion of all the above-mentioned time consuming but indispensable procedures, the texts were ready to be converted in .txt. format and the corpus was built.

3.2.2 Steps in Corpus Analysis

The corpus data were analysed using online text analysis tools *Sketch Engine* and *AntConc 3.5.8 (Windows)* combined with manual analysis. Since the research is pedagogically oriented the teachers'/researchers' interventions were necessary, as Jones and Durrant [41:387] put it: "It is important to bear in mind that corpus software is not yet able to construct pedagogically useful word lists without substantial human guidance. Teachers wishing to create such lists will need to make a number of important methodological decisions, and to make these decisions well will need to understand the issues surrounding them." For the sake of clarity and easier following of

the sequence of procedures, teachers'/researchers' interventions are described in the *Corpus findings (analysis) and discussion* section, along with each list resulting from the analysis.

With the ultimate goal of creating a list of noun pre-modifiers of the most frequent nouns (*f* ≥ *4*), or multi-noun lexical units contained in the corpus, a gradual corpus analysis was conducted in several steps involving:

a) production of the frequency list (word list) in rank order
b) production of the list of the most frequent nouns in several steps
c) production of the list of premodifiers of the most frequent nouns
d) production of the list of noun premodifiers of the most frequent nouns

4. CORPUS FINDINGS (ANALYSIS) AND DISCUSSION

4.1 Production of the Frequency List (Word List)

The first common analytical step is the production of the frequency list (word list) which demonstrates that a comparatively small set of words accounts for a large proportion of text [42], or in other words, it presents the core lexis/lexicon. The frequency list of the analysed pedagogical specialized corpus based on all the texts used for teaching/learning *English for Traffic and Transport Purposes* as the first foreign language at an institution of higher education comprises 3,028 word types, demonstrating the predominance of grammatical words. Such a distribution is expected in terms of the general distribution of different items in the English language [32]. In the frequency list (Table 1) the first lexical (content) words are ranked fairly high, *traffic* appearing already at 9th position, *transport* at 18th, *engine* at 20th, *system* at 23rd, *road* at 25th, all of them being nouns. This is not surprising, since on average every fourth word in the text is a noun [5], which makes nouns the most frequent word category in a text. Thus, the following step in the corpus analysis is the production of the list of nouns.

N	*Word*	*Freq.*	*N*	*Word*	*Freq.*	*N*	*Word*	*Freq.*
1	the	966	15	with	86	29	vehicles	51
2	of	455	16	by	83	30	at	50
3	and	448	17	be	81	31	has	48
4	to	344	18	transport	81	32	locomotives	45
5	a	311	19	from	78	33	power	45
6	in	299	20	engine	73	34	electric	43
7	is	211	21	an	70	35	car	42
8	are	147	22	it	68	36	combustion	41
9	traffic	124	23	system	68	37	have	41
10	as	112	24	which	67	38	trains	40
11	for	112	25	road	66	39	but	40
12	on	110	26	one	55	40	also	38
13	that	105	27	this	53			
14	or	97	28	can	51			

TABLE 1: Frequency list (rank order) of first 40 words.

4.2 Production of the List of the Most Frequent Nouns

In academic texts nouns are used twice as much than in everyday speech, which is explained by the informative character of the academic style of writing [5]. In addition, the strategy of learning the most frequent words appearing in ESP as very useful, since knowing them facilitates the understanding of a specialized text and saves time needed to check the meaning in the dictionary [45]. In view of the aim of this work which is of pedagogical nature, considering the learner and his/her needs as the centre of teaching/learning process [43], we proceeded by the production of the list of the most frequent nouns to be learnt by our students, extracting those appearing four times or more (*f* ≥ *4*). By part-of-speech (POS) tagging a list of 285 nouns was produced. In order to enable manual interventions, the list was converted in Excel. The manual intervention consisted of stripping off proper names, such as *France, Britain, Europe, Rimac, Brajdica* and the

like, unless they were constituent and indivisible elements of multi-word lexical units such as *Wankel rotary engine* or *Diesel engine,* which was checked by means of concordancing. The common nouns making part of a name were also subjected to further analysis by concordancing. The screenshot in Figure 1 shows concordances of the noun *concept,* giving a figure of 17 occurrences in the list of nouns.

FIGURE 1: Concordances of the noun *concept.*

The co(n)text-based information revealed that the noun *concept* appeared 15 times as a part of the name of the electric hypercar *Concept One,* produced by the manufacturer *Rimac Automobili,* and just two times as a common noun. Therefore, it was discarded from the list of the most frequent nouns (*f* ≥ *4*).

Upon completion of this procedure, the software-generated list of most frequent nouns (*f* ≥ *4*) was reduced from the original 285 nouns to 257. These nouns are going to be named *presumable nouns* for the purpose of this paper/analysis until their belonging to the category of nouns is entirely proved, because although immensely useful, POS taggers have a residual error rate of 3-5 per cent which can only be removed by manual post-editing [30]. The errors arise in the first place because of the polyfunctionality of words in the English language, i.e.one the same word can have more than one function and belong to the various parts of speech, which the software cannot always discern. For instance, *crossing, drive, track* can be both nouns and verbs, *today* can be both a noun and an adverb, while *level, public, glass* can function both as nouns and adjectives. Furthermore, some words appearing in the software-produced list of nouns, such as *case*, *order*, or *contrast,* belong to the category of conjunctions, (set) phrases or idioms e.g. *in (which) case, in order to / that / for, in/by contrast,* and therefore should be excluded from the list of most frequent nouns. Discerning nouns from other parts of speech can only be done by examining them in their original co(n)text (linguistic environment) through an extensive concordance search. For the purpose of this paper, more than 2,000 concordances have been investigated. On the basis of information obtained from concordances, the software-generated list of nouns, copied and pasted into Excel sheets, was manually corrected in steps/phases. Concordancing was also applied in the other analytical procedures which are explained in detail along with the examples in the following text (paragraphs 4.2.1 – 4.2.7).

The following presents the procedures employed to produce an accurate list of nouns (*f* ≥ *4*) appearing in the corpus. As a result, some of the 257 items appearing in the list of nouns were removed from the list if their frequency dropped under *f* ≥ *4*, while the frequency rank of others might have changed.

