Vilnius University

Faculty of Philology

Department of English Philology

Gerda Pudžiuvelytė

A Contrastive Analysis of Research Papers and Media Articles on Astrophysics in English: the Lexical Bundles Approach

Thesis submitted in partial fulfilment of requirements for the degree of MA in English Studies

Supervisor: Prof. Assoc. Dr Rita Juknevičienė

Vilnius

2023

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Rita Juknevičienė, for her patience, support, and guidance throughout this study. I am thankful for her given advice and shared knowledge while writing this research paper.

However, the remaining errors are my own.

Table of Contents

Table of Contents

List of Abbreviations

$Corpus_{ACAD}$	Corpus of academic research papers on astrophysics
$Corpus_{MEDIA}$	Corpus of media articles on astrophysics
DC	Dependent clause
DC-based LB	Lexical bundle that incorporates fragments of dependent clause
DO	Discourse-organising function
LB	Lexical bundle
LBA	Lexical bundles approach
NP	Noun phrase
NP-based LB	Lexical bundle that incorporates fragments of noun phrases and/or prepositional phrase
R	Referential function
S	Stance function
VP	Verb phrase
VP-based LB	Lexical bundle that incorporates fragments of verb phrase

Abstract

The study analyses the complexity of linguistic expression of academic and media articles on astrophysics. The main objectives of this research were to identify and examine register differences in terms of the structural and functional features. In accordance with the objectives of the study, two corpora of academic and media articles on astrophysics were compiled. The study was carried out based on the frequency-driven approach and the lexical bundles approach (Biber et al. 2004) with which the recurrent four-word sequences were determined and analysed. The results of the study reveal that the media and academic registers have more differences than similarities both structurally and functionally. The structural variation between the registers in general shows the tendency that: the lexical bundles that incorporate noun phrase and prepositional phrase fragments are more frequent in the media register; the lexical bundles that incorporate verb phrase fragments are as twice more occurrent in the academic register; and, the lexical bundles that incorporate dependent clause fragments are slightly more prominent in the academic register. Regarding the functional variation between the registers, it was identified that the media register contains more referential lexical bundles; whereas, in the academic register stance and discourse organising lexical bundles were approximately two and four times more frequently occurring respectively. Moreover, it was found one predominantly exclusive structure for each analysed register. In academic register it was *1st person + that-clause* structure and in the media register - *noun phrase + at-phrase* structure.

1. Introduction

It is noticeable that fewer studies that apply the lexical bundles approach have examined the media register. Thus, taking into consideration the lack of interest concerning this specific language variety, it would be interesting to investigate this register's features. The purpose of the study is to carry out a contrastive corpus-driven analysis of articles on astrophysics published in academic journals and in media news portals / magazines. Physics in general is considered as a challenging subject. Its sub-branch astrophysics is even more formidable because in the physics the majority of the classical laws (i.e. Newton's laws of motion) can be observed hic et nunc, on the Earth. In contrast, in astrophysics additional mathematics and computer simulations are required to examine an item of the analysis (i.e. the black hole).

Regarding the presumable linguistic complexity of a scientific astrophysics research paper and an online media article it is hypothesized that academic research articles are more complex than media articles in terms of linguistic expression considering the media writer's obligation to produce such text in which the difficult subject would be presented as easily comprehensible information containing less complex vocabulary and grammatical structures. This research paper seeks to investigate how the complexity of linguistic expression, lexical and syntactical features contrast in the academic and media articles. The study is guided by two main research questions, namely: 1) what are the differences or similarities between the two registers in terms of the structural features; 2) what are the differences or similarities between the two registers in terms of the functional features.

1.1. Literature review

1.1.1. Register analysis

In the last three decades, register has become one of the most examined discourse elements in language variation studies. The definition of register has slightly altered throughout the years since its first mention. The first definition of the term was proposed by Thomas Bertram Reid (1956) by which he emphasized that registers are varieties of language which are used in the different social situations. In other words, registers work as prescriptions of certain type of language for specific life scenarios. Crystal (1964: 149) described register similarly to the former definition by noting that it is a kind of language whose form is seen as characteristics of a social situation. And thereby, there can be detectable legal language, liturgical language etc. (ibid.). In addition to this, Ferguson (1983: 154) claimed that the language structure varies

in accordance with the particular occasions of use because every language is situated in the social context.

It can be noticed that the previous older definitions somewhat rely too heavily on a social aspect which can be explained by the fact that only in the 1960s the sociolinguistics emerged and consequently separated from the linguistics as a new branch, thus the connection of registers and social contexts hitherto are apparent. Biber et al. (1999: 8-9) describe register as a particular kind of text. A social context is not even mentioned presumably considering the new interest of linguists that was corpus linguistics (emerged in 1980s). By a particular kind the authors meant the structure and function of a language. Biber et al. (1999) present linguistic evidence for four major registers: conversation, fiction, newspaper and academic prose. Later in the 21st century the notion of social context was specified to situational variables, because register still represents a certain type of discourse. Biber and Conrad (2009: 6) defined register as a variety of language that is associated with a particular situation of use. Similarly, Crawford and Csomay (2016: 14) referred to a register as a variety of language that is defined by both the language used in the context and a specific context.

In the more recent definitions it is conspicuous that the social context became simply a specific context where the language used in that context itself becomes part of a register's formation. This means that the context is perceived as one of the variables alongside such other variables as topic, purpose of communication or mode of communication when analysing a register but not as a creator of a specific type of language.

Another important point must be observed in relation to other similar terms used in linguistic literature, namely, *genre* and *style* which are sometimes used interchangeably with *register.* However, there are several important differences. While genre analysts focus on the conventional structures that construct a text; research on style typically emphasises the aesthetic preference of linguistic features within a variety. Register analysis in light of a situational context unravels functionally motivated lexical and grammatical characteristics that are typical for a specific variety of language (Biber & Conrad 2009: 2).

Register analysis includes three objectives. The first is an analysis of a text within its context; the second, an analysis of linguistic features that are detected in the text; the third, the functional interpretation of the relationship between the language that is produced in the context and the context (Crawford & Csomay 2016: 16). Language can be used for different purposes, in different circumstances and in different contexts (Biber & Conrad 2009: 9). Thus before

commencing the analysis of typical lexical and grammatical features the situational variables of the analysed register have to be indicated. The scholars have provided a list of such variables: participants, relations among participants, channel, production circumstances, setting, communicative purpose and topic (ibid.: 40). For instance, in this current study the registers of academic and media article on astrophysics will be analysed. The major conspicuous difference between the two varieties is the different audiences (see Table 1 below). Thus it can be hypothesized that this might be the main factor which would account for the differences between the two registers. The variable similarities and differences between the registers given in Table 1 exemplify approach proposed by Biber and Conrad (2009):

Table 1. The situational variables of academic and media registers on astrophysics.

	Academic	Media
Participants	Addressor: one or more scholars Addressee: specialist audience	Addressor: one writer Addressee: wide public
Relationship	Shared knowledge	Interaction, shared knowledge
Channel	Written, online/printed	Written, online/printed
Production	Exhaustive planning, edited	Planning, edited
Setting	Public	Public
Communicative Purposes	Inform, explain, persuade	Narrate, inform about what has been discovered, summarize information
Topic	Specific Topic on Astrophysics scientific analysis	General topic on Astrophysics / Astronomical items

After describing the registers, the linguistic analysis is undertaken. It is crucial to note that analysing language variety requires a representative data set (ibid.: 10). A larger sample of texts is better than a smaller one; because of a one reason, it could possibly be not credible to make assumptions or conclusions about a language variety after analysing a small sample of texts. Each language variety has its features that have been observed and analysed by linguists.

For example, the purpose of a scientific research article is to reveal new information about the natural world by identifying new phenomena and comparing it with its previous patterns (Biber & Gray 2016: 6). It works as a tool of thinking that transmits the knowledge to people (Heller & Morek 2015: 175). Academic language is considered to be conspicuously formal, impersonal, precise and economical (Li & Li 2015: 161). In addition to this, academic register can be realized by the means of lexical and syntactical features that usually are the frequent use of technical words, abstract nouns and descriptive adjectives, acronyms, symbols or signs; also, the use of premodification rather than post-modification, nominalisation, non-predicative verbs, passive voice, long sentences (Li & Li 2015: 161-163, Parkinson 2013: 164-165). However, some specific scientific fields might reveal something unique, for instance, Tarone et al. (1981) have conducted a study of analysis of the usage of passive voice in EST (English for science and technology) register. They have examined the occurrence of verb forms of passive and active voices in two astrophysics journal articles (Tarone et al. 1981: 123). The scholars found out that the use of the passive and the structure of *we* plus active voice have a similar frequency of occurrences; they have drawn 4 conclusions which are, namely:

> (1) we indicates the author's unique procedural choice, while the passive indicates an established or standard procedure; (2) we is used to describe the author's own work and the passive to describe the work of others, unless that work is not mentioned in contrast to the author's, in which case the active is used; (3) the passive is used to describe the author's proposed studies; and (4) the use of the active or the passive is determined by focus due to the length of an element or the need for emphasis.
>
> (ibid.).

Presumably media register as such is less frequently analysed than the academic. It is commonly specified as a newspaper, news, online discourse, sports report, blog etc. (Biber et al. 1999: 8-9, Biber & Egbert 2018: 75). Notwithstanding the wide range of sub-registers within the broad and varied category of media they still do have some shared typical characteristics. Biber and Egbert (2018: 105) analysed many sub-registers of an online discourse which they referred to as the Narrative Register. The scholars have found that nearly all the mini registers are focusing on people and heavily relying on the descriptive detail.

In this current study articles produced in news portals and online magazines will be analysed. It is plausible that an online magazine sub-register was not excessively analysed yet, however, the comparison of the situational variables of Biber and Egbert's and this study's shows clear similarities: single author, written, general audience, little specialist knowledge needed (ibid.:

74). Thus it can assumed that the online magazine little register belongs to the narrative register that is an online discourse that refers to the media register.

The last step of register analysis that follows is the functional explanation why certain linguistic items are more typical in one or another register. Registers differ in their distributions of characteristics rather than in a single occurrence of a feature (Biber & Conrad 2009: 9). It does not matter if it is a conversation or newspaper article, a noun or a 1st person pronoun would be found at least once in both of these registers; however, the distributions of these two linguistic items would proportionally be different, conversation would contain more pronouns and fewer nouns and, in the newspaper language, it would be vice versa (ibid.). That is why the registers have typical, the most expected, representative, characteristics which in association with the situational variables depict a language variety.

1.1.2. Register and corpus approach

In register analyses the large-scale sample of texts usually is a corpus, either an online available one (BNC, COCA) or a specialised smaller size. According to Sinclair (1991: 171), "a corpus is a collection of naturally occurring language text, chosen to characterize a state or variety of a language". The crucial benefit of applying corpus approach for a register analysis is the large set of machine-readable texts produced naturally by humans. More specifically, a corpus-driven approach primarily systematically aims to derive linguistic characteristics from the frequency distributions and recurrent patterns that appear in the analysed language (Tognini-Bonelli 2001: 87). The frequency data provides patterns that should be explained (Biber et al. 2004: 376). Furthermore, the frequency-driven approach is particularly useful for the analysis of multi-word sequences that can reflect the reason why those word chunks are prefabricated and used in the first place (ibid.). Thus, a corpus is a reliable resource when analysing a language variety. Firstly, its composition includes a careful selection of texts and sampling criteria; thereby, a corpus is considered to be authentic and representative (Tognini-Bonelli 2001: 54). Moreover, Leech (1992: 107) outlined main features of a corpus appliance: 1) focus on linguistic performance, rather than competence; 2) focus on linguistic description, rather than linguistic universals; 3) focus on quantitative, as well as qualitative models of language; and 4) focus on a more empiricist, rather than rationalist view of scientific inquiry. These points of focus refer to the important demand of a register analysis that is to identify empirically the actual features of natural language.

Another important aspect of using a corpus in such analyses is its suitability for contrastive studies. Connor and Moreno (2005: 4) have emphasized that a corpus is beneficial for contrastive analyses. In addition, they have regarded the action of comparison by stating that "It is understood that apples should not be compared with oranges" and suggesting tertium comparationis that is in simple terms a common ground for a comparison (ibid.). When conducting a contrastive study where the two corpora are used it is mandatory to apply tertium comparationis in the processes of a corpus assembling and analysis production so that the findings would be comparable and lead to meaningful conclusions.

1.1.3. Lexical Bundles

It was found that language is commonly constructed and produced on the basis of formulaic expressions, words that go together, rather than composed of separate words (Barlow 2011: 35-37). Moreover, Wray and Perkins (2000: 1) describe a formulaic expression as "a sequence, continuous or discontinous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar". Linguistic research provide ample evidence that words naturally tend to co-select a specific lexical companion to establish a partly or fully meaningful word sequence; and, the meaning of a formulaic expression is perceived holistically. For instance, Vilkaitė (2016: 28, 35, 40) in her study where she analysed formulaic language in BNC *Baby* corpus discovered that formulaic expressions of different kinds (idioms, phrasal verbs, collocations and lexical bundles) make up about 41% of the English language and the 29% of these expressions are lexical bundles.

