The Relevance of Online Video Clips for the Online Informal Learning of English for Transport Technology – A Corpus-Based Approach

*Violeta **Jurkovič**, Faculty of Maritime Studies and Transport*
University of Ljubljana, Slovenia, violeta.jurkovic@fpp.uni-lj.si

Original scientific paper
DOI: 10.31902/fll.42.2022.18
UDK 811.111'232:004.738.5

Abstract: As a result of the vast opportunities for the use of English in everyday online life, the field of online informal learning of languages, in particular English, has attracted a new wave of research attention. Importantly for this study, it has been confirmed that online engagement in activities involving the use of English has a positive effect on the development of English language proficiency, and that watching online video clips in English is one of the most frequent activities that students engage in. Nevertheless, the number of corpus-based studies in this field remains scant. In addition, no research study has addressed the relevance of online video clips for the online informal learning of English in the specific field of transport technology. In order to bridge this research gap, the objective of this paper is to adopt a corpus-based approach to explore whether a careful selection of video clips in the specific domain of transport technology has the potential to support or supplement the core course materials. The construction of a specialised corpus of video clips adopted a six-step methodology that allowed the preparation, collection, documentation, construction, analysis, and finally evaluation of the corpus. This corpus was then compared against the corpus of the core course materials in terms of lexical density, terminology, and four-grams. The findings of the study indicate that, if carefully selected, video clips have the potential to support the formal language learning process, and should therefore contribute to the understanding of the potential of online informal learning of English in discipline-specific settings; moreover, the adopted methodology may serve as a guide for informed audio-visual material selection and analysis in any language for specific purposes.

Keywords: online informal learning of English, video clips, English for transport technology, lexical analysis, corpus-based approach

Introduction

As a result of the vast opportunities for the use of English in everyday online life, the field of online informal learning of languages, in particular English, has attracted a new wave of research attention. Importantly for this study, it was confirmed that online engagement in activities involving the use of English has,

among other benefits, a positive effect on the development of English language proficiency (Alm 12; Lee and Dressman 442; "Online Informal Learning of English" 37). In addition, it was shown that watching online video clips in English is one of the most frequent activities that online users engage in (Lai and Zheng 6; Lai et al. 7; Trinder 405; "Online Informal Learning of English" 35), which is why it can be assumed that this is one of the online activities with the greatest potential for informal language development.

Nevertheless, the number of corpus-based studies in the field of online informal learning of English remains scarce although they would be helpful in the analysis of the listening materials and subsequent testing on the most frequent lexis (Sockett and Toffoli 150) and word clusters (Sockett 10) within specific genres (Lin 167). A recent scoping review of studies in online informal learning of English published between 2017 and 2019 showed that the most frequently investigated topics involve the linguistic, cultural, affective, agency, and digital literacy dimensions (Soyoof et al. 10). This corroborates the underuse of the corpus-based approach in the research of online informal learning of English, which can at least in part be attributed to the time consuming process of specialised corpus construction.

Therefore, to contribute to the body of corpus-based studies in the field of online informal learning of English, this paper will adopt a corpus-based approach to explore whether a careful selection of video clips in transport technology has the potential to support the core course materials of Professional English I for students of the first-cycle academic study programme of Transport Technology and Logistics at the Faculty of Maritime Studies and Transport of the University of Ljubljana, Slovenia.

After this Introduction, the paper will define online informal learning of English. Then, it will describe specialised language corpora and their construction, and summarise the available corpus-based studies made within the framework of online informal learning of English. After the presentation of the research question and methodology, it will focus on the results of the corpus analysis in terms of lexical density, terminology, and four-grams of the two specialised corpora constructed for the purpose of this study. The results section will be followed by the discussion and conclusion.

Literature review

Online informal learning of English (Sockett 7) or informal digital learning of English (Lee and Dressman 435–6) refers to the online activities that users of the English language engage in while using English for a range of receptive, productive, and interactive activities in authentic communication in the digital environment (Sockett 7). Unprecedented opportunities for engagement with the English language have been enabled in particular by the onset of Web 2.0

technologies that allow users to create their own content and participate in authentic communicative events instead of being only receivers of online resources. Authenticity means that communication usually takes place without the primary intention of improving one's English language competence which does, however, become a by-product of online engagement. Importantly, Trinder (407) showed that online engagement for personal or professional purposes may involve a degree of deliberation to improve one's foreign language competence. Therefore, within this context, language development follows the main principles of language acquisition in naturalistic environments (Kusyk 92) and takes place in line with the main tenets of the usage-based theory of language acquisition (Tomasello 69–87), which emphasises that the development of language is sensitive to the frequency, recency, and contexts of use of language constructions (Bybee 28).

Language corpora can broadly be divided into reference, spoken, specialised, sample, static, dynamic, comparable, and parallel corpora (Gorjanc 8–10). Each corpus is built with a specific purpose. Specialised corpora are most often built for a particular (terminological analysis) objective within the domain of languages for specific purposes and can be very limited in terms of their scope (Gorjanc 8–9). Usually they are used when researchers would like to find an answer to a specific question within their domain of interest (Arhar Holdt 54; Baker et al. 147). Arhar Holdt (56) defines six fundamental steps for specialised corpus construction. The first is the preparation phase when decisions are made regarding the inventory, topics, and coverage of texts to be included in the corpus. The second step concerns data collection. The least time-consuming method is the collection of texts available on the internet if these are in line with the objectives of corpus analysis and copyright. Data collection is followed by the documentation of text data. Next, the corpus needs to be manually checked for any mistakes that result from data transfer between text formats. In the following step the corpus is analysed in line with the research objectives, the results are interpreted, and synthesised. The final step of specialised corpus construction is corpus evaluation, which covers all aspects of the first five steps.