4.2.1 Discerning Nouns from *-ing* Forms

In line with the observation that the software is not always sophisticated enough to pick up various similarities [44], we have used concordances to establish whether all the listed nouns

ending in *-ing* such as *engineering, steering, beginning* are really nouns, or might be present participles (in function of a participle clause), or parts of the continuous form of a verb. The examination showed that all of them were used as nouns.

4.2.2 Discerning Nouns from Verbs

One of the software advantages in the creation of the list of nouns is the automatic summing up of singular and plural forms of nouns under one single entry, subsuming the plural forms under their singular forms and showing their cumulative frequency. Yet, there remains a possible confusion whether for instance *use(s), drive(s), speed(s), design(s)* are nouns or the Present Simple tense verbs, which can only be established by concordancing, as shown in Figure 2 for the presumable noun *drive*.

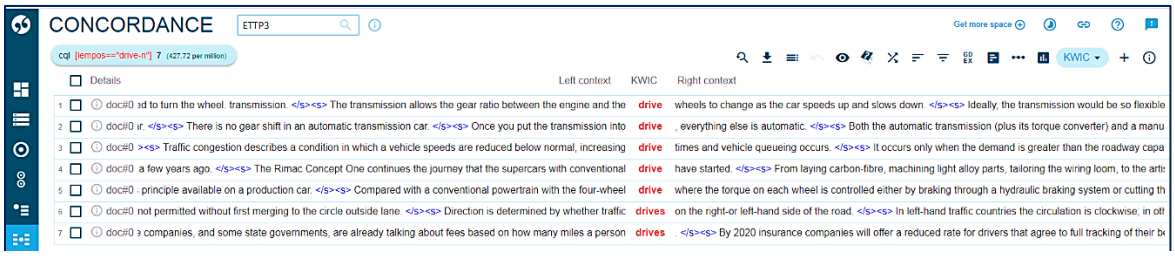

FIGURE 2: Concordances of the presumable noun *drive.*

The word *drive(s)* appears seven times on the list of the most frequent nouns, yet the concordancing proved that that five times it appears in the function of a noun and two times as a verb. Therefore, its position on the list of the most frequent nouns changes from *f = 7* to *f = 5*.

The initial frequency of the word *start* is *f = 5*. Its concordances show (Figure 3) that it appears two times as a verb and three times in the function of a noun. Consequently, it was removed from the list of the most frequent nouns because its frequency dropped under *f ≥ 4.*

FIGURE 3: Concordances of the presumable noun *start.*

Upon completion of examining the rest of presumable nouns by applying concordancing, the frequency of some nouns changed, while the list of most frequent nouns resulted in 256.

4.2.3 Discerning Nouns from Adverbs

Some words may appear as adverbs or as nouns. For instance, *today* or *lot*. Concordances showed that in all four instances word *today* appears as a temporal adverb, just as *lot* appears in all the cases as an adverb modifying quantity (Figure 4) and therefore were discharged from the list of the most frequent nouns, which after this procedure resulted in 254 presumable nouns.

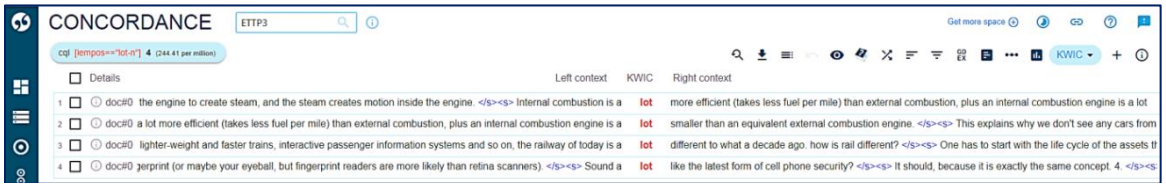

FIGURE 4: Concordances of the presumable noun *lot.*

4.2.4 Discerning Nouns from Adjectives

Next step was discerning words that can be both nouns and adjectives such as *light, fluid, level* and the like, as shown on the example of the word *motive* (Fig. 5) which is in all 13 instances used as an adjective meaning 'producing or causing movement' and not once as a noun meaning 'the reason that you do something'. Hence it was discarded from the list of the most frequent nouns.

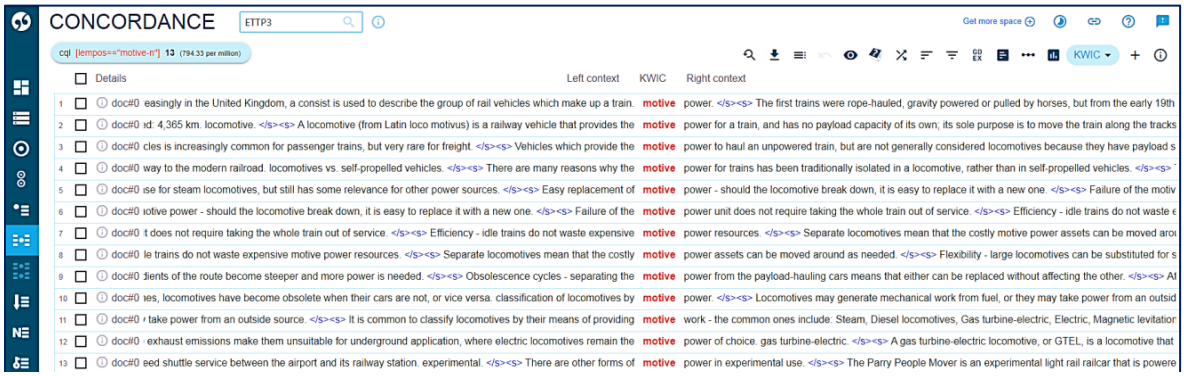

FIGURE 5: Concordances of the presumable noun *motive.*

Following these analyses done by concordancing, the list has come down to 250 most frequent nouns.