Lexical bundle as a specific type of formulaic expression was first described in the Longman Grammar (Biber et al. 1999). Over the following years, lexical bundles attracted the attention of many linguists (cf. Biber et al. 2004, Hyland 2008, Cortes 2015). According to Biber et al. (1990: 990) a lexical bundle is a sequence of three or more words that naturally reoccur together in a discourse. Biber et al. (1999) in their descriptive grammar *Longman Grammar of Spoken and Written English* have described lexical bundles that are used in the various registers of the English language. Scholars focused on a large-scale data set, Longman Spoken and Written Corpus of 40 million words (LSWE) (1991: 5). Throughout the observation of descriptively and empirically analysed data, it was pointed out that lexical bundles have three main aspects which should be considered (Biber et al. 1999: 990-991). Firstly, a small lexical bundle of three-words can be merged into a four-word lexical bundle, for instance *I don't think* can appear in *I don't think so*; thus, the smaller lexical bundles occur much more frequently in the analysed

data than the bigger ones (Biber et al. 1999: 990). In addition to this, a lexical bundle is considered as such if it has an enormous frequency in an analysed register because in the analyses of lexical bundles the paramount quality is frequency (Cortes 2015: 204). Thus, the second aspect for the operational definition of LBs is their frequency; how many occurrences is enough for a word chunk to become a lexical bundle of relevant currency? Biber et al. (1999: 992-993) considered lexical sequences as lexical bundles only if four-word sequences occurred at least ten times per million words and five/six-word sequences occurred at least five times per million words; these frequency levels differ for the reason mentioned previously that smaller lexical bundles occur more frequently because they can be incorporated into the bigger ones thereby causing the five/six-word lexical bundles to be less common. However, the occurrences were far more frequent than expected, thus, the new benchmarks for the frequency (20, 40 and 100 words per million) were proposed by Biber et al. (1999: 1001). The third aspect which must be taken into account is the distribution of lexical bundles in the analysed data. Biber et al. (1999: 990-991, 993) have emphasised that lexical bundles have to be widely spread across texts in a corpus in order to avoid individual idiosyncrasies of a speaker or writer.

In their later work Biber et al. (2004: 372) started questioning how the characteristics of lexical bundles could be defined because the empirical studies that were carried out on the topic of the lexical bundle analysis gave little attention for a such facet. Scholars adopted frequency-driven approach by which the lexical bundles were detected and subsequently qualitatively analysed (ibid.: 373). It was noticed that the lexical bundles have certain patterns and functions. The structural patterns revealed three main types of lexical bundles, namely: lexical bundles that incorporate verb phrase fragments, e.g. *is one of the*; lexical bundles that incorporate dependent clause fragments, e.g. *that there is a*; and, lexical bundles that incorporate noun and prepositional phrase fragments, e.g. *one of the things*, *at the end of* (ibid.: 381). The three main functions of the lexical bundles were identified as: stance expressions, e.g. *it is possible to*; discourse organisers, e.g. *on the other hand*; and, referential expressions, e.g. *in the case of* (ibid.: 384-388), which reflect the three functions of language as proposed by Michael Halliday in his functional grammar (in Biber 2006). This study has highlighted the importance of structure and function of lexical bundles which facilitates the clarification of how lexical bundles can functionally affect discourse in certain registers thereby letting the scholar to observe a comparison between two target registers. In addition to this, Biber et al. (2004: 397-398) noted that structure and function tend to have a relationship, for instance, the majority of stance bundles are generated of dependent clause fragments, whereas referential bundles are

usually generated of noun or prepositional phrase fragments. Only discourse organisers can be detected in all three structural types (ibid.). Cortes (2015:) contributes as well to Biber et al. (2004) insight into the types and functions of lexical bundles. She stresses that lexical bundles have strong grammatical correlates which are useful in identifying the tendencies of the use of lexical bundles in various registers (Cortes 2015: 208).

1.1.4. Previous research on register analysis

The methodology for identifying the linguistic features gradually shifted throughout the years. The early method of *field, tenor* and *mode* was proposed by Halliday et al. (1976). It was systematically applied by scholars until Ferguson (1983) in his research noticed a distinctive technique for analysing the register. In his study the language variety was analysed by applying the 'locating' method. The main principle of this method is to identify the reoccurrences of the functional linguistic features in the dataset. Shortly after a decade, the reoccurrence principle was modified and named as *lexical bundles analysis* by Biber et al. (1999).

Subsequently, the lexical bundles approach materialized rapidly in the register research. Various studies have been conducted regarding different language variations. The paramount register that has been given the most attention is the academic discourse. Register analysts have analysed and compared various disciplines such as chemistry, physics, business, biology, applied linguistics, mathematics, history, medicine (Cortes 2004; Hyland 2008; Hyland & Tse 2009; Herbel-Eisenmann et al. 2010; Farvardin et al. 2012; Grabowski 2013, 2015; Cao 2021; Hussain et al. 2021; Yin & Li 2021). Another important aspect of the academic register was the investigation of the smaller sub-registers and their salient linguistic characteristics: in PhD abstracts (Li et al. 2020); in university spoken and written language (Biber et al. 2004; Biber & Barbieri 2007; Hernández 2013); in rated learner varieties of English (Chen & Baker 2014); in research article introduction section (Noor & Anwar 2020); in academic lectures (Nesi & Basturkmen 2015); in academic writing and speaking (Conrad & Biber 2005; Byrd & Coxhead 2010; Hong & Hua 2018). Moreover, much consideration was given to the analysis of texts that were produced by native and non-native English speakers (Chen & Baker 2010; Ädel & Erman 2012; Dontcheva-Navratilova 2010; Güngör & Uysal 2016, 2020).

On the other hand, it is noticeable that there are only a few studies in which the media register is analysed with the employment of the lexical bundles approach. For example, Rujirawan (2021) in her master dissertation has conducted a study of a journalistic register. The scholar sought to investigate the functional and structural characteristics of four-word lexical bundles

that are found in the articles on AI and 5G of the Forbes online magazine. Considering the data, Rujirawan has comprised a corpus of 362 texts which were analysed in accordance with the Biber et al.'s (2004) and Hyland's (2008b) taxonomies. The results have shown that the most prevalent structure of lexical bundles was noun phrase and prepositional phrase-based items, and functions mostly were research-oriented or referential expressions. Another register research was carried out by Xu and Sun (2022) who have conducted a contrastive analysis of a news register of China and UK newspapers. The scholars wanted to explore the structural aspect of lexical bundles that are detected in the news articles on medicine. For such task authors compiled two corpora that represented two varieties of language. Scholars have employed Biber et al.'s (2004) structural classification framework. The results of this study suggest that in both Chinese and English newspapers lexical bundles that incorporate noun, preposition or verb fragment are similarly widely apparent and the least frequent structure in both corpora was the dependent clause based lexical bundles. However, it was found that UK newspaper register contains fewer verb-based lexical bundles than the China newspaper register.

2. Data and methods

2.1. Data

In order to conduct a proper contrastive study of articles on astrophysics in academic and media registers two corpora were compiled each consisting of 200 texts for each language variety. The data of the academic register was retrieved from online British / American academic journals, namely: *The Astrophysical Journal* available from https://iopscience.iop.org/journal/0004-637X; *The Astrophysical Journal Supplement* available from https://iopscience.iop.org/journal/0067-0049; and, *Monthly Notices of the Royal Astronomical Society* available from https://academic.oup.com/mnras. The data of the media register was collected from the online British / American magazines and newspaper, i.e., *Astronomy Magazine* available from https://astronomy.com/; *Quanta Magazine* available from https://www.quantamagazine.org/; and, *The Guardian* available from https://www.theguardian.com/uk?INTCMP=CE_UK. More detailed information about each source can be found in Appendix 1.

In the academic astrophysics corpus ($Corpus_{ACAD}$ hereafter) consists of 1,762,641 tokens and the size of the media astrophysics corpus ($Corpus_{MEDIA}$ hereafter) is 250,720 tokens. The texts were manually revised to remove irrelevant sections, namely, reference lists, images and comments under images that are not linked to the text (e.g., *this picture can be downloaded from*) etc. The average lengths of an academic and media articles are approximately 8,000 and 1,200 words respectively. Moreover, the corpora were compiled on the basis of three criteria: firstly, texts are written in English; secondly, each corpus has to represent each different register, specifically academic and media; lastly, they have to be published in the time period of 2020–2022.

2.2. Methods

The study was carried out based on the lexical bundles approach (LBA) (Biber, Conrad & Cortes 2004). The lexical bundles analysis represents the corpus-driven approach to language study which involves the processing and examination of quantitative and qualitative data retrieved from corpora. Frequency lists were generated automatically with the help of AntConc 4.2.0 software (Anthony 2023). Whereby the verification of structural and functional categories within the context was accessible.

To delimit the study sample the operational definition of a lexical bundle (LB hereafter) had to be developed. Firstly, it was decided to analyse four-word LBs due to their optimal frequencies;

as mentioned earlier two or three-word bundles usually are often incorporated in the longer four, five-word bundles, thus their frequencies are very high and the five, six- word bundles are relatively rare. Secondly, it was needed to establish the cut-off point for the normalised frequencies. For this study a less conservative cut-off point of 24 occurrences per million was set (cf. Biber et al. (2004) used 40 occurrences per million) which corresponds to 42 raw instances in the $\text{Corpus}_{\text{ACAD}}$ and 6 raw instances in the $\text{Corpus}_{\text{MEDIA}}$. And thirdly, no cut-off point for dispersion was set for the reason that the greater attention was given for the frequency whose cut-off point framed the lowest dispersion points for each corpora, 11% (21 raw) of texts in the academic corpus and 3% (5 raw) of texts in the media corpus. To avoid idiosyncrasies LBs should occur in at least five different texts (Biber et al. 1999: 992). Moreover, it has to be noted that the contracted word forms incorporated in LBs were manually separated and treated as two individual words, e.g., don't => do not. Contractions were identified only in the $\text{Corpus}_{\text{MEDIA}}$. In addition to this, apostrophes also indicated the genitive case, thus these bundles were excluded because the suffix -s expresses grammatical and not lexical meaning. Lastly, numerals in LBs are marked with a number sign #. In the following sections of this thesis, the term *lexical bundles* and its acronym *LBs* will refer to items that meet the operational definition in terms of frequency and dispersion given in this paragraph.

However, it must be indicated that in the analysed data some specific structure types were found that slightly differed from the structural types of LBs proposed by Biber et al. (2004). Firstly, LBs that incorporate $\underline{1^{st}/2^{nd}\text{ person pronoun + VP fragment}}$ and from the first sight resemble verb fragment-based bundle are actually classified as dependent clause-based bundles which take form of $\underline{1^{st}/2^{nd}\text{ person pronoun + dependent clause fragment }}$ if they had a subordinator pointing to the dependent clause, e.g.:

we	*assume*	*that the*
1^{st} p.p.	VP fragment	that-clause fragment

In the analysed data such LBs indicate the dependent clause that subordinates the verb as in:

(1) ***We assume <u>that the</u>*** *<u>statistical errors on the polar and azimuthal angles are such</u>* <…> ($\text{Corpus}_{\text{ACAD}}$)

The second aspect which has to be pointed out is concerned with the LBs that incorporate *WH*-clause, *to*-clause and *that*-clause fragments. Biber et al. (2004) give examples of such bundles that begin with the subordinator (i.e. *that there is a, what I want to*) with an exception of *to*-

clause bundles that can be preceded by an adjective or verb. However, in this study it has been found that dependent clause-based bundles are also preceded by noun phrases, e.g.:

a black hole *that*

noun phrase that-clause fragment

These LBs show the noun post-modification by a subordinated clause in the analysed data:

(2) <…> *it leaves behind* ***a black hole*** ***that*** *sinks to the middle of the star cluster*. (Corpus$_{MEDIA}$)

Thus, such LBs were classified as dependent clause-based rather than noun phrase-based. Needless to say, such structure is apparent in the structural sub-categories of noun phrase-based bundles, i.e., noun phrase with other post-modifier fragment, nevertheless a post-modifier is not always necessarily clausal, it can be a noun in apposition, adverb, adjective, prepositional phrase, e.g.:

matter *in the universe*

noun phrase post-modifier as a prepositional phrase

2.3. Procedure

Firstly, the corpora compilation began by downloading the texts from the online platforms into the Microsoft Word documents which were later converted into the plain text documents. After the data was compiled the lexical bundle lists were generated by AntConc 4.2.0 software (Anthony 2023). Then followed the comparisons of normalised frequencies, structural and functional categorizations in accordance with Biber et al.'s (2004) LBA theory. The test of statistical significance (χ^2) was performed using the online calculator (https://www.socscistatistics.com/tests/chisquare2/default2.aspx) (2023).

3. Results and discussion

The results and discussion section will be structured in the following order. Firstly, the quantitative results will be presented in lists subsequently with the established comments and inferences. Secondly, the qualitative findings of structural and functional characteristics will be provided and illustrated with the charts, tables and examples.

3.1. Quantitative results

The results show that media register holds a greater number of lexical bundles than the academic register. It was found that the $Corpus_{MEDIA}$ contains twice as many LBs as the $Corpus_{ACAD}$ which is 304 LBs and 144 LBs respectively. The full lists of lexical bundles of each register can be found in the section of appendices.

Furthermore, as it can be seen from Table 2, it is relatively evident that the normalised frequencies in the media register seem to be considerably higher than in the academic register. For instance, out of the given 50 LBs the majority of LBs in media register are approximately twice and more frequent than in the academic register. A possible explanation for this might be that in the media articles the vocabulary that is used can be characterized as containing popular words corresponding to the topic. It means that the media articles present a content by which the general knowledge about the subject is discussed. Thereby certain words are excessively recurrent resulting in the composition of frequently recurrent LBs that are topic related.