Within the domain of online informal learning of English, Sockett (1–15) adopted a corpus-based approach to describe four-grams that online informal learners of English would be exposed to when watching television series in the English language. For this purpose, a specialised 500,000 token corpus of the most frequently watched television series was built. The four-grams extracted from this corpus were then compared against the most frequent four-grams in the spoken portion of the British National Corpus. The main findings indicated that the predominant structures among the 50 most frequent four-grams were pronoun-verb structures (8). Next, a significant correlation between the frequency of occurrence of the same four-grams in both corpora was found,

which means that the frequency of occurrence of four-grams in the specialised corpus significantly matched their frequency of occurrence in real-life spoken language (9).

Lin (164–176) built a specialised 7.68 million token corpus of internet television series to investigate the validity of watching internet television for the acquisition of formulaic sequences in a foreign language. More specifically, the research study compared the formulaic sequences used in internet television with those used in everyday speech, and explored which internet television genres showed the highest degree of similarity to everyday speech in terms of formulaic sequence patterns. The main findings indicated a high degree of similarity between the speech patterns in internet television series and real life communication (170–1). The genres that displayed the highest degree of similarity with real-life conversations were factual, drama, and comedy series (171–2).

Jurkovič (2021) adopted a corpus-based approach to investigate the potential effect of online informal learning of English on the development of English language competence within the specific domain of maritime English as a language for specific purposes. Two specialised language corpora were used in this study: a spoken corpus of a medical television series, and a written corpus of a medical manual for seafarers. The two corpora were compared in terms of lexical density, lexical diversity, terminology, and four-grams. The results showed significant differences between the two corpora and indicated that watching a medical television series may have a more significant effect on the development of formulaic sequences used in spoken maritime communication rather than on the development of terminological competence in maritime medical English.

Methodology

Research question and hypotheses

The main research question that this paper aims to answer is: "Does a careful selection of video clips in transport technology have the potential to support the core course materials of Professional English I for students of the first-cycle academic study programme of Transport Technology and Logistics at the Faculty of Maritime Studies and Transport of the University of Ljubljana, Slovenia?" To provide an answer to this research question, the following null hypotheses were formulated:

- $H_0$1: The lexical density does not differ significantly between the "Transport Course Book" corpus and the "Transport Video Clips" corpus.
- $H_0$2: The terminology does not differ significantly between the "Transport Course Book" corpus and the "Transport Video Clips" corpus.
- $H_0$3: The most frequent four-grams do not differ significantly between the "Transport Course Book" corpus and the "Transport Video Clips" corpus.

Setting

At the first-cycle degree level, the Faculty of Maritime Studies and Transport of the University of Ljubljana, Slovenia, offers three professional degree study programmes and one academic degree study programme in Transport Technology and Logistics. One of the mandatory courses in the second year of studies of the latter is the 5-ECTS course of Professional English I. The main objectives of this course include reading, analysing, and writing academic and professional texts, developing the students' general lexical and grammatical knowledge and skills as well as subject-specific terminology. While the logistics portion of the course is covered through a researched essay approach, the transport technology portion is catered for by the course book *Course Book Title*. The difficulty level of the course book is set at B2 according to the Common European Framework of Reference for Languages (Council of Europe 24) while the general English language proficiency entry level of the students ranges between A2 and C1.

Specialised language corpora

Two specialised language corpora were used in this study. The first is the corpus of texts in the course book *New Insights into Transport English*. For easier reference, this corpus was named "Transport Course Book".

The course book is divided into four core sections that reflect the main modes of transport: road, rail, maritime, and air transport, each further subdivided into subtopics:

- Road transport (Road traffic safety, Highway Code, Road accident statistics, Vision Zero programme, and Road transport vehicles),
- Rail transport (History and future of rail transport, Rail freight transport, Railway terminals, High-speed passenger trains, and Human resources in railway transport),
- Maritime transport (Merchant navy ships, Maritime ports, Bill of Lading, and Marine pollution), and
- Air transport (Airports, Passenger terminals, and Airplanes).

The course book includes three additional sections: Advantages and disadvantages of transport modes, Intermodal transport, and Transport and urban development.

The "Transport Course Book" corpus was created by deleting the instructions to language tasks and thus retaining the reading, listening, and task texts. The number of tokens (running words) and types (different words) in the "Transport Course Book" corpus is presented in Table 1.

Table 1: Number of tokens and types in the "Transport Course Book" corpus

	"Transport Course Book" corpus
Number of tokens	15,235
Number of types	2,922

The second specialised corpus used in this study is the corpus of transcripts of a selection of video clips in transport technology. A comparison against the "Transport Course Book" corpus will allow us to test the formulated null hypotheses and provide an answer to the research question. For easier reference, this corpus was named "Transport Video Clips". Its construction followed the six steps of specialised corpus construction suggested by Arhar Holdt (56).