4.2.5 Discerning Nouns from Other Parts of Speech

A number of nouns listed as the most frequent nouns in the corpus belong to the other parts of speech, appearing in conjunctions, (set) phrases or idioms, like *case* in *in case (in which case, in all the cases, in the case),* or *order* in *in order to/that/for, or contrast* in *in/by contrast,* which is illustrated in Figure 6 showing the concordances of the word *order.*

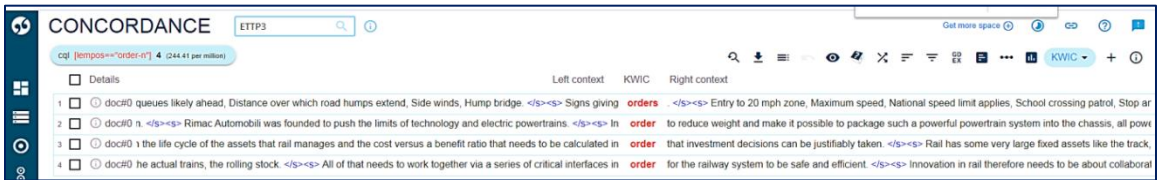

FIGURE 6: Concordances of the word *order.*

After this analytic procedure, there remained 247 nouns on the list of the most frequent nouns (*f* ≥ *4*) which after all the researchers'/teachers' interventions need not be called *presumable* any more, since they have been discerned from the other parts of speech and established as actual nouns.

After nouns had been discerned from the other parts of speech, some further minor researchers'/teachers' interventions proved necessary, which is presented in the following two paragraphs.

4.2.6 Capitalised Words, British and American Variants, Orthography

Since the software treats capitalized and non-capitalized tokens as different, one more intervention on the part of researchers/teachers was needed, such as in case of the noun *transport* which appears 76 times as non-capitalized (*transport*) and four times as capitalized (*Transport*). So, we listed it as one single entry with *f = 80*.

Furthermore, the corpus being compiled of the texts written in both British and American variant of the English language, the same noun appears as British English (BE) *transport* 80 times, and 8 times as American English (AE) *transportation*. So, we summed the frequencies of both the variants up into one type, using brackets: *transport(ation), f = 88,* which now becomes 4th most frequent noun on the list.

As for the other orthographic differences between the two variants they were solved in a way that both variants were maintained as one entry, the more frequent one being at the initial position, while their frequencies were summed up, as for instance *centre/center f = 4.*

The synonymous nouns were summed up under one entry, maintaining both the nouns, as for instance British *railway* which appears 46 times, and American *railroad* appearing three times. We consider it important that the students learn both the variants of the English language. On the list of the most frequent nouns they appear as *railway/railroad (49),* the more frequent one being written first.

Subsequent to these interventions, the final list of the most frequent nouns comprises 244 nouns.

4.2.7 Singular vs Plural Nouns

As mentioned in paragraph 4.2.2 of this paper, the SE software lemmatizes nouns in sense of subsuming immediately plural nouns under their singular counterparts. We can see it by clicking on concordances. For the majority of nouns it is plausible, such as for instance *car/cars,* but it is not suitable for pluralia tantum nouns as *goods* (*f = 16*) appearing in the corpus only in the sense of 'merchandise', or for the invariable nouns ending in *-s, for* instance *logistics* (*f = 6*) wrongly lemmatized by the software into *good* and *logistic*. Interesting is also the example of a foreign noun *datum* (*f = 16*), lemmatized by the software in the singular form, although in the corpus it occurs exclusively in its plural form i.e. *data*. Regarding the last one *data,* we hold it important that for teaching purposes ESP teachers know whether individual nouns appear in the corpus (teaching materials) in singular form at all. These nouns have been corrected by hand and therefore appear in their plural forms in the final list of the most frequent nouns (Table 2) with no impact on the final number of 244 most frequent nouns.

Noun	***Freq.***	***Noun***	***Freq.***	***Noun***	***Freq.***
traffic	121	driver	25	service	16
engine	100	country	25	goods	16
road	89	motor	24	direction	16
transport	88	time	23	section	15
system	81	speed	22	unit	15
train	75	transmission	22	design	15
vehicle	74	gas	22	air	15
car	65	terminal	21	bus	15
locomotive	64	way	21	junction	14
railway	49	construction	20	use	13
motorway	43	area	20	network	13
container	42	route	19	line	13
power	41	wheel	18	congestion	12
combustion	39	corridor	17	flow	12
roundabout	37	type	17	rule	12
rail	35	turbine	17	percent	12
fuel	30	track	17	water	12
passenger	30	data	16	cylinder	12
steam	30	technology	16	oil	12
part	27	intersection	16	form	12

TABLE 2: Final frequency list (rank order) of first 60 nouns (*f ≥ 4*).

4.3 Production of the List: Premodifiers of the Most Frequent Nouns – Multi-word Lexical Units (MWLUs)

The next step towards the creation of the list of the most frequent noun premodifiers of the most frequent nouns (MNLUs) was the creation of the list of all the other premodifiers of the most frequent nouns, that is, the list of multi-word lexical units appearing in the corpus. In order to provide this list, extensive manual analysis was required as well, since, as Jones and Durrant [41:388] noted, "automated corpus analysis tools are not yet able adequately to distinguish between different senses of words." It was proved by our analyses when choosing very useful, but not perfect option of *keywords – multi-words,* offered by the software. By applying our decent knowledge of the subject matter and using concordances, it resulted that many of the computer-generated multi-words were not multi-words at all, having no meaning, such as: *driver can*, *urban area close, geometrical road, term internal combustion, hour traffic congestion, long distance container freight.* Figure 7 illustrates the concordances of the last one, which in the multi-word list should have been presented as *long distance container freight transport*, but the head noun *transport* is obviously missing. Therefore, for the teaching/learning purposes, such a list cannot be used without eye-and-hand interventions.

FIGURE 7: Concordances of the multi-word *long distance container freight.*

For this reason, we resorted to time demanding but indispensable manual extraction of the most frequent modifiers of the most frequent nouns, that is MWLUs, by using concordances.

The following is the extract from our handmade list of premodifiers of the most frequent nouns ($f \geq 4$) in the corpus. The display of all the premodifiers of all the most frequent nouns would exceed the scope of this paper, so as an example, in Table 3 we show a list of all premodifiers (arranged in alphabetical order) of only one of the most frequent nouns, *engine* ($f = 100$). It results that this noun is premodified 33 times by various parts of speech.

Many authors such as [5], [19], [21] and [45] agree that four most common structural types of noun premodification in English are: adjective, *-ed* participle, *-ing* participle and noun. In our list of premodifiers of the most common nouns (the extract from which is shown Table 3), we find examples of each of the individual structural types of premodification. The following are the examples taken from the corpus (premodifiers marked in italics):

- by adjective: *electric* car, *ideal* motorway, *natural* gas, *small* engine
- by *-ed* participle: *unnamed* locomotive, *motorized* vehicle, *loaded* container
- by *-ing* participle: *braking* system, *stacking* area, *compressing* turbine,
- by noun: *steam* locomotive, *iron* rail, *liquid* fuel, *traffic* speed, *gas* engine.