Whereas in academic articles, the vocabulary that is used is essentially academic language despite the topic-specific words. The topic-specific words are not so conspicuous because each research article is usually specialized on a particular sub-topic of the discipline. Thus in this case there are no excessively recurrent ‘popular astrophysical’ words because the spectrum of the topic is much more wider and deeper than in the media article which results in containing more different words and less LBs.

Moreover, it is worth mentioning that some frequencies of lexical bundles in both registers are relatively high reaching 40 occurrences per million and more which are considered a high frequency by Biber et al. (2004). In $Corpus_{ACAD}$ such remarkably frequent LBs take 35% (51 raw) and in the $Corpus_{MEDIA}$ such LBs constitute 30% (91 raw) of all analysed LBs. Thus, taking into account the previous statement that media register contains more LBs than academic in quantity, nonetheless, it can be assumed in terms of the quality that academic register possesses slightly more, by 5%, high-frequency LBs than the media register.

Table 2. The most frequent 50 lexical bundles of each register with normalised frequency per 1,000,000 words.

NO	Academic	NormF	Media	NormF
1	as a function of	447	at the university of	431
2	in the case of	151	after the big bang	267
3	with respect to the	138	of the milky way	263
4	a function of the	120	at the center of	231
5	on the other hand	116	in the early universe	164
6	by a factor of	113	james webb space telescope	156
7	the line of sight	112	an astronomer at the	152
8	as well as the	103	the hubble space telescope	148
9	per cent of the	98	times the mass of	136
10	in this paper we	95	the end of the	124
11	we find that the	94	the mass of the	116
12	in this work we	90	of a black hole	116
13	of the magnetic field	81	an astrophysicist at the	112
14	is consistent with the	79	of the black hole	112
15	in this section we	75	the cosmic microwave background	112
16	we note that the	71	the event horizon telescope	112
17	the evolution of the	68	in the astrophysical journal	108
18	the presence of a	65	of the solar system	108
19	is shown in fig	64	one of the most	104
20	the fact that the	64	a supermassive black hole	104
21	are shown in fig	56	black hole at the	104
22	the mass of the	55	astronomer at the university	96

23	in the context of	54	mass of the sun	96
24	the effect of the	54	in the milky way	96
25	the position of the	54	center of the milky	96
26	the size of the	53	light years from earth	92
27	an order of magnitude	51	of the university of	92
28	in this case the	50	hole at the center	92
29	the shape of the	50	million years after the	92
30	of the order of	48	for the first time	88
31	the total number of	48	of gas and dust	88
32	on the order of	47	the european space agency	84
33	the origin of the	47	in the solar system	80
34	the properties of the	47	the speed of light	80
35	we assume that the	46	the large magellanic cloud	80
36	the location of the	45	when the universe was	80
37	along the line of	45	over the course of	76
38	as shown in fig	45	the university of California	76
39	can be used to	45	goddard space flight center	76
40	in the absence of	44	the size of a	76
41	is due to the	44	at the heart of	72
42	can be found in	43	of the early universe	72
43	in terms of the	43	into a black hole	72
44	of the x ray	43	the center of our	72
45	are listed in table	42	million light years away	68
46	in addition to the	42	at the end of	68
47	the value of the	42	the supermassive black hole	68
48	to that of the	41	the milky way and	68

49	the ratio of the	40	in our solar system	64
50	in the presence of	40	it is hard to	64

3.1.1. Shared LBs

Another significant aspect of the distribution of LBs in registers is the evidence that some of them are shared. Out of all (144 in $Corpus_{ACAD}$ and 304 in $Corpus_{MEDIA}$) analysed LBs from each corpus only 22 (15% of $Corpus_{ACAD}$ and 7% of $Corpus_{MEDIA}$) shared LBs were identified. They are presented in Table 3 below.

Table 3. The shared lexical bundles with normalised frequency per 1,000,000 words.

NO	Type	Academic NormF	Media NormF
1	in the case of	151	24
2	on the other hand	116	44
3	as well as the	103	36
4	the mass of the	55	116
5	the size of the	53	44
6	in this case the	50	32
7	the shape of the	50	36
8	the location of the	45	24
9	can be used to	45	24
10	in the form of	37	32
11	it is important to	35	24
12	at the end of	34	68
13	the end of the	33	124
14	the surface of the	32	32
15	a wide range of	31	24
16	at the same time	31	56
17	the speed of light	31	80
18	it is possible that	29	26
19	at a distance of	27	32
20	the formation of the	27	28
21	for the first time	27	88
22	the rest of the	27	44

This list of the shared LBs reflects something of a general idea about the astrophysics discipline/subject. However, the major variation of these LBs can be observed because their frequencies presumably are dependent on the register type. For example, LBs (1) *in the case of,* (2) *on the other hand* and (3) *as well as the* are more prominent in the academic register whereas LBs (4) *the mass of the* and (13) *the end of the* are more frequently occurring in the media register. In accordance with Biber et al.' (2004) study, it can be stated that the first three LBs are particularly typical 4-word bundles found in the academic register and the latter two are less common.

3.2. Qualitative results

In the following section the results of the structural analysis of lexical bundles in academic and media registers will be reported. Firstly, the structural distribution of LBs in astrophysics academic register will be provided. Secondly, the classifications in astrophysics media register will follow and subsequently, the comparison of the LB structure results of both registers will be given. Before proceeding to examine the results, it is important to note that the long terms of LB structural types will refer to, namely, NP-based LB (lexical bundles that incorporate noun phrase and/ or prepositional phrase fragments) VP-based LB (lexical bundles that incorporate verb phrase fragments) and DC-based LB (lexical bundles that incorporate dependent clause fragments) hereafter. Further labelling of the sub-categories of these three main ones, will be commented on while observing the examples.

A chi-square test of independence was performed to examine the relation between registers and structural / functional types of lexical bundles. The test showed that differences in the distribution of LBs across structural and functional types in academic and media registers is different, and this quantitative difference is statistically significant both for structural (χ^2 = 10.4035, p = .005507) and functional (χ^2 = 8.5333, p = .014028) categories of LBs with $p < .05$ in both cases. The following sections of this paper will report results of the qualitative analysis.

3.2.1. Structural types of LBs in academic register

In $\text{Corpus}_{\text{ACAD}}$ it has been found that lexical bundles that incorporate noun phrase fragments (NP-based LBs) which account for 62% of the study sample were the most frequent LBs that incorporate verb phrase fragments (VP-based LBs) and dependent clause fragments (DC-based LBs) constitute 27% and 11% which is approximately twice and six times less respectively in comparison with the NP-based LBs. The structural classification of LBs in astrophysics research article register can be seen in Figure 1.

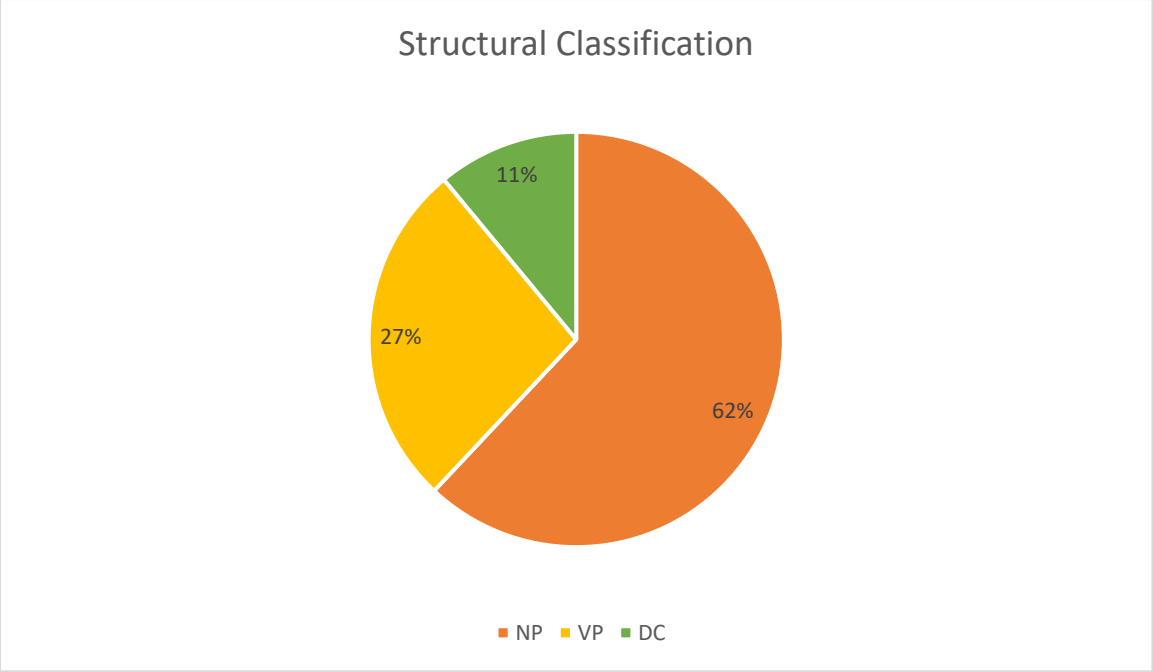

Figure 1. The structural classification of lexical bundles in academic register.

Taking into consideration the previous research academic register is pictured as 'noun-centric' thus the domination of NP-based LBs is common (cf. Biber & Conrad 2005, Byrd & Coxhead 2010). This is due to the purpose of the scientific research article which is to provide information about the analysed object and to explain why or how it is happening. For example, in this study, the majority of NP-based LBs include prepositions. A more detailed account of NP-based LBs is given in the following paragraph. In the chart below, Figure 2, five sub-categories of NP-based LBs are shown.

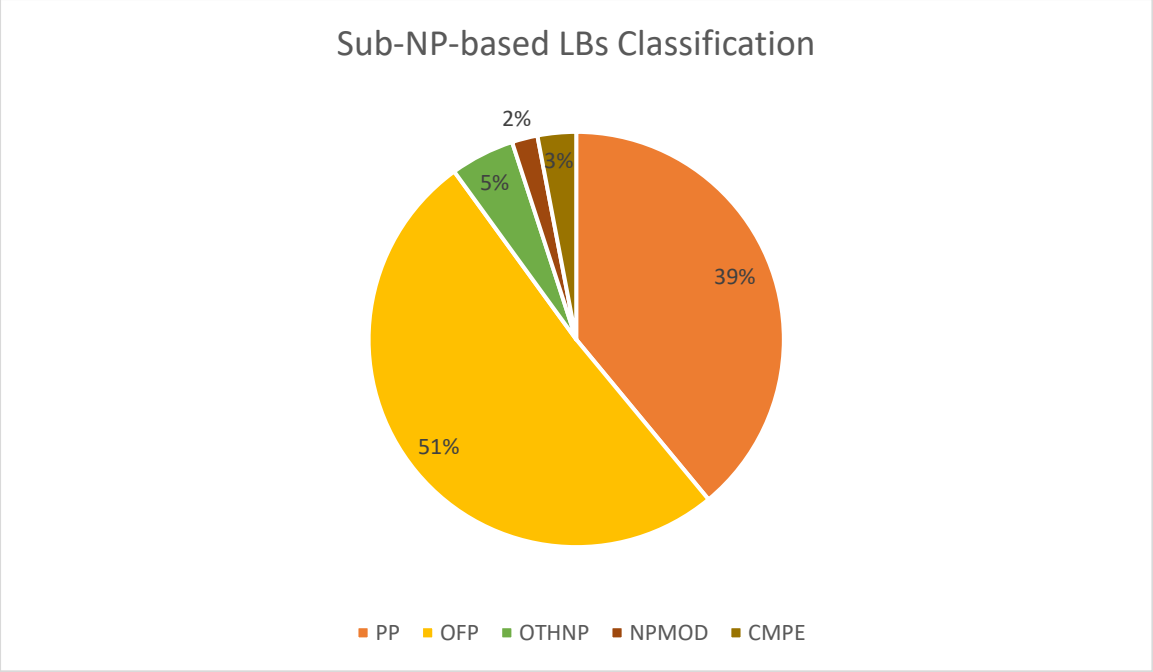

Figure 2. The sub-structural classification of noun phrase fragment based lexical bundles in academic register.

As it can be seen in the vertical bar chart, the LB that incorporates a noun phrase with an *of*-phrase fragment (OFP) is the most frequent sub-type which takes 51%. Another large percentage of 39% is taken LBs that incorporate a prepositional phrase fragment (PP). The less notable three sub-types were LBs that included other noun phrase expressions (OTHNP), comparative expressions (CMPE) and post-modifiers (NPMOD) that constitute 5%, 3% and 2% accordingly. The examples of LBs are provided in accordance with the frequency in Table 4.

Table 4. Examples of sub-classes of noun phrase fragment based lexical bundles in academic register.

Sub-structural type	**Examples**
(OFP) NP + of-phrase expression	*a function of the* *per cent of the* *the evolution of the*
(PP) NP + prepositional phrase expression	*with respect to the* *in this paper we* *in this case the*
(OTHNP)	*the magnetic field strength*

Other noun phrase expression (CMPE)	*the same as in*
Comparative expression (NPMOD)	*as well as the*
NP + post-modifier	*the differences between the* *the anonymous referee for*

Regarding the classification of VP-based LBs four sub-categories were identified. They are presented in Figure 3.