First, decisions were made concerning the inventory, topics, and coverage of video clips to be included in the corpus. Using the YouTube search box and filters, three video clips were identified for each of the subtopics covered by the course book based on the following selection criteria:

- the video clip is freely available on YouTube,
- the video clip examines relevant and up-to-date carrier content,
- the length of the video clip ranges between four and twenty minutes,
- the video clip is narrated in standard and correct English,
- an animation or slide presentation is not considered to be a video clip, and
- the video clip is distinguished by high production quality.

Data collection took place so that automatically generated captions in English for each of the three video clips were downloaded, checked, segmented into sentences, and corrected for mistakes.

Next, to facilitate the decision which among the three video clips for each subtopic to include in the corpus, word profiling software (*AntWordProfiler*) was used as a tool to conduct a preliminary analysis of the academic/off-list-type-per-token ratio of each video clip. *AntWordProfiler* is able to distribute the corpus tokens into Nation's General Service Lists 1000 and 2000, and Coxhead's Academic Word List. The off-list tokens were considered to be terminological units. Terminological density, defined as the share of academic and off-list tokens per minute of video clip, was calculated for each video clip. As an example, Table 2 presents the data for the video clip titled "How container ports work: logistics of intermodal transport"[1] .

[1] https://www.youtube.com/watch?v=2JcHMhtH6_s

Table 2: Data for "How container ports work: logistics of intermodal transport"

	Number	**Share (%)**
Tokens	972	100
General Service List 1000 tokens	737	76
General Service List 2000 tokens	56	6
Academic Word List tokens	73	7
Off-list tokens	106	11
AWL + off-list tokens	179	18
AWL + off-list token-per-minute ratio	28.8	

The data presented in Table 2 were used to illustrate the first step of the selection process in which the academic/off-list-token-per-minute ratio of the selected video clips was calculated. The transcript of the sample video clip contains a total of 972 tokens: 737 of these belong to the General Service List 1000 (e.g., contain, efficient) and 56 to the General Service List 2000 (e.g., yard, calculate). There are 73 Academic Word List tokens in this video clip (e.g., mode, implement) while 106 tokens do not appear in any of the applied word lists (e.g., containerisation, intermodal). It was assumed that, with the exception of proper names (e.g., Rotterdam), the latter most likely represented terminological units specific of the presented domain. This video clip is six minutes and thirteen seconds long, which means that the calculated academic/off-list-token-per-minute ratio of 28.8 indicates that almost 29 Academic Word List or off-list tokens are used per video clip minute.

Second, the video clip with the highest academic/off-list-token-per-minute ratio was chosen among the three identified in the preliminary step. After a careful repeated verification of the compliance of each selected video clip with the pre-set selection criteria, the video clip was replaced by another if it included a high number of proper names that had contributed to the academic/off-list-token-per-minute ratio, or if the relevance of the carrier content was found to be low. The specialised "Transport Video Clips" corpus was then built of the transcripts of fifteen video clips, the documented data for which are presented in Table 3 (the links to the selected video clips are available in the footnotes).

Table 3: Transport technology course book subtopic, video clip title and length, and academic/off-list-token-per-minute ratio of the video clips included in the "Transport Video Clips" corpus

Course book subtopic	**Video clip title**	**Length (in seconds)**	**Academic/ off-list-type-per-token ratio per minute**
Advantages and disadvantages of transport modes	N/A		
Intermodal transport	Demonstrating intermodal containerised transport in North-West Europe[2]	1195	36.8
Road transport safety	How modern road barriers keep our roads safe[3]	286	29.0
Highway Code	Understanding road markings[4]	561	13.4
Road accident statistics	N/A		
Vision Zero programme	Systematic safety: the principles behind vision zero[5]	480	24.2
Road transport vehicles	How long-haul trucking works[6]	751	29.2
History and future of rail transport	How freight trains connect the world[7]	621	31.0
Rail freight transport	Boosting efficiency - optimising rail freight operations[8]	376	32.7
Railway terminals	N/A		

[2] https://www.youtube.com/watch?v=y5ouL8ODpIE
[3] https://www.youtube.com/watch?v=XVh_hG4Jkfw
[4] https://www.youtube.com/watch?v=84NkaNkZOrI
[5] https://www.youtube.com/watch?v=5aNtsWvNYKE
[6] https://www.youtube.com/watch?v=QlPrAKtegFQ
[7] https://www.youtube.com/watch?v=9poImReDFeY
[8] https://www.youtube.com/watch?v=gWWRuCr5hfg

High-speed passenger trains	Top 10 fastest passenger trains in the world[9]	363	42.7
Human resources in railway transport	N/A		
Merchant navy ships	Why are billions of dollars-worth of ships being intentionally destroyed[10]	824	25.2
Maritime ports	How Antwerp is a model for smart shipping technology[11]	264	66.6
Bill of Lading	N/A		
Marine pollution	Why 99% of ocean plastic pollution is missing[12]	539	27.6
Airports	Airport layout & structure: aviation lesson 09[13]		40.5
Passenger terminals	Take a tour of the new LaGuardia Airport Terminal B[14]	322	33.5
Airplanes	How airplanes are made[15]	313	31.6
Transport and urban development	AMSURB1x 2016 2.4c Urban growth and urban transport[16]	418	23.1
Total length (in seconds)		**7,313**	
Average academic/off-list type-per-token ratio			**32.5**

The data presented in Table 3 show that relevant video clips that would meet the pre-set selection criteria were not found for the following course book subtopics: Road accident statistics, Railway terminals, Human resources in railway transport, Bill of Lading, and Advantages and disadvantages of transport

[9] https://www.youtube.com/watch?v=kxRbej_8hng
[10] https://www.youtube.com/watch?v=qo-2gDg-37w
[11] https://www.youtube.com/watch?v=JTbUQYINNEU
[12] https://www.youtube.com/watch?v=fsjvwQclGLo
[13] https://www.youtube.com/watch?v=t5SJ37z8UHA
[14] https://www.youtube.com/watch?v=RK3Ndg8F7fM
[15] https://www.youtube.com/watch?v=7rMgpExA4kM
[16] https://www.youtube.com/watch?v=7mns21BrEH0

modes. The total viewing length of the "Transport Video Clips" corpus is 7,313 seconds or, in other words, two hours, one minute, and 53 seconds. The average academic/off-list token-per-minute ratio is 32.5.