Premodifiers				Noun
1)	air-cooled	17)	high-performance	**engine** (*f = 100*)
2)	car	18)	internal combustion	
3)	cold	19)	jet	
4)	compression ignition	20)	reciprocating	
5)	diesel	21)	reciprocating piston	
6)	external combustion	22)	rocket	
7)	fine	23)	rotary	
8)	commercially successful internal combustion	24)	six-stroke piston	
9)	internal combustion	25)	small	
10)	modern internal combustion	26)	spark ignition	
11)	four-cylinder	27)	stationary	
12)	fuel-injected	28)	steam	
13)	gas	29)	turboshaft	
14)	gas turbine	30)	two-stroke gasoline	
15)	gasoline car	31)	two-stroke piston	
16)	heat	32)	Wankel	
		33)	Wankel rotary	

TABLE 3: List of premodifiers of the noun *engine.*

In ESP chemical formulas and abbreviations appear as premodifiers [23]. We, on the other hand, find units of measurement and numbers, both cardinal and ordinal: *one-way road, three-lane roundabout, four-wheel drive, four-stroke engine, six-stroke piston engine, twenty-foot equivalent unit (TEU), first-class compartment, third rail, 12-volt power, 250,000 miles, the first motorways, 8 passenger seats, 500 sensors, eleven different systems*, etc. In our list of premodifiers of the most common nouns in the corpus, we have kept only those numbers and units that are integral parts of MWLUs and the removal of which would compromise their meaning. The rest of numbers were discarded as premodifiers. For instance, *third rail* is an indivisible two-word lexical unit meaning 'an additional rail supplying electric current', used to supply traction vehicles with electricity, while the *four-stroke engine* is an indivisible three-word lexical unit meaning 'the most common type of internal combustion engines', and *twenty-foot equivalent unit (TEU)* is 'an inexact unit of cargo capacity often used to describe the capacity of container ships and container terminals, based on the volume of a 20-foot-long (6.1 m) intermodal container'. By removing number premodifiers, these MWLUs would lose their meaning. Precisely such MWLUs including numbers in premodification are one of the features of technical ESP, and thus of ETTP as well. Accordingly, examples of premodification by numbers and units of measurement such as in the aforementioned *8 passenger seats, 500 sensors, eleven different systems,* where numbers are not constituent elements of MWLUs, have been removed from the list. We also removed determiners (articles, demonstrative pronouns, personal pronouns, possessive pronouns and quantifiers), but retaining as premodifiers proper names like *Wankel* and *Diesel* in the two-noun lexical units *Wankel engine* and *Diesel locomotive*, or three-word lexical units *Wankel rotary engine* and *Diesel-electric unit*. These MWLUs feature in specialized technical dictionaries as inseparable entries.

Besides, there have been no manual interventions in the orthography of multi-word premodifiers. They were included in the list of premodifiers of the most common nouns in their original form, as they were found in the text. For this reason, the list of premodifiers of the most common nouns sometimes includes premodifiers written with a hyphen (*internal-combustion-engine, air-cooled, high-performance*), and sometimes without it (*internal combustion engine, air cooled, high performance*). However, for the purposes of this research, hyphenated premodifiers were treated as separated words. In view of teaching/learning, we consider it necessary for students to notice the orthographic diversity and get used to it in order to find their way in the specialized literature, and this corpus analysis is intended primarily for them and their ETTP teachers.

The results of further analysis of the most frequent nouns in our specialized-genre corpus (traffic and transport) indicate that 44 (18.03%) out of 244 most frequent nouns (*f ≥ 4*) are not

premodified at all, while the other 200 (81,97%) are premodified by the total of 798 premodifiers belonging to various parts of speech. These findings do not coincide with Biber et al. [21] who found that in academic prose almost 60% of all nouns have some modifier, but are closer to Štambuk's [46] conclusion that MWLUs (premodified nouns) constitute 40-70% entries in the specialized electronics dictionary.

Biber et al. [21] found that in all the four registers they investigated (conversation, fiction, news and academic prose) 70 - 80% premodified nouns have only one premodifier, about 20% have two premodifiers, while just 2% have three or four premodifiers.

Our findings indicate that in our specialized corpus dealing with traffic and transport 563 most frequent nouns are premodified by one premodifier (70.56%), most of them constituting 2 word lexical units[1], 176 (22.06%) have two premodifiers constituting 3 word lexical units, 43 (5.38%) have three premodifiers constituting 4 word lexical units, 15 (1.87%) have four premodifiers constituting 5 word lexical units, and 1 (0.13%) noun has five premodifiers, constituting a 6 word lexical unit, as illustrated in Figure 8.

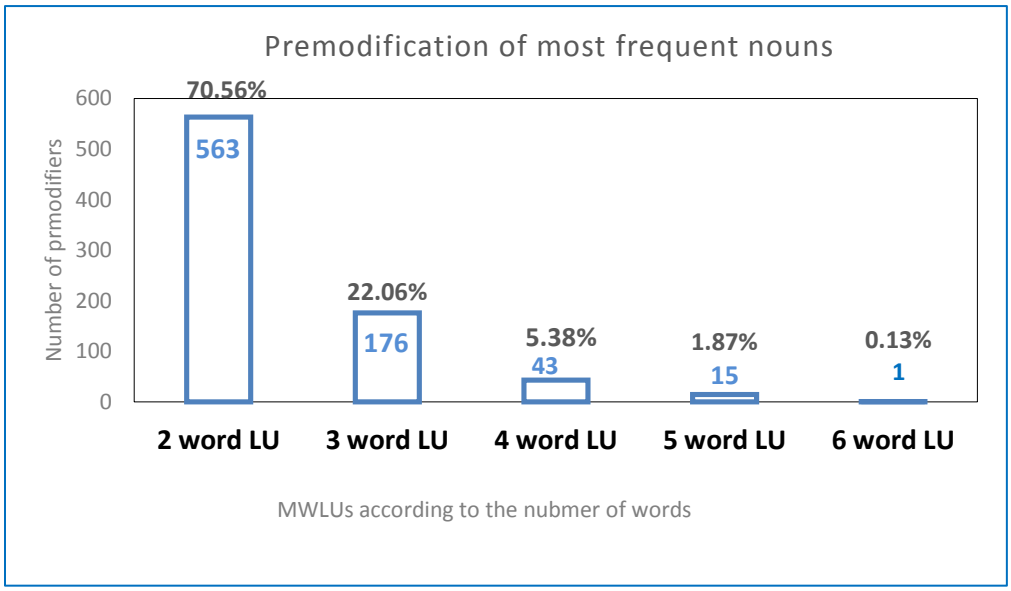

FIGURE 8: Premodification of the most frequent nouns.