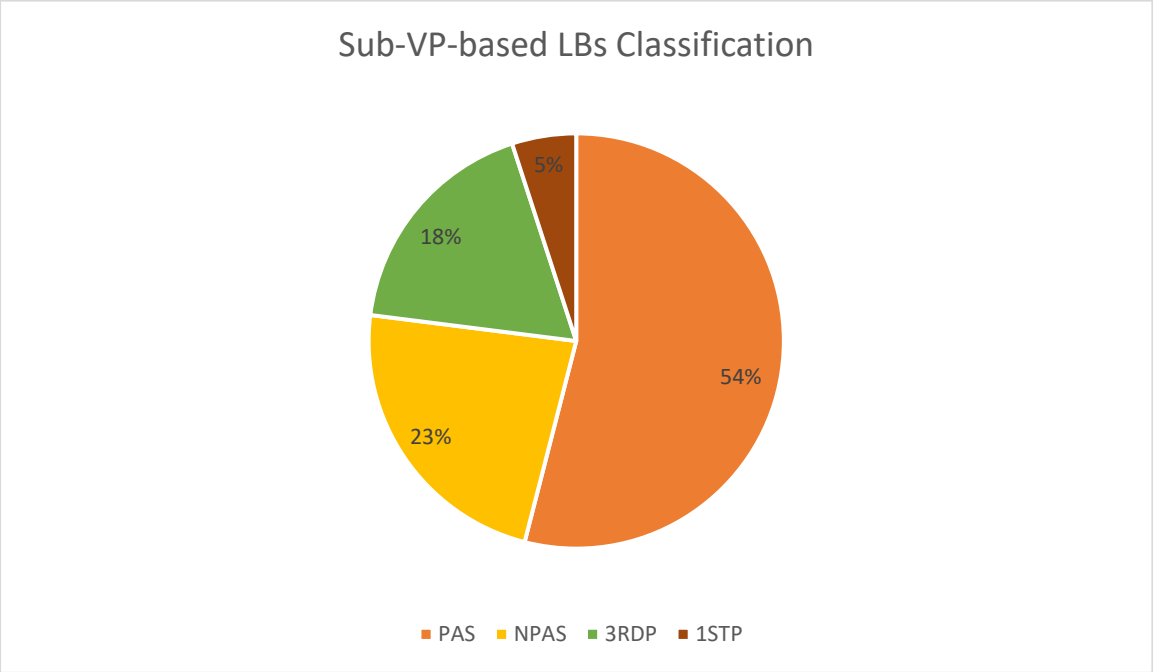

Figure 3. The sub-structural classification of verb phrase fragment based lexical bundles in academic register.

The top sub-type of verb structure is the verb phrase with a passive verb (PAS) which reaches 54% out of all the instances. Subsequently, the structures of verb phrase with a non-passive verb (NPAS) and 3rd person pronoun + verb phrase fragment (3RDP) follow with 23% and 18% respectively. And lastly, the least occurrent structure of 1st person pronoun + verb phrase fragment (1STP) appeared only in the 5% of the analysed VP-based LBs. The examples of LBs are provided in the order of decreasing frequency in Table 5.

Table 5. Examples of sub-classes of verb phrase fragment based lexical bundles in academic register.

Sub-structural type	Examples
(PAS)	*is shown in fig*
VP + passive verb	*can be found in*
	can be seen in
(NPAS)	*is due to the*
VP + non passive verb	*is the number of*
	is in agreement with
(3RDP)	*the magnetic field is*
3rd person pronoun + VP	*it is important to*
(1STP)	*in fig we show*
1st person pronoun + VP	*in section we describe*

The last major class of the DC-based LBs contains five sub-classes. The results are given in Figure 4.

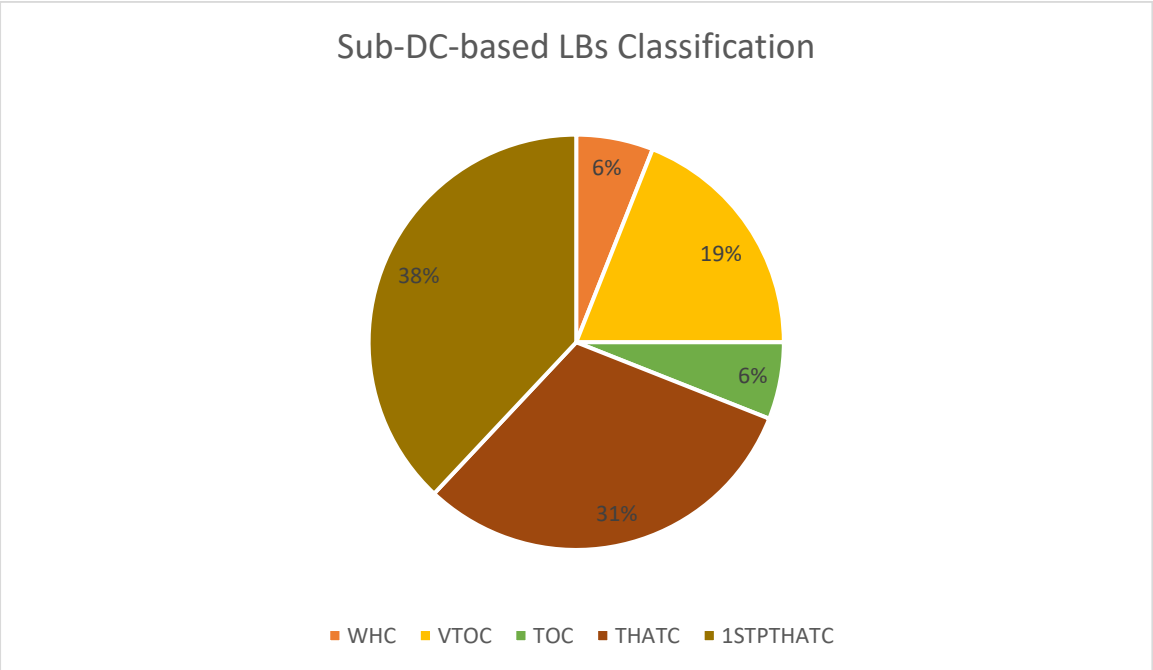

Figure 4. The sub-structural classification of dependent clause fragment based lexical bundles in academic register.

The chart presents phenomenal results where the most dominating structure is the 1st person pronoun + that clause fragment (1STPTHATC) which constitute to the 38%. The next following structure is that-clause (THATC) with the 31%. Another interesting part of these results is that the structure of verb + to-clause (VTOC) is approximately three times more

frequent than the bare to-clause (TOC) structure which is the least frequent structure together with the WH-clause (WHC) structure that make up only the 6% each of all DC-based LBs. The examples of LBs are provided in accordance with the frequency in Table 6.

Table 6. Examples of sub-classes of dependent clause fragment based lexical bundles in academic register.

Sub-structural type	**Examples**
(1STPTHATC) 1st person pronoun + that clause fragment	*we find that the* *we note that the* *we assume that the* *we found that the*
(THATC) That-clause	*the fact that the* *that there is a* *this means that the*
(VTOC) Verb + to-clause	*can be used to* *is assumed to be* *is expected to be*
(WHC) WH-Clause	*which is consistent with*
(TOC) To-clause	*to account for the*

Such distribution of frequency of DC-sub-structural categories is not coincidental. The most dominant structure contains 1st person pronoun with an active verb that is followed by a subordination. It is important to note that in the introduction section the typical features of hard sciences register were highlighted as: the frequent use of Science Technology words, abstract nouns and descriptive adjectives, acronyms, symbols / signs; and, the use of premodification rather than post-modification, nominalisation, non-predicative verbs, passive voice, long sentences (Li & Li 2015: 161-163, Parkinson 2013: 164-165). However, academic astrophysics register seems to agree not with all the typical characteristics which is specifically the usage of active voice instead of passive. Tarone and her colleagues have argued why authors choose an active voice over a passive; because, in many cases where form *we plus active voice* could be written in the *passive* form:

> (2a) Previously (Stoeger 1976b) we pointed out that the usual way of placing boundary conditions at rms was inadequate and misleading. (p. 216)

where the passive was clearly possible, as in our own paraphrase of
(2a): (2b) Previously (Stoeger 1976b) it was pointed out that . . . (our paraphrase)
(1981: 125)

Scholars came to the conclusion that the preference depends on what the author is writing about (Tarone et al. 1981: 135). In the current analysed data the generalizations of the preference are also evident. For example, the generalization no. 2, *we* is used to describe the author's own work (example (3)) and the passive to describe the work of others (example (4)):

(3) ***We note that the*** *destruction that we see is somewhat less extreme than in Garrison-Kimmel et al. (2017b).* ($Corpus_{ACAD}$)

(4) *However,* ***it should be noted*** *that Nesvorný et al. (2010) showed from their dynamical model* <…> ($Corpus_{ACAD}$)

Moreover, if we look at the normalised frequencies of these two specific LBs, the active voice is more prominent than the passive, 71 versus 28 occurrences per million words. Thus, the passive / active voice usage in astrophysics research article register presumably did not change drastically over the 40 years, taking into account Tarone et al.' and current study's results.

3.2.2. Structural types of LBs in media register

The results have revealed that in astrophysics media register the most common structure is the NP-based accounting for 82%. The VP-based and DC-based structures constitute 11% and 7% respectively. This register shows an immense degree of the structural differentiation where the NP-based LBs are roughly 7 and 12 times more frequent than the VP-based LBs and DC-based LBs. The structural classification of LBs can be observed in Figure 5 below.

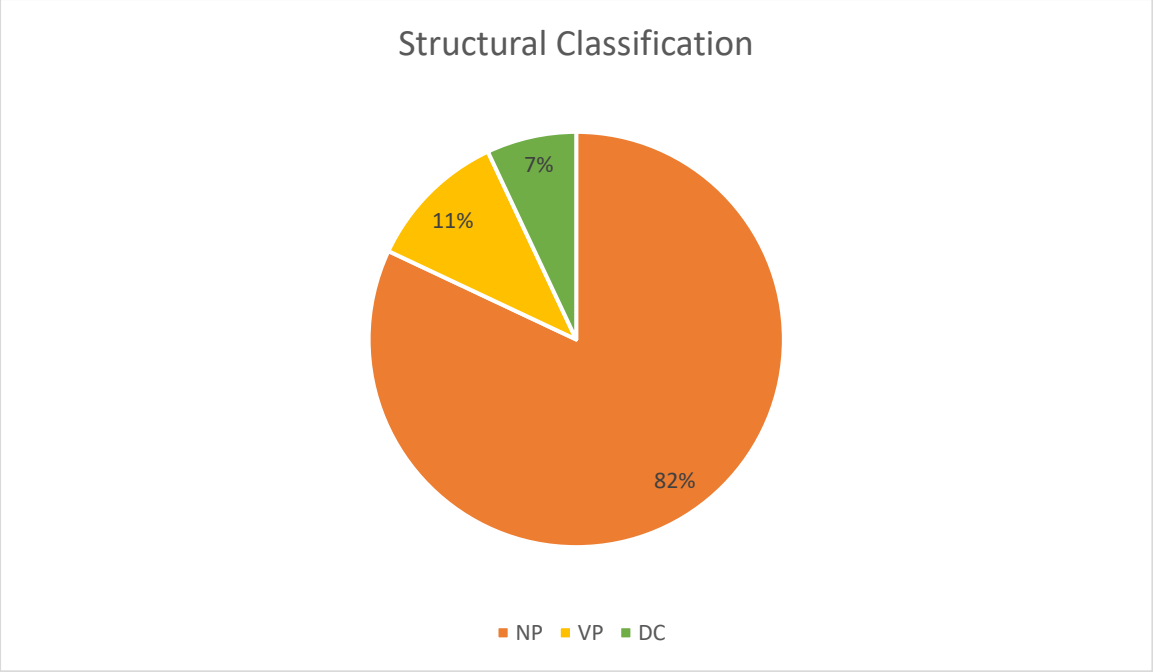

Figure 5. The structural classification of lexical bundles in media register.

Comparing the results of this study with the findings of the previous research which were not conducted much, it can be suggested that the structural distribution of LBs in media register is somewhat similar. As mentioned formerly in the introduction section in Rujirawan's (2021: 28) study the percentages showed such order: NP-based LBs 63%, VP-based LBs 30% and DC-based LBs 8%. The results in Xu and Sun's (2022: 205) research displayed this order: NP-based LBs 80% VP-based LBs 13% and DC-based LBs 7%. The frequency of the VP-based structure varies, however it is clear that the dominant structure is NP-based and the least prominent one is the DC-based structure which does not even reach 10% out of all analysed LBs. A more specific sub-classification of NP-based LBs is presented in Figure 6 below.