The number of tokens and types in the "Transport Video Clips" corpus is presented in Table 4.

Table 4: Number of tokens and types in the "Transport Video Clips" corpus

	"Transport Video Clips" corpus
Number of tokens	20,644
Number of types	3,505

Third, the "Transport Video Clips" corpus was analysed by comparing it against the "Transport Course Book" corpus in terms of lexical density, terminology, and four-grams. The results will be presented in the next section by examining separately each null hypothesis formulated for this study. Each hypothesis will be introduced by the relevant theoretical background.

Results

$H_0$1: The lexical density does not differ significantly between the "Transport Course Book" corpus and the "Transport Video Clips" corpus.

Lexical density is defined by the proportion of content words (nouns, adjectives, main verbs, and adverbs) in a text (Baker et al. 106). One of the most common ways for the calculation of lexical density, which will also be adopted in this study, is to divide the number of content words in a text by the total number of tokens and then multiply the calculated value by 100 (Stubbs 41). The similarity of the text to a spoken text means that it will most likely have a lower lexical density (Halliday 654–6). Lexical density not only depends on the mode of text production but also on the text genre (Stubbs 224).

In order to compare the lexical density of the two specialised corpora, they were first tagged using tagging software (*TagAnt*) that uses the Treebank tags (Santorini 1–37). Concordancing software (*AntConc*) was then used to count all content word tags in each specialised corpus. Finally, following Stubbs' (41) method, the number of content words was divided by the total number of tokens in each corpus and multiplied by 100 (Table 5).

Table 5: Lexical density in the "Transport Course Book" and "Transport Video Clips" corpora

	"Transport Course Book" corpus	**"Transport Video Clips" corpus**
Number of tokens	15,235	20,644
Number of content words	8,719	11,569
Lexical density	57	56

Table 5 shows a high similarity of lexical density values for both corpora, namely 57 percent for the "Transport Course Book" corpus and 56 percent for the "Transport Video Clips" corpus. This means that 57% of all tokens in the first corpus are nouns, adjectives, main verbs, and adverbs, while the share of these word classes in the second corpus is 56%.

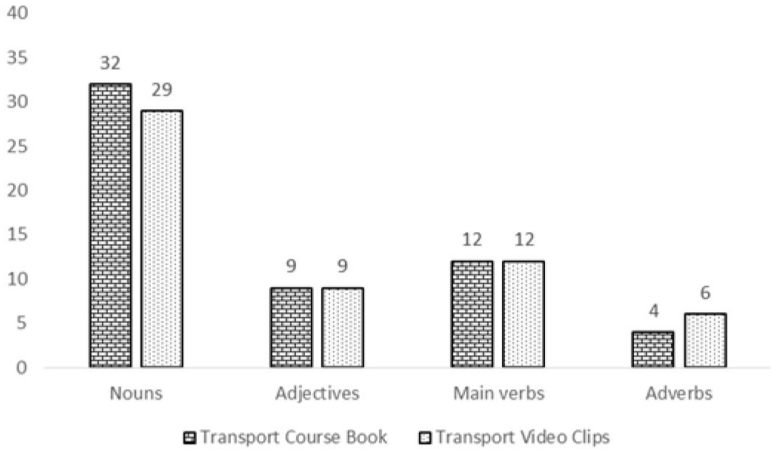
Next, the numbers and shares of individual content word class tokens in relation to the total number of tokens in each corpus were calculated (Table 6 and Picture 1).

Table 6: Numbers and shares of content word class tokens in the "Transport Course Book" and "Transport Video Clips" corpora

	"Transport Course Book" corpus		**"Transport Video Clips" corpus**	
	Number	**Share**	**Number**	**Share**
Nouns (total)	**4880**	**32**	**6078**	**29**
NN – noun, singular or mass	2978	20	3747	18
NNS – noun plural	1351	9	1728	8
NP – proper noun, singular	533	3	598	3
NPS – proper noun, plural	18	0	5	0
Adjectives (total)	**1346**	**9**	**1765**	**9**
JJ - adjective	1232	8	1560	8
JJR – adjective, comparative	91	1	139	1
JJS – adjective, superlative	23	0	66	0
Main verbs (total)	**1850**	**12**	**2487**	**12**
MD – modal	227	1	314	2
VV – verb, base form	441	3	654	3
VVD – verb, past tense	78	1	126	1
VVG – verb, gerund/present participle	313	2	424	2
VVN – verb, past participle	479	3	477	2

VVP – verb, singular present, non-3rd	126	1	234	1
VVZ – verb, 3rd person singular present	186	1	258	1
Adverbs (total)	**643**	**4**	**1239**	**6**
RB – adverb	521	3	1041	5
RBR – adverb, comparative	25	0	51	0
RBS – adverb, superlative	8	0	15	0
WRB – wh-adverb	89	1	132	1
Content words (total)	**8,719**	**57**	**11,569**	**56**

The highest shares among content word class tokens in both corpora belong to nouns (32% in the "Transport Course Book" corpus and 29% in the "Transport Video Clips" corpus), followed by main verbs (12% in both corpora), adjectives (9% in both corpora), and finally adverbs (4% in the "Transport Course Book" corpus and 6% in the "Transport Video Clips" corpus). Picture 1, which summarises the data presented in Table 6, visually corroborates the high level of similarity in the shares of content word class tokens in both analysed corpora.