The reason for higher incidence of nouns premodified by three, four and even five words in our corpus might lie in the fact that our corpus deals with ESP, namely technical English (English for Traffic and Transport Purposes), and proves what was long ago established by Bartolić [47:260] that in technical English information tends to be "conveyed in a more condensed form which has a greater impact upon the reader".

4.4 Production of the List: Noun Premodifiers of the Most Frequent Nouns – Multi-Noun Lexical Units (MNLUs)

The ultimate goal of this research was the creation of the list of noun premodifiers of the most frequent nouns. The reason behind it is the possible "teachability" of multi-noun lexical units (MNLUs) as MWLU's hyponymous category, as presented in works [3] and [4]. Since reception and production of MNLUs can be taught and learnt to a decent level, it serves as a basis for

[1] The authors of this paper are aware of the fact that in English grammar some of the two-word lexical units appearing on our list, consisting of an adjective premodifying a head noun (for instance *cold engine*) are not really MWLUs, but merely premodified nouns.

teaching MWLUs. Once the students have mastered MNLUs reception and production, dealing with MWLUs as hyponymous category to MNLUs, becomes easier. When the students had learnt how to establish semantic relationships among the constituent nouns in a MNLUs, establishing the semantic relationship between the other premodifiers (belonging to other parts of speech, predominantly adjectives) and a head noun in a MWLU is facilitated. Long ago, it was established that students who learn ESP as a foreign language (L2) cope well with the MWLUs in which a head noun is premodified by an adjective, but experience difficulty with those where a head noun is premodified by noun(s), i.e. MNLUs [27]. Premodification by an adjective can be deceiving not just for ESP learners, but for native non-expert readers as well. The knowledge of the subject matter is indispensable in establishing to which of a MWLU constituent nouns a premodifying adjective refers to.

For instance, in a MWLU *General System Theory*, only the field-expert reader knows whether the *system* is *general* or is the *theory*. Although such expert knowledge is also very helpful in reception and production of MNLUs, there exist several rules and hints as well, which can be very useful [4]. Prior to teaching/learning MNLUs, the teacher's task is to analyse the pedagogical corpus and extract MNLUs the head nouns of which are the most frequent nouns in the corpus.

As in case with the production of the list of MWLUs, i.e. premodifiers of the most frequent nouns in the corpus that belong to various parts of speech, the software proved itself not being sophisticated enough to provide a completely perfect list of noun premodifiers, that is of MNLUs. The production of pedagogically useful word lists requires considerable human guidance, since semantic tagging is a hard task for a computer because it does not understand the context, and words can be ambiguous for both grammatical category and semantic field [38]. We proved this by concordances (Figure 9) of presumable MNLUs offered by the software.

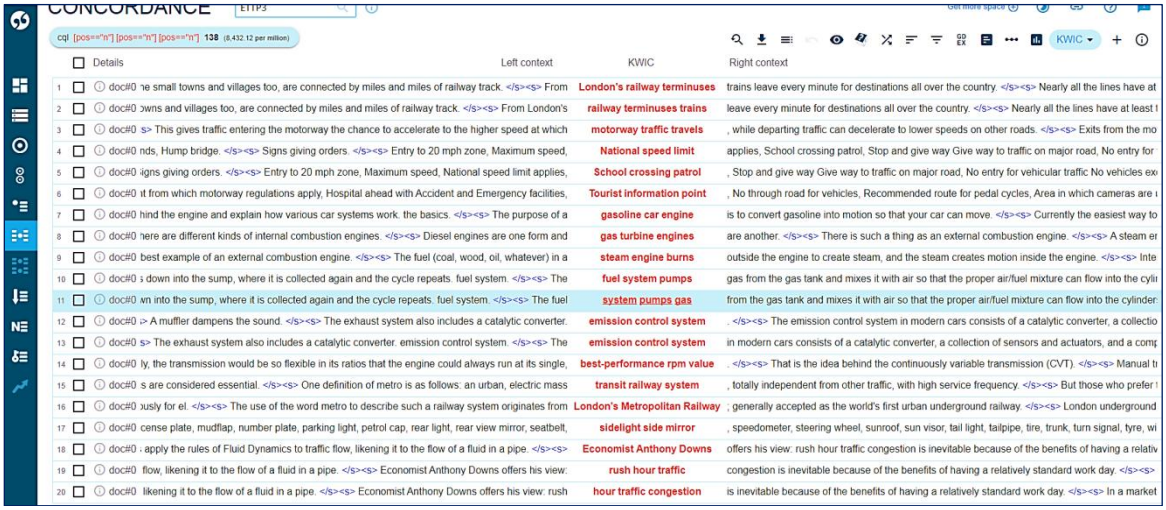

FIGURE 9: Concordances of MNLUs.

The highlighted concordance in Figure 9 indicates that the software extracted *system pumps gas* as a MNLU, considering *pumps* a noun instead of a verb, and not recognizing *fuel* as a noun premodifier in a two-noun lexical unit *fuel system*.

Following guidelines for the corpus analysis suggesting to take into account all the examples in our corpus relevant to what we are investigating if we want our analysis to be totally accountable to the corpus data [30], we created the list of the most common noun modifiers by an extensive hand-and-eye intervention. As an example, from the list of noun premodifiers of the most frequent nouns we extracted the list of noun premodifiers (arranged in alphabetical order) of the noun *system (f =81),* shown in Table 4. This noun is premodified 42 times by (a) noun(s). It is interesting to observe that the noun *system* is not the most frequent noun in the corpus, yet it is

most frequently premodified by nouns than any other noun. In other words, the noun *system* is the most frequent head noun appearing in MNLUs in the analysed corpus. This may be explained by the fact that the same noun in MNLUs can appear both as a premodifier and as a head noun [5]. In our corpus the noun *system* appears in premodifying function as well: *system tunnels, drainage system improvement*.