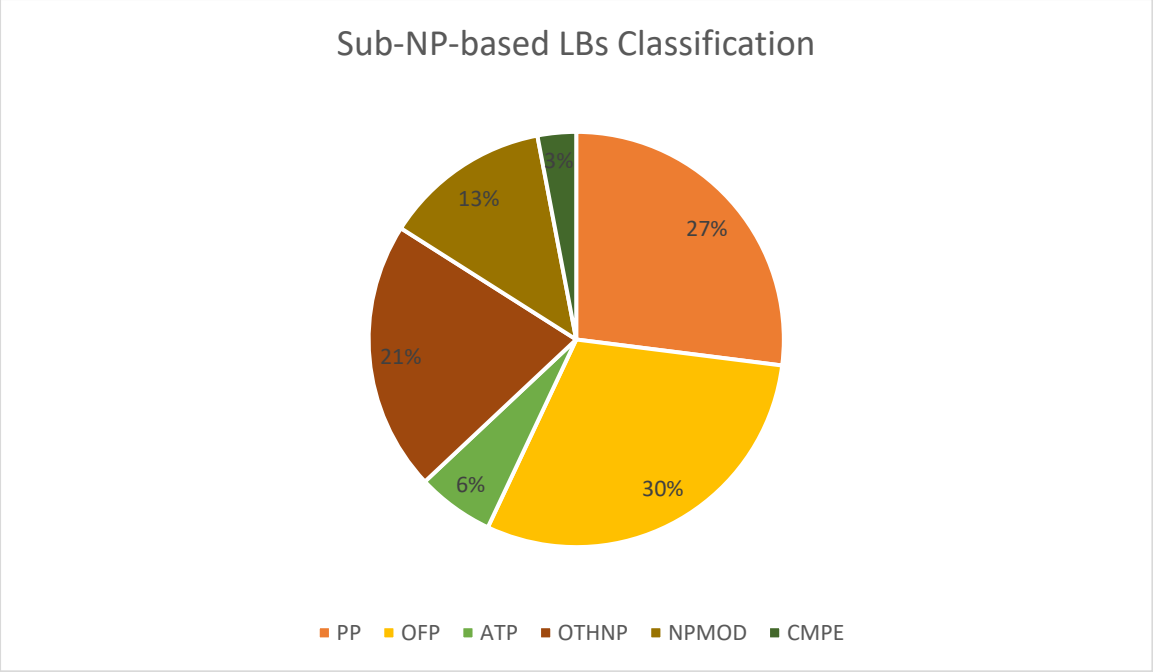

Figure 6. The sub-structural classification of noun phrase fragment based lexical bundles in media register.

Altogether six sub-classes of NP-based structure were established. The structure of LBs that incorporate of-phrase fragment (OFP) and prepositional phrase fragment (PP) took the largest percentages of 30% and 27% respectively. Surprisingly LBs that express other noun phrase (OTHNP) constitute quite high percent of 21%. After these leading sub-categories, structures of noun phrase + post-modifier (NPMOD), 13%, at-phrase fragment (ATP), 6%, and comparative expression (CMPE), 3%, follow. The examples of LBs are provided in accordance with the frequency in Table 7.

Table 7. Examples of sub-classes of noun phrase fragment based lexical bundles in media register.

Sub-structural type	**Examples**
(OFP) NP + of-phrase expression	*the end of the* *one of the most* *the speed of light* *the centre of our*
(PP) NP + prepositional phrase expression	*at the university of* *after the big bang* *in the early universe* *of the black hole*
(OTHNP)	*james webb space telescope*

Other noun phrase expression	*the hubble space telescope* *the cosmic microwave background* *the event horizon telescope*
(NPMOD) NP + post-modifier	*million years after the* *matter in the universe* *the black hole in*
(ATP) NP + at-phrase	*an astronomer at the* *an astrophysicist at the* *black hole at the*
(CMPE) Comparative expression	*times more massive than* *times stronger than earth* *galaxies like the milky way*

It is necessary to indicate that the structure of LB that incorporate at-phrase fragment (noun phrase + at-phrase) was separated from the structure of noun phrase + post-modifier (NPMOD) as a distinct sub-category due to its notable frequency. LBs that contain at-phrase fragment usually behave as a post-modifier of a noun, e.g.:

(5) ***an astronomer at the*** *University of Leicester*. ($Corpus_{MEDIA}$);

or as a place adverbial, e.g.:

(6) ***black hole at the*** *centre of a neighbouring galaxy*. ($Corpus_{MEDIA}$).

The evidence suggests that the half instances of structure of noun phrase + post-modifier in the astrophysics media register which signal the post-modification and additional place information addition is commonly conveyed through the at-phrase.

The sub-structures of VP-based LBs will be explained in the following paragraph. In the figure 7 the sub-categorization can be observed.

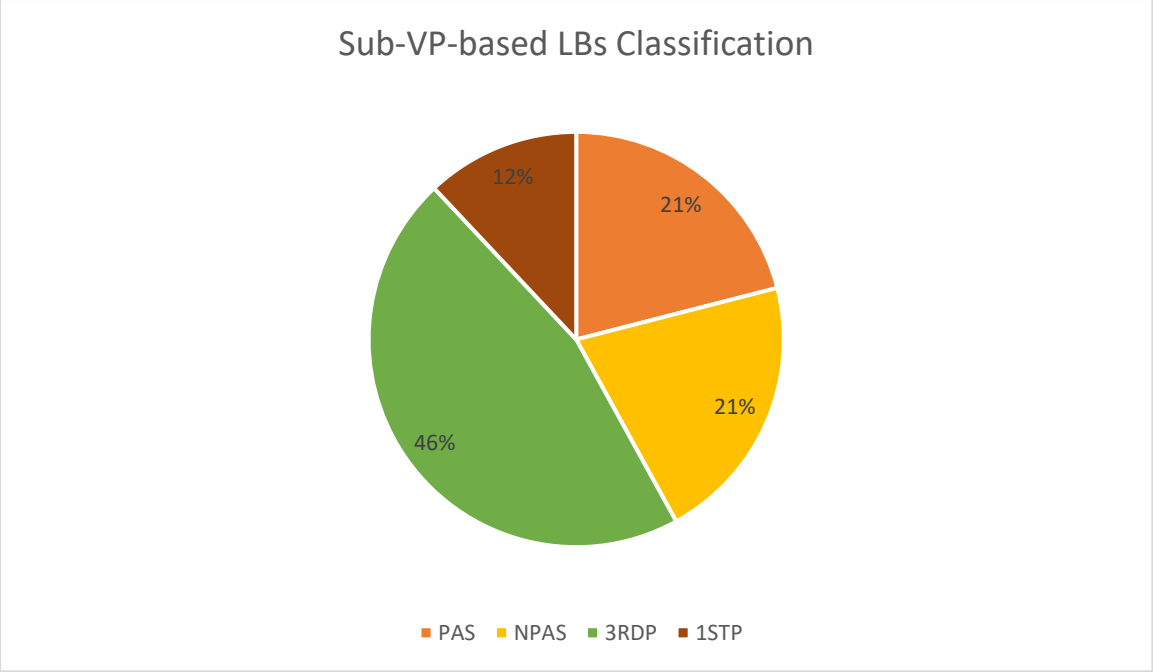

Figure 7. The sub-structural classification of verb phrase fragment based lexical bundles in media register.

The largest percentage of 46% is taken by LBs that possess the structure of 3rd person pronoun + verb phrase fragment (3RDP). The structures of verb phrase with a non-passive (NPAS) and passive verb (PAS) constitute the same percentage of 21% each. The least occurrent structure is 1st person pronoun + verb phrase fragment (1STP) which makes up to only 12%. The examples of LBs are presented in accordance with the frequency in Table 8.

Table 8. Examples of sub-classes of verb phrase fragment based lexical bundles in media register.

Sub-structural type	**Examples**
(3RDP) 3rd person pronoun + VP	*it is hard to* *the milky way is* *the universe is expanding* *a black hole is*
(NPAS) VP + non passive verb	*is one of the* *are some of the* *is not the only* *must see cosmic objects*
(PAS) VP + passive verb	*in a paper published* *was not involved in* *is known as the*

(1STP)	*we do not know*
1st person pronoun + VP	*we do not have*
	I do not think

Putting aside the usage of dummy pronoun *it*, the frequent use of the sub-structure 3rd person pronoun + verb phrase fragment could be explained by one factor. After a careful reading of astrophysics online magazine articles, it has been noticed that some parts of the produced text resemble a narrative by which certain stories or descriptions about the scientists, their findings (example (7)) or about the creation of the universe or the existence of particular cosmic objects (example (8)) are told and for that reason the 3rd person is naturally used:

(7) *A Surprise Discovery Points to the Source of Fast Radio Bursts. After a burst lit up their telescope "like a Christmas tree," **astronomers were able to** finally track down the source of these cosmic oddities. On the morning of April 28, a newly built radio telescope was monitoring the quiet skies over British Columbia when it caught the flash that would change everything. One of the telescope's duties was to search for fast radio bursts — millisecond-long blips that, until then, had always come from distant galaxies.* (Corpus$_{MEDIA}$)

(8) *N. Bartmann theory predicts that **supermassive black holes are** born in the hearts of dusty starburst galaxies. And given enough time, they throw off the surrounding shroud of gas and dust to reveal themselves as extremely luminous quasars.* (Corpus$_{MEDIA}$)

Thus, the usage of the structure of the 3rd person + verb phrase fragment is presumably common in astrophysics media register.

The sub-categorization of DC-based LBs is given in Figure 8 below.

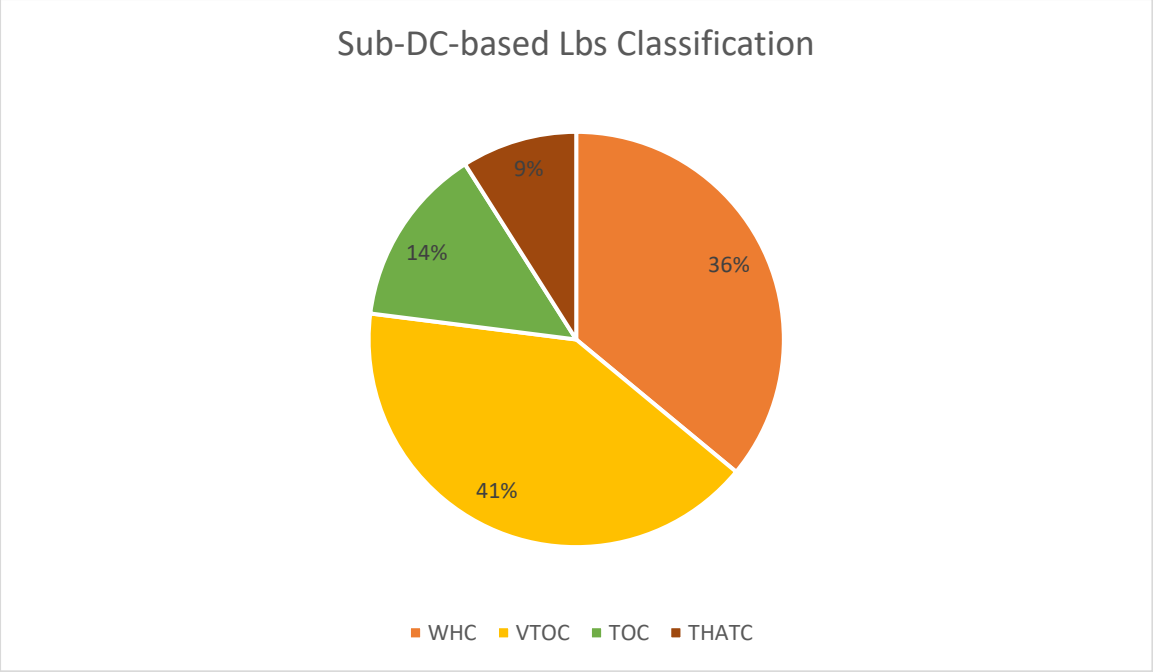

Figure 8. The sub-structural classification of dependent clause fragment based lexical bundles in media register.

Most noticeable sub-structures are verb + to-clause (VTOC), 41%, and WH-clause (WHC), 36%. The less frequent were bare to-clause (TOC) and that-clause (THATC) structures which constitute 14% and 9% respectively. The examples of LBs are provided in accordance with the frequency in Table 9.

Table 9. Examples of sub-classes of dependent clause fragment based lexical bundles in media register.

Sub-structural type	**Examples**
(VTOC)	*have been able to*
Verb + to-clause	*may be able to*
	be able to see
	might be able to
(WHC)	*when the universe was*
WH-clause	*what is going on*
	when it comes to
(TOC)	*to learn more about*
To-clause	*to figure out how*
	to come up with
(THATC)	*gas and dust that*
That-clause	*a black hole that*

It should be noted that the structures of the verb + to-clause (VTOC) and WH-clause (WHC) are so salient in the media register for the same reason mentioned in the previous paragraph, that the astrophysics magazine articles largely focus on the description of events which leads to the conspicuously apparent usage of 3rd person, and its ability / skill to do something. DC-based LBs that incorporate verb + to-clause are almost always preceded by the 3rd pronoun subject in the analysed corpus:

(13)	Researchers		measure
	Astronomers	*have been able to*	deduce
	Theorists		home

(Corpus$_{MEDIA}$)

Furthermore, the majority of DC-based LBs with WH-clauses behave as adverbials which add up to the astrophysics magazine 'narrative':

(9) *Blazing at **a time when the** universe was just 770 million years old, ULAS J1120+0641 contains a supermassive black hole 2 billion times the mass of our Sun.* (Corpus$_{MEDIA}$)

(10) *Most scientists think the dinosaurs — along with countless other creatures — were wiped out some 6 **million years ago when** a space rock slammed into Earth.* (Corpus$_{MEDIA}$)

3.2.3. Comparison of structural types of LBs between academic and media registers

In this section the similarities and differences of the structural classification of lexical bundles in astrophysics academic and media registers will be discussed. The comparison of the main three categories, NP-based, VP-based, DC-based LBs is presented in Figure 9.