Picture 1: Shares of content words class tokens in the "Transport Course Book" and "Transport Video Clips" corpora

The level of similarity was then verified through the adoption of the statistical significance test, which corpus studies most often employ for the extraction of key words or, in our case, key tags. The statistically significant differences are determined by comparing the frequency of occurrence of a word type (or tag) in one corpus with the frequency of occurrence of the same word type (or tag) in another (McEnery and Hardie 51). Using *AntConc*, the tagged "Transport Video Clips" corpus was compared with the tagged "Transport Course Book" corpus in order to extract the positive key tags for content word classes that appear statistically significantly more often in the "Transport Video Clips" corpus, and negative key tags for content word classes that appear statistically significantly more often in the "Transport Course Book" corpus. At the level of statistical significance of $p<0.0001$, set for this study, no statistically significant differences were found between the two specialised corpora in terms of the frequency of occurrence of content word class tags. This means that there are not any statistically significant differences in the frequency of occurrence of content word class subcategories (e.g., singular or mass nouns, adjectives in the comparative form, verbs in the past tense, or wh-adverbs) in the two specialised corpora.

$H_0$2: The terminology does not differ significantly between the "Transport Course Book" corpus and the "Transport Video Clips" corpus.

As mentioned in the previous section, key words may be extracted by comparing their frequency of occurrence in two corpora. Therefore, a key word can be defined as "a word which appears in a text or corpus statistically significantly more frequently than would be expected by chance when compared to a corpus which is larger or of equal size" (Baker et al. 98). If the two analysed corpora are lexically similar, a low number of (positive or negative) key words will be extracted. If words appear with a similar frequency, it may safely be assumed that the two corpora are lexically related (John et al. 7).

The log-likelihood test integrated in *AntConc* was used to derive the positive key words from the two corpora. First the "Transport Video Clips" corpus was used as the reference corpus to extract the positive key words from the "Transport Course Book" corpus. Then the same procedure was repeated in the opposite direction. In other words, in the second step the "Transport Course Book" corpus was used as the reference corpus to extract the positive key words from the "Transport Video Clips" corpus.

Table 7 shows the 19 positive key words by keyness for the "Transport Course Book" corpus.

Table 7: Key words for the "Transport Course Book" corpus

Frequency	**Keyness**	**Key word**	**Example extract from the "Transport Course Book" corpus**
32	42.48	development	… on the basis of regional *development* benefits which …
20	34.69	deck	Below the weather *deck* are the cargo holds.
33	30.30	vehicle	Single road *vehicle* accident
16	27.75	motorway	… check the traffic on the *motorway* and …
37	26.43	goods	When may the *goods* be discharged?
104	25.29	transport	… the economic performance of a *transport* chain …
31	25.09	maximum	… with a *maximum* laden weight exceeding 3.5 tonnes …
21	25.01	required	*Required* qualifications
17	22.85	yards	Shunting *yards* are considered as first generation yards.
13	22.54	ballast	… ships must either exchange or treat their *ballast* water.
13	22.54	locations	Accessible *locations* and interchanges that allow …
13	22.54	substantial	… has a *substantial* impact on the economy.
44	21.24	cargo	Where can containerised *cargo* be carried?
16	21.23	EU	Even once the *EU* has its corridors in operation, …
12	20.81	cranes	The *cranes* move standardised loading units …
137	20.33	be	Otherwise, there will always *be* a free rider problem.
15	19.62	private	… is the biggest European *private* rail operator.
11	19.07	qualifications	Required *qualifications*
19	18.70	benefits	But if there are substantial *benefits* to companies …

AntWordProfiler was used to check to which word list the extracted key words belong. Five were found to belong to the General Service List 1000 (be, development, or, private, substance), two to the General Service List 2000 (qualify, yard), six to the Academic Word List (benefits, locations, maximum, required, transport, vehicle) while nine were off-list words (cranes, EU, cargo, ballast, yards, motorway, goods, deck, qualifications).

Table 8 shows the 13 positive key words by keyness for the "Transport Video Clips" corpus.

Table 8: Key words for the "Transport Video Clips" corpus

Frequency	**Keyness**	**Key word**	**Example extract from the "Transport Video Clips" corpus**
118	53.75	we	… *we* will reduce CO2 emissions by …
108	39.69	they	… *they* can help us drive safely.
42	38.07	plastic	… tons of our *plastic* debris has accumulated …
180	30.58	it	… *it* stays fairy near the shore.
27	29.46	airports	*Airports* are divided into landside and airside areas.
23	25.09	runway	A *runway* is a defined rectangular area …
29	24.60	ocean	… will end up in the middle of the *ocean*.
47	24.25	world	… the list of top 10 fastest trains in the *world*.
28	23.58	miles	There are nearly 400,000 *miles* of coastline …
35	22.40	just	One can't *just* drive into a city and park …
56	20.33	train	If a *train* crew reaches twelve hours …
18	19.64	barrier	The spiked *barrier* is simply a steel framework …
55	19.60	per	… is designed for 217 miles *per* hour.