Noun premodifiers				Noun
1)	air-intake	22)	metro	**system** *(f = 81)*
2)	torque vectoring	23)	motorway	
3)	braking	24)	passenger information	
4)	car	25)	passenger transport	
5)	conditions observing	26)	power	
6)	cooling	27)	powertrain	
7)	data transfer	28)	public address	
8)	drainage	29)	railway	
9)	driver override	30)	road	
10)	emission control	31)	road transport	
11)	exhaust	32)	sprinkler	
12)	fire alarm	33)	starting	
13)	fire extinguishing	34)	steering	
14)	freight transport	35)	trackside	
15)	fuel	36)	traffic flow monitoring	
16)	ignition	37)	traffic sings operation	
17)	information	38)	train power	
18)	infotainment	39)	transportation	
19)	lighting	40)	underground	
20)	lubrication	41)	ventilation	
21)	mass transit railway	42)	video surveillance	

TABLE 4: List of noun premodifiers of the noun *system.*

Our research has shown that 103 (42.21%) out of 244 most frequent nouns (*f* ≥ *4*) in the corpus are not premodified by (a) noun(s), while the remaining 141 (57.79%) have noun premodification. This finding does not coincide completely with the findings of other researchers dealing with different fields of ESP: Biber et al. [21] find that nouns account for almost 40% of all premodifiers in news, and about 30% of all premodifiers in academic prose. Gačić [5], [45] finds 30 – 40% noun premodifications in academic prose. Seljan and Gašpar [48], analysing legislative documents, find that 22 – 24% nouns are premodified by a noun. Seljan, Dunđer and Gašpar [49] dealing with philosophical and sociological texts find that nouns account for only 2.1% of all premodifiers. Ang, Tan and He [28] found 44% - 52% noun premodifiers in *International Business Management* scientific articles, precisely in Methods and Results section, while in the Discussion section they appear only in 8% - 9% cases. The reason for such discrepancies is probably to be found in genre diversity. It is obvious that noun premodification is much more common in technical English, tending to present information in the compressed rather than analytical form, without losing clarity (e.g. *ignition started by means of sparks → spark ignition*).

As opposed to the other works which predominantly study two-noun lexical units (noun + noun sequences), our paper presents distribution of noun premodifiers according to the number of nouns in premodification, establishing that 141 (57.79%) most frequent nouns in the corpus (*f* ≥ *4*) are premodified by the total of 368 noun premodifiers consisting of various number of nouns. Depending on the number of nouns in premodification, together with the head noun these premodifiers form two-noun lexical units[2], three-noun lexical units, four-noun lexical units and five-noun lexical units. Their distribution in the corpus in shown in Figure 10.

[2] Some of the two-noun lexical units appearing in our list may be subsumed under compounds in English grammar (e.g. *traffic lights, power plant*), but in this paper, they are treated as two-noun lexical units (MNLUs).

The findings indicate that 298 nouns are premodified by one noun premodifier (80.98%), constituting 2-noun lexical units, 59 (16.03%) have two premodifiers constituting 3-noun lexical units, 8 (2.17%) have three premodifiers constituting 4-noun lexical units, and 3 (0.82%) have four premodifiers constituting 5-noun lexical units. As expected, the more nouns in premodification, the lower percentage of their occurrence.

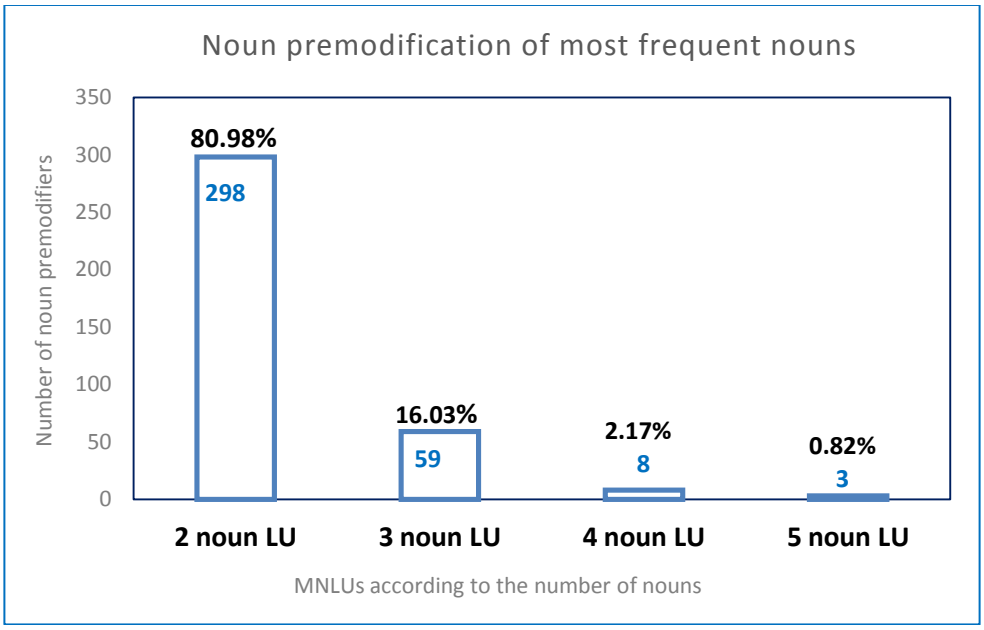

FIGURE 10: Noun premodification of the most frequent nouns.

To illustrate the way of systematizing MNLUs to be used by teachers in order to use them in the classroom, in Table 5 we present the excerpt from the list of noun premodifiers, presenting all the examples of three-noun premodification found in the corpus, i.e. the list of four-noun lexical units.

3 noun premodifiers **(four-noun lexical units)**	1) wheel torque vectoring **system** 2) mass transit railway **system** 3) traffic flow monitoring **system** 4) traffic sings operation **system** 5) freight road transport **vehicle** **6)** goods road motor **vehicle** **7)** working fluid flow **circuit** 8) rush hour traffic **congestion**

TABLE 5: List of four-noun lexical units (head noun + three noun premodifiers).