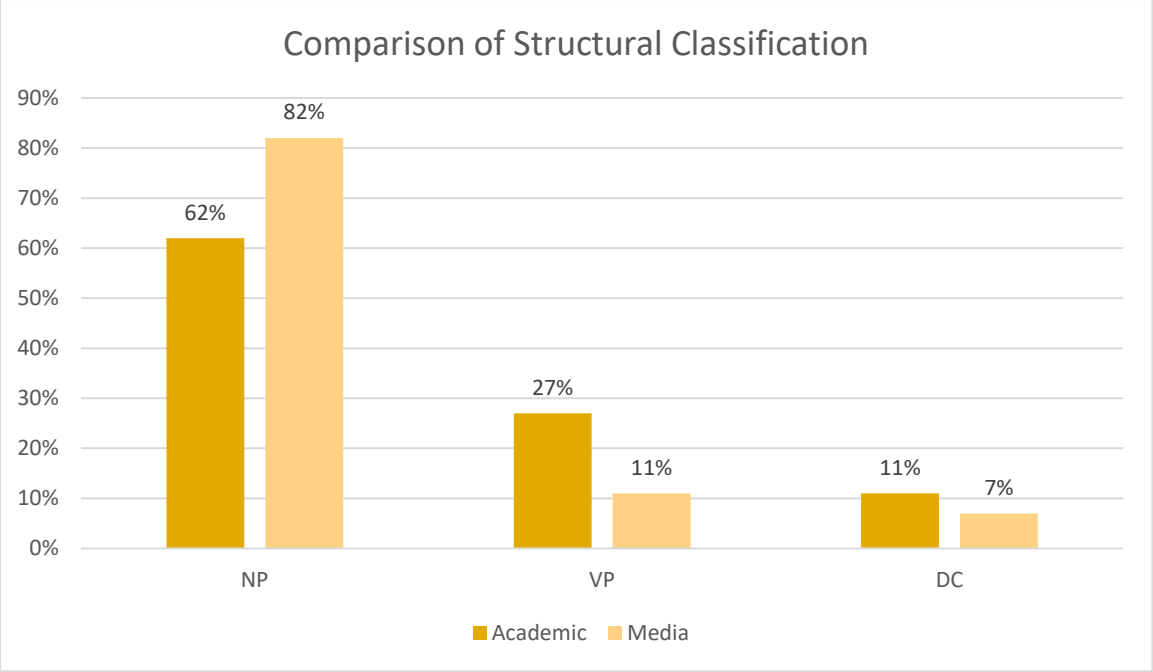

Figure 9. The comparison of structural classification of lexical bundles in the two registers.

From the general structural classification observation, it can be stated that the LBs that incorporate noun phrase fragment (NP) is 20% more frequent in the media register than in the academic. However, the VP-based LBs (VP) are as twice more occurrent in the academic register. Moreover, the frequencies of DC-based LBs (DC) constitute to the lowest percentages in both registers, nonetheless, the dependent clause category is slightly more prominent in the academic register. Overall, we can see from the chart that the distribution of the major structural classes in the two registers has an identical order, from the most frequent to the least: NP, VP. DC-based structures. In order to have a better understanding about this distribution a more detailed examination is needed, thus in the following paragraph, the comparison of structural sub-categories will be provided. In Figure 10 the first comparison of NP-based LBs sub-classes is given.

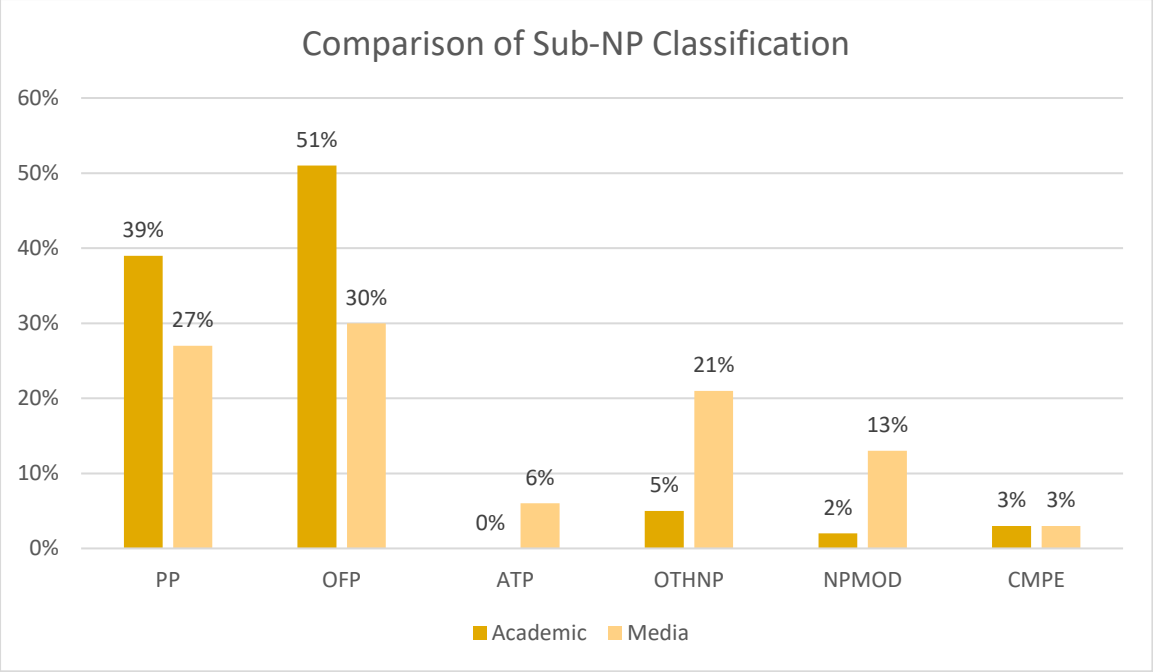

Figure 10. The comparison of sub-structural classification of noun based lexical bundles in the two registers.

The foremost difference between the registers is the presence of the noun phrase + at-phrase (ATP) structure which was established in the media register and not a single instance of that structure was identified in the academic register. Moreover, concerning the academic register on the one hand the quite higher frequencies of prepositional phrase (PP) and of-phrase (OFP) structures are apparent; on the other hand, it is evident that the structures of other noun expression (OTHNP) and noun phrase + post-modifier (NPMOD) are significantly less occurring when comparing them with the frequencies of the media register. Only the structure of comparative expression (CMPE) is equivalent in both registers. According to Biber et al. (2004: 382) in academic register LBs are more phrasal rather than clausal. Thus, it seems that the media register is even 'nounier' than the academic. For example, in the media register many instances of proper nouns or bare noun phrases are extremely frequent, e.g., *james webb space telescope, a supermassive black hole, the european space agency, our own milky way* or the LBs with a post-modification are also frequent, e.g., *million years after the, the milky way and, matter in the universe, the noble prize in.* Whereas, in academic register the preference of of-phrase and prepositional phrase bundles is evident, e.g., *the effect of the, the properties of the, in the case of, by a factor of.*

The second comparison of VP-based LBs sub-categories is presented in Figure 11.

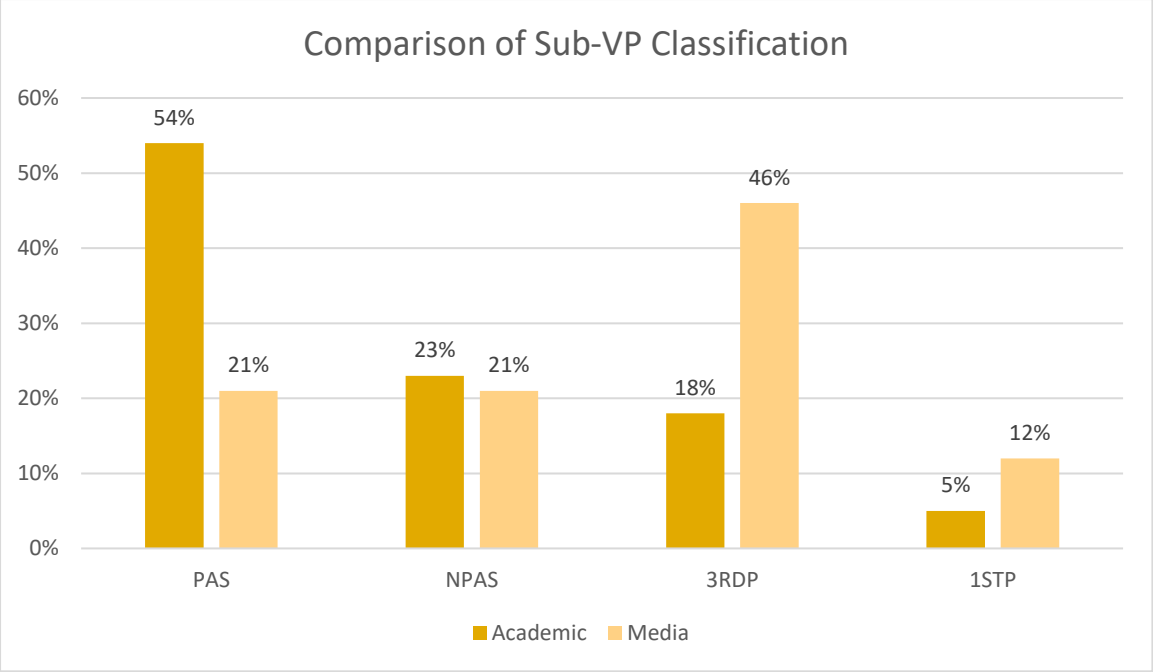

Figure 11. The comparison of sub-structural classification of verb based lexical bundles in the two registers.

Two main differences were identified while comparing the VP-based LBs. Firstly, the chart shows that the structure of verb phrase with a passive verb (PAS) is approximately as twice more frequent in the academic register than in the media. Moreover, the frequencies of the structure of verb phrase with a non-passive voice (NPAS) are practically similar. Secondly, in the media register the usage of 3rd person pronoun + verb phrase (3RDP) structure is roughly two times more occurring than in the academic register. In contrast to the academic register where such LBs would usually contain the 3rd person pronoun functioning as a dummy *it* that fills subject's syntactical place (e.g. *it is clear that*), it has been noticed that in the media register that type LBs mostly have a normal referable subject as the 3rd person (e.g. *a black hole is*). In addition to this, the structure of 1st person pronoun + verb phrase (1STP) is again more preferable in the media register. This distribution of VP-based LBs might suggest the fact that in the astrophysics academic prose personal pronouns and active voice is less common, whereas in the media articles it is more conventional. Nonetheless, the observation of the classification of academic DC-based LBs shows the 1st person pronoun usage from another angle through a different sub-structure. Considering the categorization of VP and DC-based LBs in the data and methods section it was proposed that a structure of 1st person pronoun + verb + that-clause fragment would be determined as a dependent clause based LB, thus, the instances of these LBs were deported to the DC-based LBs' classification cluster thereby, the

personal pronoun manifestation that is observed in the figure 11 might be a bit misleading when considering the overall usage of the passive and active voice in the analysed registers.

The third comparison of DC-based LBs sub-categories is presented in Figure 12.

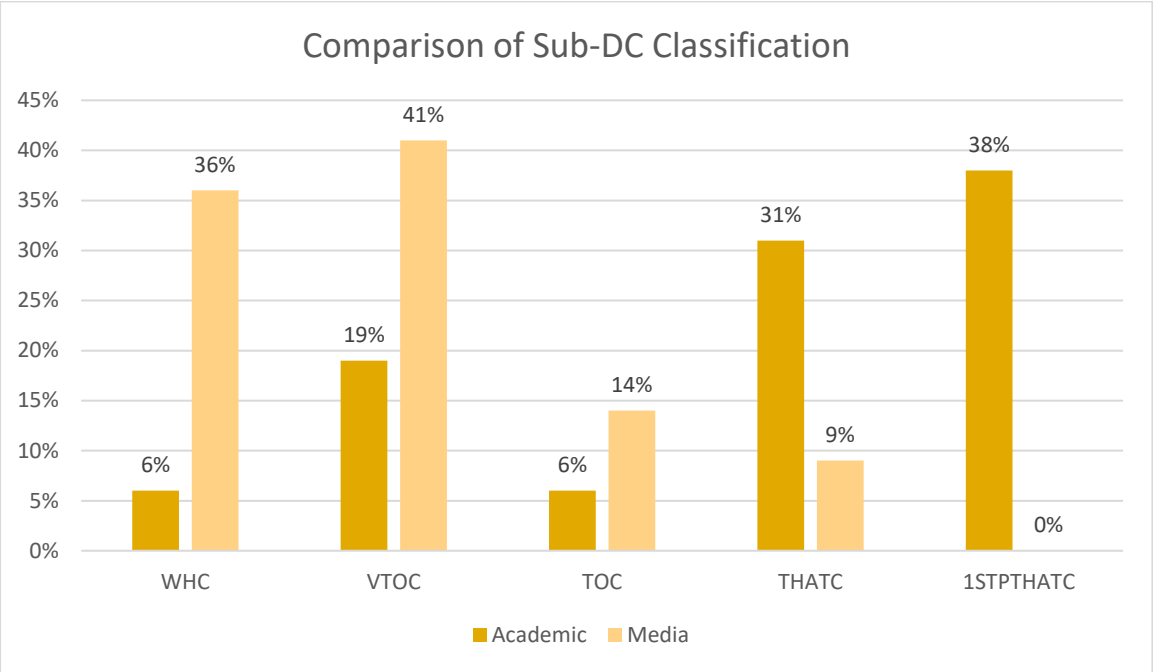

Figure 12. The comparison of sub-structural classification of dependent clause based lexical bundles in the two registers.