AntWordProfiler was used again to check to which word list the key words presented in Table 8 belong. Eight were found to belong to the General Service List 1000 (it, just, miles, per, they, train, we, world), one to the General Service List 2000 (ocean), none to the Academic Word List while four were off-list words (plastic, airports, runway, barrier).

Given that the key word analysis revealed a low number of key words for each of the two specialised corpora used in this study, the most frequent content words were extracted from each to further examine their lexical similarity.

Table 9 presents thirty most frequently occurring content words in the "Transport Course Book" and "Transport Video Clips" corpora. An asterisk was used to label the content words that appear among the most frequent thirty content words in both corpora, including different lemmas of each content word (e.g., ship/ships).

Table 9: Most frequent content words in the "Transport Course Book" and "Transport Video Clips" corpora

"Transport Course Book" corpus			**"Transport Video Clips" corpus**		
Content word	**Frequency**	**Share (%)**	**Content word**	**Frequency**	**Share (%)**
transport*	103	0.7	more	76	0.4
road*	69	0.5	road*	69	0.3
rail*	56	0.4	transport*	66	0.3
traffic*	46	0.3	train	55	0.3
cargo	44	0.3	traffic*	54	0.3
containers*	42	0.3	freight*	53	0.3
speed*	42	0.3	port	52	0.3
new	37	0.2	trains	47	0.2
goods	36	0.2	speed*	45	0.2
freight*	33	0.2	aircraft	44	0.2
vehicle	33	0.2	container*	42	0.2
vehicles	33	0.2	plastic	42	0.2
development	31	0.2	ships*	39	0.2
maximum	31	0.2	time	39	0.2
may	31	0.2	would	39	0.2
should	31	0.2	use*	38	0.2
lane	27	0.2	terminal	36	0.2
ship*	27	0.2	truck	36	0.2
weight	27	0.2	world	35	0.2
modes	26	0.2	trucks	33	0.2
use*	25	0.2	around	32	0.2
only	24	0.2	most	32	0.2
container*	23	0.2	now	31	0.2
high*	23	0.2	rail*	31	0.2
infrastructure	23	0.2	shipping	31	0.2

roads	23	0.2	system	30	0.1
safety	23	0.2	high*	29	0.1
costs	22	0.1	make	29	0.1
means	22	0.1	take	29	0.1
must	22	0.1	containers*	28	0.1

The data presented in Table 9 indicate that eleven content words appear among the thirty most frequent content words in both corpora. These are: transport, road, rail, traffic, container/s, speed, freight, ship/s, use, and high. On the other hand, nineteen most frequent content words of the first thirty are specific of each specialised corpus employed in this study.

$H_0$3: The most frequent four-grams do not differ significantly between the "Transport Course Book" corpus and the "Transport Video Clips" corpus.

Word clusters (n-grams, lexical bundles, or lexical chunks), in this paper four-grams, are simply defined as groups of words that appear in a sequence. This means that cluster analysis can be used to assess the degree of similarity between texts or corpora (Baker et al. 34) because it implies that similar word clusters will appear in two corpora with a similar frequency. Hyland (42) suggests that one of the signals that a language user is a competent member of a particular discourse community is their competent use of word clusters.

In order to enable a comparison with previous studies (Sockett 1–15; Lin 164¬–176), four-grams with a frequency of five or more were extracted from both corpora using *AntConc*. The results are presented in Table 10.

Table 10: Four-grams with a frequency of occurrence of 5 or more in the "Transport Course Book" and "Transport Video Clips" corpora

"Transport Course Book"		**"Transport Video Clips"**	
Four-gram	**Frequency**	**Four-gram**	**Frequency**
new transport infrastructure has	6	the port of Zebrugge	6
the left hand lane	6	centre of the road	5
and the development potential	5	of the world's	5
as a result of	5	the centre of the	5
economy and the development	5	the port of Dublin	5
qualifications high school diploma	5		

A limited number of four-grams appear in both specialised corpora at least five times, as the data presented in Table 10 indicate. While the four-grams from the "Transport Course Book" corpus display different structures (e.g., the formulaic sequence *as a result of*), those in the "Transport Video Clips" corpus all contain of-phrases used to clarify the nominal headword (e.g., centre of the road).

Discussion and conclusion

The research question that this paper addressed is whether a careful selection of video clips in transport technology has the potential to support the core course materials covered during the classes of Professional English I for students of the first-cycle academic study programme of Transport Technology and Logistics at the Faculty of Maritime Studies and Transport of the University of Ljubljana, Slovenia.

In the discussion section, the results will be synthesised and the corpus construction and analysis process evaluated, thus completing the six phases of specialised corpus construction defined by Arhar Holdt (56). In addition, the results will be integrated with findings of previous studies, and implications for pedagogy will be drawn.

First, the results showed a similar lexical density of both corpora. Although they may be different in terms of their mode of production (a spoken corpus of video clips compared against a written corpus of the course book), the high lexical density of the video clips corpus may depend on this specific genre (Stubbs 224), which emulates the structure of traditional expository documentaries in which a written text is narrated. This specific nature of the video clip narrative was corroborated by the analysis of content words classes in both corpora, which yielded no statistically significant differences in terms of their frequency. Therefore, the first null hypothesis can be accepted. It was shown that the lexical density does not differ significantly between the "Transport Course Book" and the "Transport Video Clips" corpora.