5. CONCLUSION

Developing lexical competencies in ESP, primarily by learning the most frequent nouns as the most common type of lexical (content) words, and the competencies of receiving/understanding and producing MWLUs and MNLUs is one of the basic tasks of ESP teaching, since good command of the most frequent nouns and mastering of the afore-mentioned condensed structures contributes to developing specialized texts reading and writing skills. These structures are used far more frequently in ESP than in general English and typical of texts that students encounter at the tertiary education level. Yet, their teaching is often unfairly neglected in ESP syllabi, although their usage contributes to the impression of a competent English speaker, which is one of the goals that students strive for.

ESP teachers at institutions of higher education are often in doubt as to which features of the language and which vocabulary to pay more attention to, in order to meet the needs of their students. This paper offers methods of analysing the pedagogical specialized corpus compiled of teaching materials (texts) used in ESP teaching which can provide objective data that will facilitate the selection of nouns, MWLUs and MNLUs which should be taught and learned, without the need for making intuitive choices, but relying on objective and precise corpus-based data instead. Since there is no perfect software to provide completely reliable data, the teacher/researcher must engage into further manual (hand-and-eye) analysis. This paper leads them through the procedure step by step. The analysis allows insight into the production of the frequency list (word list) in rank order, the production of the accurate list of the most frequent nouns, the production of the list of premodifiers of the most frequent nouns (resulting in MWLUs) and finally the production of the list of noun premodifiers of the most frequent nouns (resulting in MNLUs). The paper answers the research questions proving that no computer software can extract all the nouns, MWLUs and MNLUs with 100% certainty without the human aid. It provides their accurate number in the specialized pedagogical corpus of English for Traffic and Transport Purposes (ETTP), comparing the results with the previous works, where possible.

The offered original methodology can be adopted by teachers and applied to the corresponding ESP they teach. The use of data obtained from the analysis of a pedagogical specialized corpus will significantly improve the quality of teaching and, more importantly, the student output competencies. Furthermore, the obtained data can be used by ESP teachers to create glossaries and specialized minimum dictionaries, as a basis for creating syllabi and exercises, designing formative and summative assessment language tests - everything with a scope of enhancing students' lexical competencies in a specific field of study.

Unlike the other works which predominantly study just two-noun lexical units (noun + noun sequences), our paper presents distribution of noun premodifiers (both in numbers and percentages) according to the number of nouns in premodification, which to our knowledge, has never been done before, particularly not for ETTP.

6. REFERENCES

[1] S. Kereković. “Višerječni nazivi u tehničkome engleskom jeziku i njihove prijevodne istovrijednice u hrvatskome jeziku [Multi-word Lexical Units in Technical English and their Equivalents in Croatian].” Doctoral dissertation, Faculty of Humanities and Social Sciences University of Zagreb, Zagreb, 2012.

[2] I. Špiranec. “Višečlani nazivi u engleskom gradjevinskom nazivlju s posebnim osvrtom na imenske složenice [Multi-word lexical units in English for Civil Engineers with Special Emphasis on Nominal Compounds].” Doctoral dissertation, Faculty of Humanities and Social Sciences University of Zagreb, Zagreb, 2011.

[3] T. Polić. “Predmodifikacija imenica u višerječnim nazivima u nastavi engleskoga jezika prometnih struka [Noun premodification in multi-word lexical units in teaching English for traffic and transport purposes].” Doctoral dissertation, Faculty of Humanities and Social Sciences University of Zagreb, Zagreb, 2019.

[4] T. Polić. “Teaching Multi-Word Lexical Units in English for Specific Purposes.” International Journal on Studies in English Language and Literature, Vol. 8, pp. 26-37, Jun. 2020.

[5] M. Gačić. Gramatika engleskoga jezika struke. Zagreb: Učiteljski fakultet Sveučilišta u Zagrebu i Školska knjiga d.o.o., 2009a

[6] J. C. Sager, D. Dungworth and P. F. McDonald. English special languages: principles and practice in science and technology. Amsterdam: John Benjamins Publishing Company, 1980.

[7] O. Chirobocea. "The good and the bad of the corpus-based approach (or data-driven learning) to ESP teaching." Mircea cel Batran Naval Academy Scientific Bulletin, Vol. 20. No. 1, pp. 364-371, Jun. 2017.

[8] L. Engwall, E. Aljets, T. Hedmo and R. Ramuz. "Computer Corpus Linguistics: An Innovation in the Humanities." in Organizational Transformation and Scientific Change: The Impact of Institutional Restructuring on Universities and Intellectual Innovation (Research in the Sociology of Organizations), Vol. 42. R. Whitley and J. Glaser, Ed. Bingley: Emerald Group Publishing, 2014. pp. 331-365.

[9] S. Alsop, V. King, G. Giaimo and X. Xu. "Uses of Corpus Linguistics in Higher Education Research: An Adjustable Lens." Theory and Method in Higher Education Research, Vol. 6, pp. 21-40, Nov. 2020.

[10] L. Bowker. "Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research." Library Hi Tech, Vol. 36, No. 2, pp. 358-371, Apr. 2018.

[11] T. A. Alhabuobi. (2020, October 1). “Differences in Frequencies between Linking Verbs and Relative Pronouns in Written Language”. International Journal of Computational Linguistics (IJCL), Vol. 11, No. 2, pp. 34-48. Available: https://www.cscjournals.org/manuscript/Journals/IJCL/Volume11/Issue2/IJCL-115.pdf [Jul. 3, 2021].

[12] S. N. M. Olimat. (2019, June 30). “Euphemism in the Qur'an: A Corpus-based Linguistic Approach.” International Journal of Computational Linguistics (IJCL), Vol. 10, No. 2, pp. 16-32. Available: https://www.cscjournals.org/manuscript/Journals/IJCL/Volume10/Issue2/IJCL-97.pdf [Jul. 2, 2021].

[13] M. S. Alrabiah, A. M. Al-Salman, E. Atwell and N. Alhelewh. (2014, June 1). “KSUCCA: A Key to Exploring Arabic Historical Linguistics.” International Journal of Computational Linguistics (IJCL), Vol. 5, No. 2, pp. 27-36. Available: https://www.cscjournals.org/manuscript/Journals/IJCL/Volume5/Issue2/IJCL-58.pdf [Jul. 3, 2021].

[14] R. Huddleston and G. K. Pullum. A Student's Introduction to English Grammar, 6th ed. Cambridge: Cambridge University Press, 2010.

[15] A. J. Thomson and A. V. Martinet. A Practical English Grammar. A Practical English Grammar, 4th ed. Oxford: Oxford University Press, 2004.