The comparison of DC-based LBs has shown a discernible dichotomy between the preferred sub-structures of LBs in the two registers. In the media register LBs that incorporate WH-clause fragment, verb + to-clause and bare to-clause fragments prevail, whereas, in the academic register that-clause fragments, bare and with 1st person, are more apparent. The sub-structure of 1st person + that-clause (1STPTHATC) is found nearly exclusively in the academic register. In the media corpus only one LB was identified which has an analogous structure of 1st/2nd person + WH-clause (e.g. *do not know what*). As stated previously, in this structural classification we can additionally observe that the active voice together with the 1st person pronoun in astrophysics academic register is used quite often. It can be assumed that the identified DC-based LBs (as well as NP and VP-based) in both registers somehow but certainly represent the intention of each register: a media article describes what is done and what can be done; while an academic paper explains how and why it is done in the field.

3.2.4. Functional types of LBSs in academic register

Moving on now to consider the functional classification of lexical bundles in astrophysics academic and media registers. The functional classification rather mirrors the structural distribution where NP-based LBs usually function as referential bundles, VP and DC-based LBs as stance bundles (Biber et al. 2004: 397-398). The organization of this paragraph follows such order: firstly, the general classifications of LBs in the two registers will be provided; secondly, the more detailed comparison of functional categorisations together with the sub-classes of LBs of both registers will be given. All findings will be presented in charts together with comments and examples.

The results have shown that the referential (R) function is the most frequent, 72%, in the academic register. Stance (S) and discourse organising (DO) functions are taking 20% and 8% respectively. The presentation of the classification is provided in Figure 13.

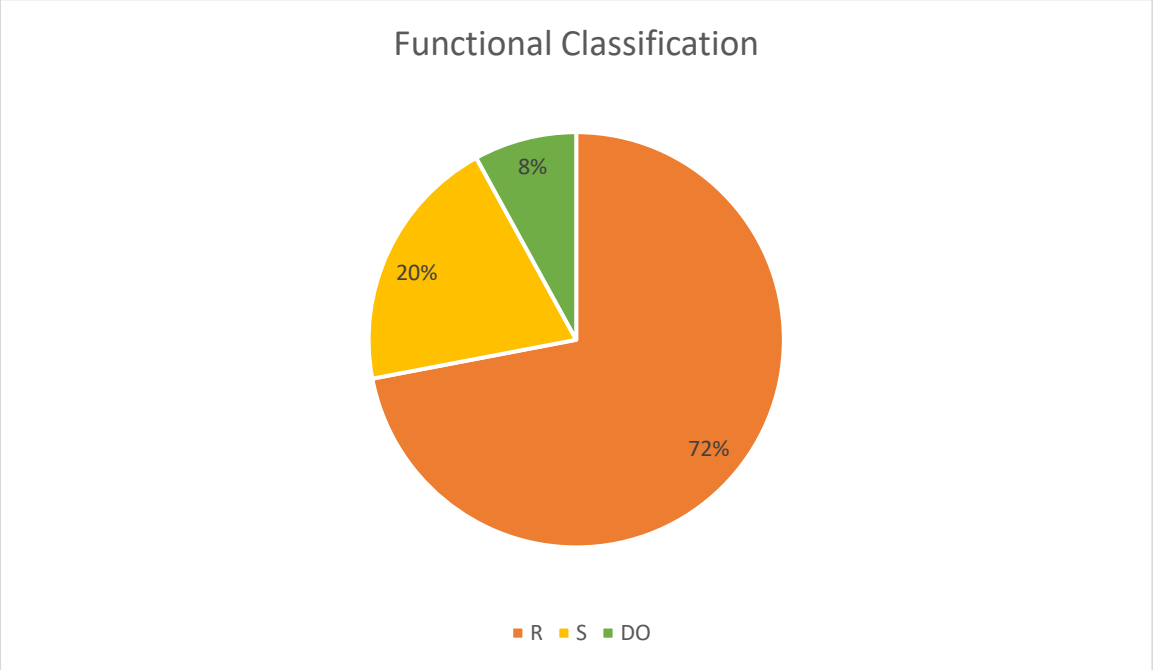

Figure 13. The functional classification of lexical bundles in academic register.

Commonly referential bundles function as specifications of attributes, e.g., *the presence of the*. This means that they are used to specify the information which is to be explained. Thereby the usage of referential LBs is a critical necessity when carrying out such tasks while producing an academic article. Stance bundles profoundly express dynamic and epistemic modalities, e.g., *can be used to, the fact that the*. LBs that function as discourse organisers in most cases convey topic elaboration and introduction features, e.g., *on the other hand, in this paper we*. The functional distribution of LBs is relatable to the previous research, for instance almost identical

interpretation of functions of LBs in the academic register was stated in Biber and Conrad's (2005: 68-69) study.

3.2.5. Functional types of LBs in media register

In the media register referential (R) bundles peaked to the top with the 88%. The function of stance (S) was identified in the 10% of the analysed LBs; and, the function of discourse organising (DO) accounts for only 2%. The functional categorisation is presented in Figure 14 below.

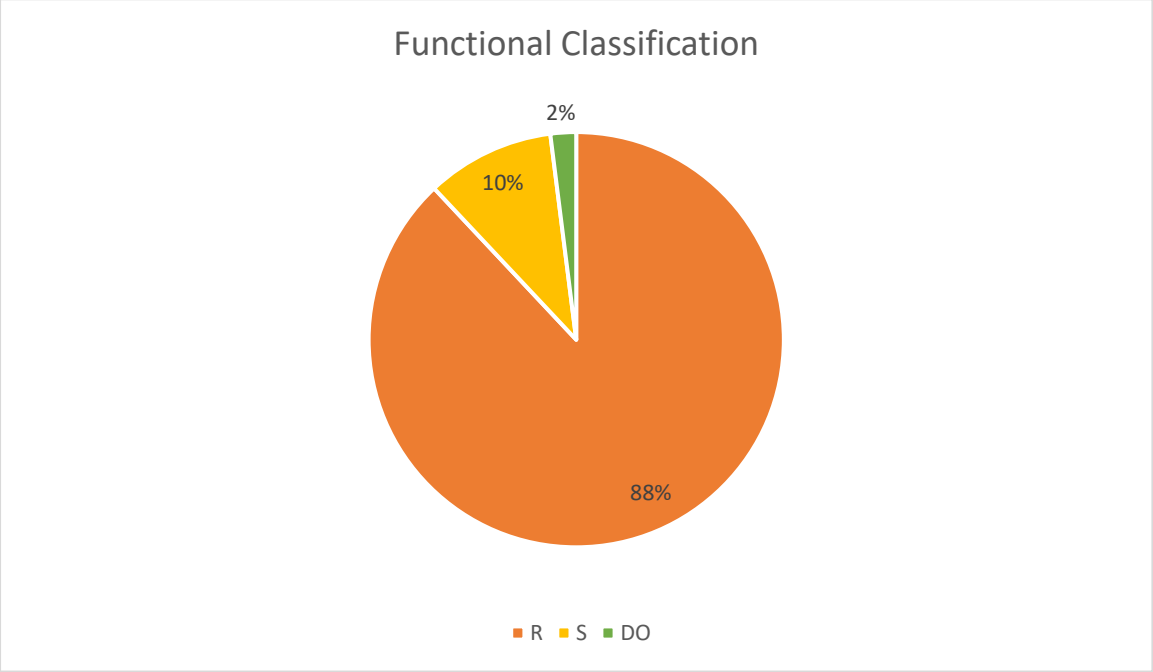

Figure 14. The functional classification of lexical bundles in media register.

LBs that are referential mostly function as identifications or specifications of attributes, e.g., *a black hole is, a wide range of.* Stance bundles essentially express dynamic modality and attitudinal stance, e.g., *can be used to, it is hard to.* Discourse organising bundles predominantly function as topic elaborations, e.g., *when it comes to*. The proportion of functional classification distribution is similar to Rujirawan's (2020: 39-41) findings. In general, the salient feature of the media register could be the highly low frequency of LBs that behave as discourse organisers.

3.2.6. Comparison of functional types of LBs between academic and media registers

In this section the similarities and differences of the functional classification of lexical bundles in astrophysics academic and media registers will be discussed. The comparison of the main three categories, referential, stance and discourse organising LBs is presented in Figure 15.

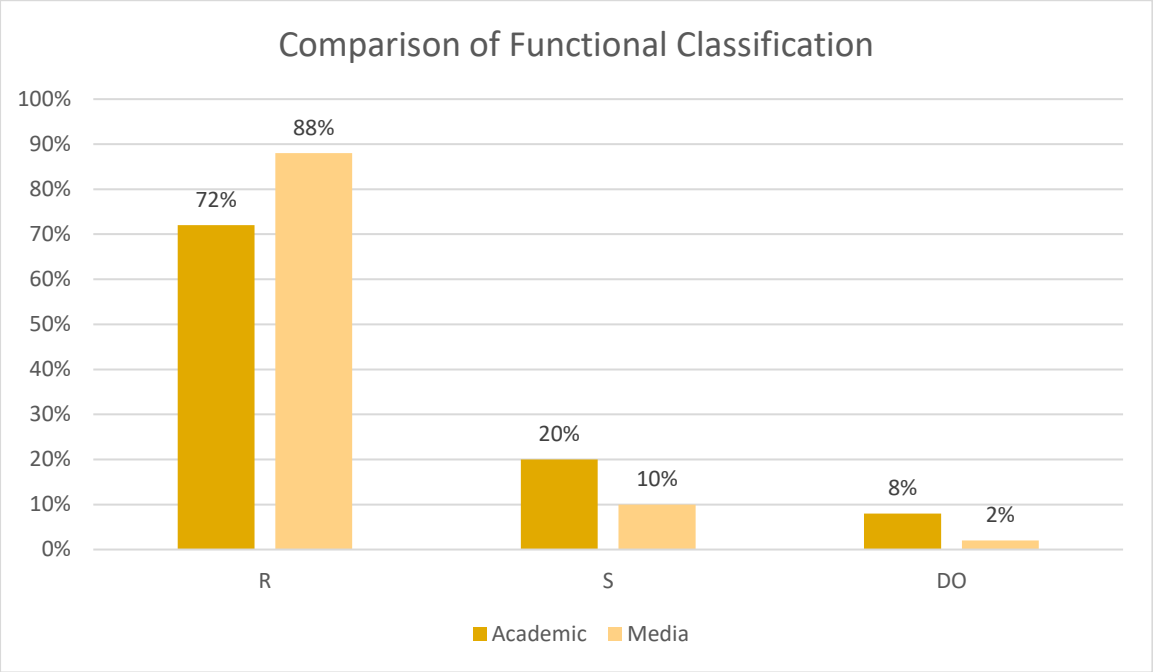

Figure 15. The comparison of functional classification of lexical bundles in the two registers.

Firstly, the slight differences of frequency of referential can be seen in the vertical bar graph. As stated previously, the media register is considered to contain more NP-based bundles, thus the referential function frequency percentage grows simultaneously. Stance function is twice more frequent in the academic register than in the media. The interesting side of this classification is the distribution of the LBs that function as discourse organisers. DO bundles are four times more frequent in the academic register than in the media. It can be suggested that in the analysed media register the bundles that function as discourse organisers are not greatly used because the online article structure does not require a firm complex structure, whereas, for the structure of the academic research paper it is inevitable; in other words, the media register texts presumably are looser in the organization than the academic register texts.

Moreover, it is important to be aware of the size and type of the audience which is noted as one of the differences between media and academic registers; media / news register focus on a wide-public, non-specialist audience, whereas academic register targets more specialist audiences (Biber et al. 1999: 16). In addition to this, these analysed media articles usually contain informative texts about general knowledge of astrophysical components, objects, events, such as telescopes, solar system, milky way galaxy, universe, black holes, light years,

the Big Bang etc. For example, in Corpus$_{MEDIA}$ roughly 9 per cent of all lexical bundles is lexical bundles that contain *black hole*, for instance *of a black hole, of the black hole, a supermassive black hole*. Another example which takes approximately 6 per cent of all instances highlights lexical bundles that contain universe, for instance, *in the early universe, when the universe was, of the early universe*. Thereby, presumably these topic-specific word clusters are highly generated and then they are repeated over and over in that type of articles, thus the particularly high percentage of the referential LBs is evident in the media register.

It is needed to take a closer look at the sub-classes, to observe the functional classification differences and similarities between the registers. In Figure 16 the first comparison of referential LBs sub-classes is given.