Second, the relatively low number of key words extracted for both specialised corpora when compared against each other indicated their lexical relatedness (John et al. 7). Among the twenty positive key words found for the "Transport Course Book" corpus, as many as fifteen belong to the Academic Word List or were identified as off-list discipline-specific words. On the other hand, most key words extracted from the "Transport Video Clips" corpus belong to the General Service List 1000, which indicates a lower frequency of this list words in the "Transport Course Book" corpus. In addition, the lexical relatedness of the two corpora was corroborated by the analysis of the most frequent content words. As a result, the second null hypothesis can be accepted. It was shown that the terminology does not differ significantly between the "Transport Course Book" and the "Transport Video Clips" corpora.

Finally, the analysis of the most frequent word clusters was made. The results of the analysis of the "Transport Video Clips" corpus were different from those found by Sockett (7–9), Lin (174–5), and Jurkovič (2021) that analysed the language that online users of English are exposed to when watching online television series. This corroborated the importance of the consideration of genre in corpus analysis (Stubbs 224; Lin 173) and the existence of significant differences between the genres of online video clips on the one hand and television series on the other. However, for both corpora used in this study a low number of four-grams found should at least in part be attributed to limited corpora sizes. Nevertheless, based on the structure of the extracted four-grams, the third null hypothesis can be rejected and the alternative hypothesis accepted. The most frequent word clusters differ significantly between the "Transport Course Book" and the "Transport Video Clips" corpora.

After the synthesis of the results, both specialised corpora built for the purpose of this study need to be evaluated to complete the six phases of corpus construction proposed by Arhar Holdt (56). First of all, the results might have been affected by the relatively small size of both corpora. This was made evident in particular by the analysis of four-grams. Second, the small size of both corpora evidenced the problem of the occurrence of very high frequency types in a particular corpus segment. For example, the type "qualifications" was extracted as a key word for the "Transport Course Book" corpus. However, a verification of this corpus indicated that all eleven occurrences of this type could be observed in a very limited corpus segment.

Based on the lexical density, terminology, and word cluster analysis of the two specialised corpora built for the needs of this study, it can be safely claimed that this selection of video clips in transport technology has the potential to partly support and partly supplement the core course materials of Professional English I for students of the first-cycle academic study programme of Transport Technology and Logistics at the Faculty of Maritime Studies and Transport of the University of Ljubljana, Slovenia.

If encouraged to watch these video clips, the links to which can be made available to them via the Virtual Learning Management system, students will be provided with another relevant resource for the consolidation of their language knowledge and skills developed during their professional English class. As mentioned in the literature review section, language constructions that are developed though communicative usage events are sensitive to frequency, recency, and contexts of occurrence (Bybee 28). Therefore, the limited frequency of occurrence of a particular type in one resource that the students are exposed to (e.g., course book) may be compensated by the repeated occurrence of the same type in another resource (e.g., video clip), which should contribute to the consolidation of learning, also by providing a reminder of a recent lexical item within a different yet similar context of use.

Finally, the present study stemmed from the needs of a single teacher/researcher within a unique teaching and learning context, which may limit the generalisability of the results. Nevertheless, the adopted methodology of specialised corpus construction may serve as a guide for informed audio-visual materials selection and analysis ("Relevance of Online Video Clips" 295) in any language for specific purposes in other teaching and learning contexts.

WORKS CITED

Alm, Antonie. "Facebook for Informal Language Learning: Perspectives from Tertiary Language Students." *The EUROCALL Review* 23.2 (2015): 3–18. Print.

Anthony, Laurence. *TagAnt (Version 1.2.0)*. Tokio: Waseda University, 2015. Web. 8 July 2022.

Anthony, Laurence. *AntConc (Version 3.5.9)*. Tokio: Waseda University, 2020. Web. 8 July 2022.

Anthony, Laurence. *AntWordProfiler* (Version 1.5.1). Tokio: Waseda University, 2021. Web. 8 July 2022.

Arhar Holdt, Špela. "Gradnja specializiranega korpusa." *Jezik in slovstvo* 51.1 (2006): 53–67. Print.

Baker, Paul, Andrew Hardie, and Tony McEnery. *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press, 2006. Print.

Bybee, Joan. *Language, Usage, and Cognition*. Cambridge: Cambridge University Press, 2010. Print.

Council of Europe. *Common European Framework of Reference for Languages*. Strasbourg: Council of Europe, 2001. Print.

Coxhead, Averil. *An Academic Word List*. Wellington: Victoria University of Wellington, 1998. Print.

Gorjanc, Vojko. *Uvod v korpusno jezikoslovje*. Domžale: Založba Izolit, d.o.o., 2005. Print.

Halliday, Michael. *An Introduction to Functional Grammar*. London: Arnold, 2004. Print.

Hyland, Ken. "Academic Clusters: Text Patterning in Published and Postgraduate Writing." *International Journal of Applied Linguistics* 18.1 (2008): 41–62. Print.

John, Peter, Benjamin Brooks, and Ulf Schriever. "Profiling Maritime Communication by Non-native Speakers: A Quantitative Comparison between the Baseline and Standard Marine Communication Phraseology." *English for Specific Purposes 47* (2017): 1–14. Print.