[16] E. Walker and S. Elsworth. Grammar Practice for Intermediate Students: With Key, 3rd ed. London: Longman, 2000.

[17] R. Huddleston. English Grammar: An Outline, 13th ed. Cambridge: Cambridge University Press, 2005.

[18] R. Quirk and S. Greenbaum. A University Grammar of English. Pearson India, 2016.

[19] G. Leech and J. Svartvik. A Communicative Grammar of English, 3rd ed. London: Routledge, 2013.

[20] M. A. Halliday, C. M. Matthiessen and C. Matthiessen. An Introduction to Functional Grammar, 3rd ed. Abingdon: Routledge, 2014.

[21] D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan and R. Quirk. Longman Grammar of Spoken and written English. London: Longman, 1999.

[22] P. Newmark. The Application of Case Grammar to Translation. Trier: Linguistic Agency, University of Trier, 1985.

[23] M. L. Carrió Pastor. “English complex noun phrase interpretation by Spanish learners.” Revista espanola de linguistica aplicada (RESLA), No. 21, pp. 27-44, 2008.

[24] I. Ferčec and Y. Liermann-Zeljak. “Nominal compounds in technical English” in The Practice of Foreign Language Teaching: Theories and Applications. A. Akbarov, Ed. Newcastle Upon Tyne: Cambridge Scholars Publishing, 2015, pp. 268-277.

[25] A. Stipetić. Rječnik željezničkog nazivlja. Zagreb: Institut prometa i veza, 1994.

[26] M. Lauer. “Designing statistical language learners: Experiments on noun compounds.” Doctoral dissertation, Macquarie University, Sidney, 1996.

[27] B. Leiva de Izquierdo and D. Bailey. “Complex noun phrases and complex nominals: Some practical considerations.” TESL Reporter, Vol. 31, No. 1, pp. 19-29, 1998.

[28] L. H. Ang, K. H. Tan and M. He. “A Corpus-based Collocational Analysis of Noun Premodification Types in Academic Writing.“ 3L: Language, Linguistics, Literature, Vol. 23, No. 1, pp. 115-131, Jan. 2017.

[29] R. Huddleston and G. K. Pullum. The Cambridge Grammar of the English Language, 5th ed. Cambridge: Cambridge University Press, 2012.

[30] A. Hardie. “*Corpus linguistics”* in The Routledge Handbook of Linguistics. K. Allan, Ed. London: Routledge, 2015, pp. 502-515.

[31] E. Vaughan. “How can teachers use a corpus for their own research” in The Routledge Handbook of Corpus Linguistics, A. O’Keeffe and M. McCarthy, Ed. London: Routledge, 2010, pp. 471-484.

[32] S. Adolphs and M. S. Phoebe. “Corpus linguistics” in The Routledge Handbook of Applied Linguistics. J. Simpson, Ed. London: Routledge, 2011, pp. 597-610.

[33] R. M. T. Trimble and L. Trimble. “The development of EFL materials for occupational English” in English for Specific Purposes: An International Seminar. H. L. B. Moody and J. D. Moore, Ed. Bogota, Columbia: The British Council, 1977, pp. 52-70.

[34] M. Limaye and R. Pompian. “Brevity versus clarity: The comprehensibility of nominal compounds in business and technical prose.” International Journal of Business Communication, Vol. 28, No. 1, pp. 7-21, Jan 1991.

[35] P. Master. “Noun compounds and compressed definitions.” English Teaching Forum, Vol. 41, No. 3, pp. 2-9, 2003.

[36] L. Gavioli and G. Aston. “Enriching reality: language corpora in language pedagogy”. ELT Journal, Vol. 55, Issue 3, pp. 238-246, Jul. 2001.

[37] J. Sinclair. “Corpus and Text. Basic Principles” in Developing Linguistic Corpora: a Guide to Good Practice. Wynne M., Ed. Oxford: Oxbow Books, 2005, pp. 1-16.

[38] *. “Common European Framework of Reference for Languages (CEFRL)”. Internet: https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages [Nov. 21, 2020].

[39] @SEPDEK's. "Superscript circles or "non-breaking spaces." Internet: http://georgepavlides.info/superscript-circles-or-non-breaking-spaces/, Dec. 3, 2013. [Sep. 15, 2020].

[40] M. Scott. "What can corpus software do" in The Routledge Handbook of Corpus Linguistics. A. O'Keeffe and M. McCarthy, Ed. London: Routledge, 2010, pp. 136-51.

[41] M. Jones and P. Durrant. "What can a corpus tell us about vocabulary teaching materials" in The Routledge Handbook of Corpus Linguistics. A. O'Keeffe and M. McCarthy, Ed. London: Routledge, 2010, pp. 387-400.

[42] R. Moon. "What can a corpus tell us about lexis" in The Routledge Handbook of Corpus Linguistics. A. O'Keeffe and M. McCarthy, Ed. London: Routledge, 2010, pp. 197-211.

[43] T. Dudley-Evans and M. J. St John. Developments in English for Specific Purposes: A Multi-disciplinary Approach. Cambridge: Cambridge University Press, 1998.

[44] J. Evison. "What are the basics of analysing a corpus?" in The Routledge Handbook of Corpus Linguistics. A. O'Keeffe and M. McCarthy, Ed. London: Routledge, 2010, pp. 122-135.

[45] M. Gačić. Riječ do riječi (lingvistička istraživanja odnosa engleskoga i hrvatskog jezika na području prava i srodnih disciplina). Zagreb: Učiteljski fakultet Sveučilišta u Zagrebu i Profil, 2009b

[46] A. Štambuk. Jezik struke i spoznaja. Split: Književni krug, 2005.

[47] L. Bartolić. "Nominal compounds in technical English" in English for specific purposes: science and technology. M. Trimble, L. Trimble and K. Drobnic, Ed. Oregon: Oregon State University, 1978, pp. 257-277.

[48] S. Seljan and A. Gašpar. "First Steps in Term and Collocation Extraction from English-Croatian Corpus" in Proceedings of 8th International Conference on Terminology and Artificial Intelligence, Toulouse, France, 2009.

[49] S. Seljan, I. Dunđer and A. Gašpar. "From digitisation process to terminological digital resources." Proceedings of the 36th International Convention MIPRO 2013, Rijeka, Croatia, 2013, pp. 1053-1058.