Jurkovič, Violeta. "Relevance of Online Video Clips for Autonomous Learning of Maritime English." *Languages for Special Purposes in a Multilingual, Transcultural World*. Ed. Gerhard Budin and Vesna Lušicky. Vienna: University of Vienna, 2014. 290–297. Print.

Jurkovič, Violeta. "Online Informal Learning of English through Smartphones in Slovenia." *System* 80 (2019): 27–37. Print.

Jurkovič, Violeta. "Možnost vplivanja priložnostnega učenja angleščine preko spremljanja medicinske televizijske serije na razvoj jezikovne zmožnosti v medicinski pomorski angleščini – korpusni pristop." *Vestnik za tuje jezike* 13.1 (2021): 445–465. Print.

Jurkovič, Violeta. *New Insights into Transport English*. Portorož: Faculty of Maritime Studies and Transport, 2021. Print.

Kusyk, Meryl. "The Development of Complexity, Accuracy and Fluency in L2 Written Production through Informal Participation in Online Activities." *CALICO Journal* 34.1 (2017): 75–96. Print.

Lai, Chun, and Dongping Zheng. "Self-directed Use of Mobile Devices for Language Learning beyond the Classroom." *ReCALL* 30.3 (2017): 299–318. Print.

Lai, Chun, Xiao Hu, and Boning Lyu. "Understanding the Nature of Learners' Out-of-Class Language Learning Experience with Technology." *Computer Assisted Language Learning* 31.1–2 (2017): 114–143. Print.

Lee, Ju Seong, and Mark Dressman. "When IDLE Hands Make an English Workshop: Informal Digital Learning of English and Language Proficiency." *Tesol Quarterly* 52.2 (2018): 435–445. Print.

Lin, Phoebe. "Investigating the Validity of Internet Television as a Resource for Acquiring L2 Formulaic Sequences." *System* 42 (2014): 164¬–176. Print.

McEnery, Tony, and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press, 2011. Print.

Nation, Paul. *Teaching and Learning Vocabulary*. Massachusetts: Newbury House, 1990. Print.

Santorini, Beatrice. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. 3rd Revision. Philadelphia: University of Pennsylvania, 1990. Print.

Sockett, Geoffrey. "From the Cultural Hegemony of English to Online Informal Learning: Cluster Frequency as an Indicator of Relevance in Authentic Documents." *Asp* (2011): 1–15. Print.

Sockett, Geoffrey, and Dayzee Toffoli. "Beyond Learner Autonomy." *ReCALL* 24.2 (2012): 138–151. Print.

Soyoof, Ali, et al. "Informal Digital Learning of English (IDLE): a Scoping Review of what Has Been Done and a Look towards What is to Come." *Computer Assisted Language Learning* (2021): 1–27. Print.

Stubbs, Michael. *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell Publishing, 2002. Print.

Tomasello, Michael. "The Usage-based Theory of Language Acquisition." *The Cambridge Handbook of Child Language*. Ed. Edith Bavin. Cambridge: Cambridge University Press, 2009. 69–88. Print.

Trinder, Ruth. "Informal and Deliberate Learning with New Technologies." *ELT Journal* 71.4 (2017): 401–412. Print.

LA RILEVANZA DEI VIDEO PER L'APPRENDIMENTO INFORMALE DELL'INGLESE IN RETE PER STUDENTI DI TECNOLOGIA DEI TRASPORTI – UN APPROCCIO BASATO SUL CORPUS

Le vaste opportunità per l'uso dell'inglese in rete nella vita di tutti i giornianno dato spunto ad un nuovo campo di ricerca: l'apprendimento informale in rete delle lingue straniere, in particolare dell'inglese. Per la presente ricerca è importante sottolineare che i risultati conseguiti dalle ricerche precedenti hanno confermato l'effetto positivo sulla conoscenza dell'inglese in seguito alle diverse attività in rete. Una delle attività più comuni compiute dagli utenti in rete è guardare video su varie piattaforme social. Nonostante ciò, la ricerca in questo campo si è servita solo raramente dell'approccio basato sul corpus. Rimane inoltre del tutto tralasciato il tema della rilevanza dei video per l'apprendimento informale dell'inglese in rete nel campo specifico della tecnologia dei trasporti. Lo scopo di questo articolo è colmare questa lacuna di ricerca attraverso un approccio basato sul corpus con l'intento di constatare se una selezione rilevante di video nel dominio specifico della tecnologia dei trasporti possa essere da supporto o essere integrata nel corso d'inglese per gli studenti della tecnologia dei trasporti e della logistica. La costruzione di questo corpus specializzato di video è stata divisa in sei fasi: preparazione, raccolta, documentazione, costruzione, analisi e valutazione. Nella fase successiva questo corpus è stato paragonato con il corpus del libro di testo usato durante il corso in termini di densità lessicale, terminologia e n-grammi. I risultati sembrano indicare il potenziale dei video che, se accuratamente selezionati, possono essere di supporto per l'apprendimento formale delle lingue straniere anche in contesti disciplinari specifici. Inoltre, la metodologia adottata per la costruzione di questo corpus specializzato può servire da modello per la selezione e scelta informata di materiale audiovisivo in qualsiasi lingua per scopi specifici.

Parole chiave: apprendimento informale dell'inglese in rete, video, inglese per la tecnologia dei trasporti, analisi lessicale, approccio basato sul corpus.