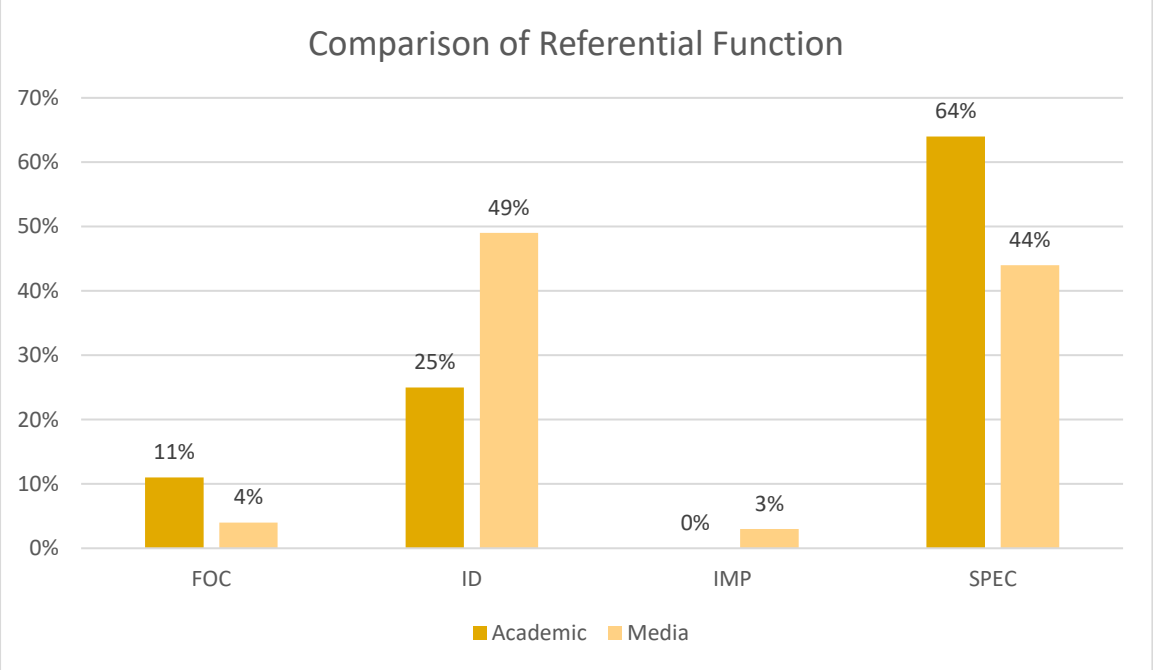

Figure 16. The comparison of referential lexical bundles in the two registers.

The prime difference in accordance with the frequency percentages is the occurrences of the sub-classes of focus (FOC) and imprecision (IMP). In the academic register focus-function LBs are almost three times more occurrent than in the media; additionally, imprecision-function LBs were not identified in the academic corpus, whereas in the media they are apparent, e.g., *about # percent of, are some of the*. This indicates that the centrality and precision is claimed more often in the academic paper rather than in the online article, for instance focus bundles:

(11) a) ***We note that the*** *destruction that we see is somewhat less extreme than in Garrison-Kimmel et al. (2017b)* <…>

(11) b) *We plot the ionizing luminosity* ***with respect to the*** *superorbital phase for each observation in Fig. 8.* (Corpus$_{ACAD}$)

(12) *While previous research has suggested that small amounts of hydrogen peroxide and other oxidants can be formed by stressing or crushing of rocks in the absence of oxygen,* ***this is the first*** *study to show the vital importance of hot temperatures in maximizing hydrogen peroxide generation.* (Corpus$_{MEDIA}$)

imprecision bundle:

(13) *As the most luminous kilonova event on record, the find was already groundbreaking, resulting in* ***some of the most*** *detailed observations to date*. (Corpus$_{MEDIA}$)

Furthermore, the sub-function of identification (ID) is more preferable in the media register rather than in the academic. Usually the 'identified' things are the names of institutions, descriptions of scientists or naming of the cosmic objects, for example:

(14) a) *Aaron Parsons, a radio astronomer at* ***the University of California****, Berkeley, who wasn't involved in either experiment* <…>

(14) b) *Other stars came from small "dwarf" galaxies that slammed into* ***the Milky Way and*** *aligned with an emerging disk.* (Corpus$_{MEDIA}$)

Another fascinating difference can be observed through the manifestation of LBs that function as specifications of attributes (SPEC). The two registers show a preference for specific specifications of attributes. For instance, in academic corpus, we could find salient LBs that specify intangible framing (example (15)), text deixis (example (16)), in contrast, in the media corpus salient LBs would be those that specify place/time deixis (example (17)), tangible framing (example (18)), even the quantity attributes and multifunctional referencing (example (19)).

(15) a) <…> *we study each outlier to understand* ***the origin of the*** *discrepancy.*

(15) b) *Each class here is an encoded number which changes* ***the nature of the*** *problem from classification to regression.* (Corpus$_{ACAD}$)

(16) a) *A vector point diagram for these stars* ***is shown in Fig****. 1.*

(16) b) *Derivatives are calculated directly and smoothed* ***as described in section*** *3 <…>* (Corpus$_{ACAD}$)

(17) a) *That groundbreaking snapshot featured the* ***supermassive black hole at*** *the center of M87, a massive elliptical galaxy 55 million light-years from Earth.*

(17) b) *But after the gas begins to evaporate, a few* ***million years after the*** *star's birth, the balance changes.*

Instances of these LBs (SPEC) must be interpreted with caution because sometimes one LB can have more than one function, for example it can express specification of time:

(17) c)1) *It was* ***the middle of the*** *night in a very isolated Utah location <…>*

or place:

(17) c)2) *<…> it leaves behind a black hole that sinks to* ***the middle of the*** *star cluster <…>* (Corpus$_{MEDIA}$)

(18) *In this case,* ***the size of the*** *asteroid is more plausible*. (Corpus$_{MEDIA}$)

(19) a) *The resulting friction heated it to* ***tens of thousands of*** *degrees Fahrenheit.*

(19) b) *<…> is set to go into operation by* ***the end of the*** *decade.* (Corpus$_{MEDIA}$)

In Figure 17 the comparison of sub-classes of stance LBs is provided.

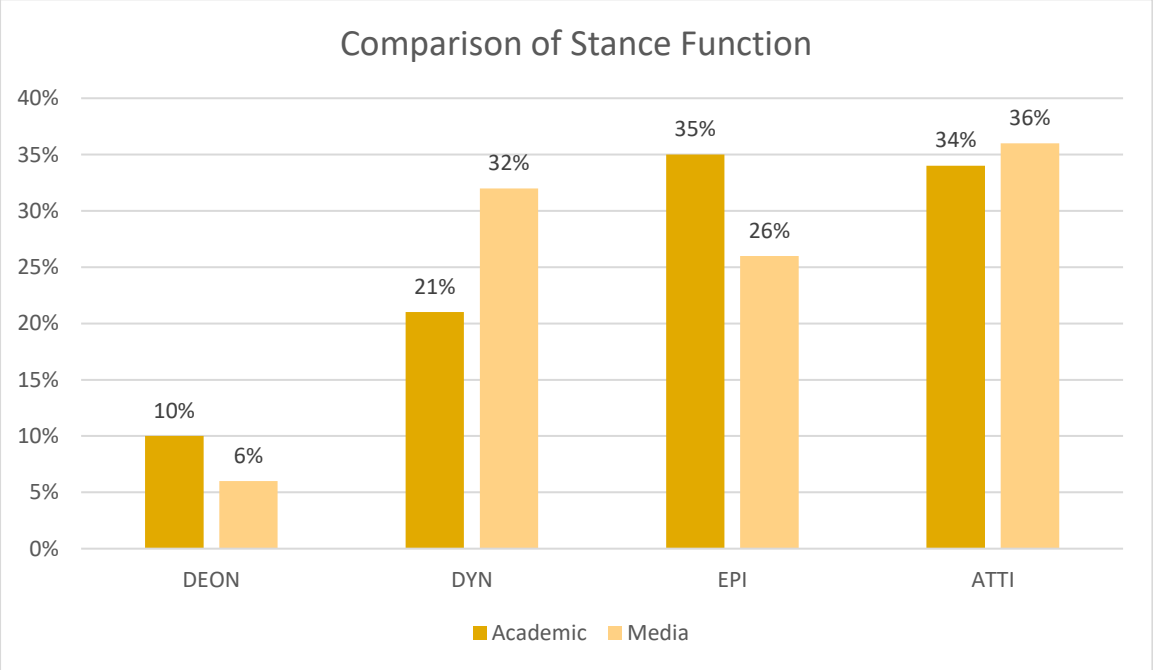

Figure 17. The comparison of stance function of lexical bundles in the two registers.

Firstly, the sub-class of dynamic modality (DYN) markers is approximately two times more occurrent in the media register rather than in the academic. Through the dynamic modality the ability of a subject is expressed. The only difference is that in the academic corpus theses LBs are impersonal while in the media corpus they are mostly personal, for instance:

(20) a) *Smaller scales in the velocity field* ***can be seen in*** *the bottom panel* <…>

(20) b) *This ambient state* ***can be used to*** *create superadiabatic entropy gradients that simulate heat flux from the bottom boundary.* (Corpus$_{ACAD}$)

(21) a) *We* ***will be able to*** *see the history of things take place during reionization.*

(21) b) *This time, though,* ***astronomers were able to*** *compare both the star and the light from its supernova blast to the expected profile of an electron-capture supernova.* (Corpus$_{MEDIA}$)

Secondly, through the epistemic modality a certain degree of subject's knowledge towards a proposition is expressed. LBs that function as epistemic modality (EPI) markers were more frequent in the academic register. In addition to this, both, high probability and high certainty, are evident in the academic corpus, whereas in the media corpus the certainty degree is omitted:

(22) a) *Thus,* ***we assume that the*** *priors are independent of each other* <…>

(22) b) *From these figures,* ***it is clear that*** *precise tidal deformability measurements will contribute substantially to* <…> (Corpus$_{ACAD}$)

(23) a) *Another issue is that* ***we do not know*** *enough about gamma-ray bursts* <…>

(23) b) *The level of disagreement is enough to make people uncomfortable, but* ***I think it is*** *far from the end of the story* <…> (Corpus$_{MEDIA}$)

Furthermore, deontic modality (DEON) LBs were more apparent in the academic register. Through the deontic modality obligation and permission is indicated. In the academic corpus usually a weak obligation or a directive can be found; and, in the media corpus mostly directives are common.

(24) ***It should be noted that*** *masses of the stellar components are hard to estimate* <…> (Corpus$_{ACAD}$)

(25) ***It is important to*** *establish observationally whether this gap is real, or whether it's an observational artifact* <…> (Corpus$_{MEDIA}$)

Lastly, in both registers the similarity of the usage of LBs that function as attitudinal stance (ATTI) markers can be observed. Through the attitudinal stance attitudes, feelings and judgements of a subject are conveyed. In the media register more evaluative, feelings-based LBs are occurring, while in the academic register the majority of LBs are expressing prediction.

(26) a) ***It is hard to*** *find locations where other things are exposed*.

(26) b) *It's* ***one of the biggest*** *mysteries of black hole research*. (Corpus$_{MEDIA}$)

(27) *This scenario* ***is expected to be*** *more frequent*. (Corpus$_{ACAD}$)

In Figure 18 the comparison of sub-classes of discourse organising LBs is presented.

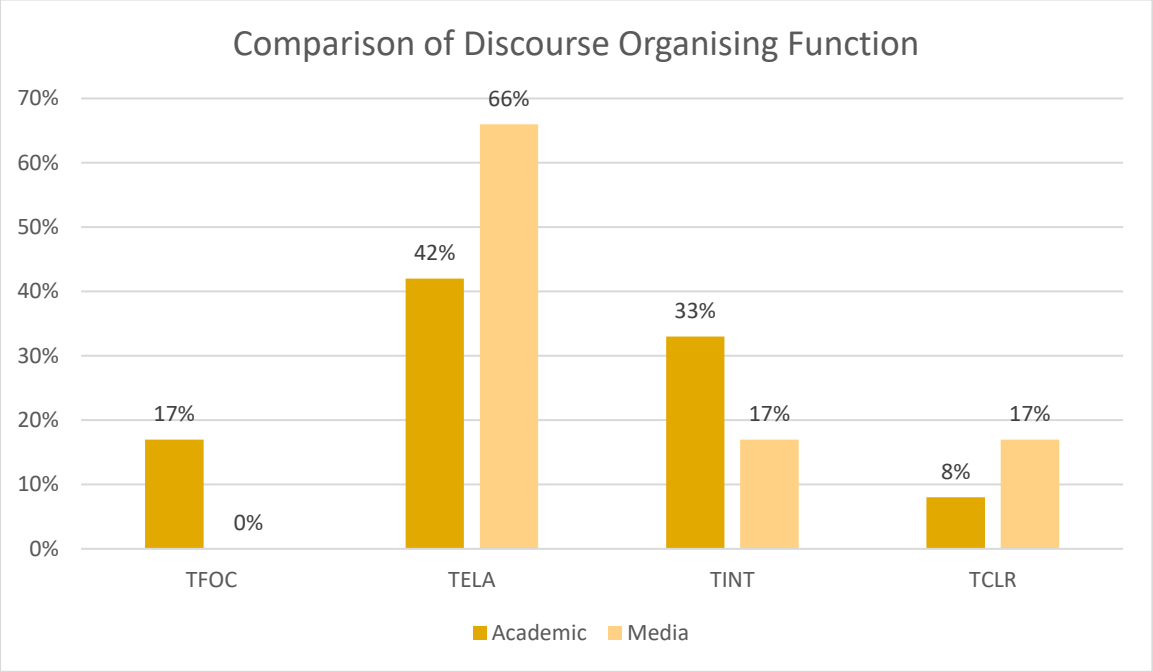

Figure 18. The comparison of discourse organising function of lexical bundles in the two registers.

The distribution of the last major functional sub-classes shows some significant differences as well. Albeit the most frequent usage in both registers, the sub-class of topic elaboration (TELA) is more preferable in the media register. And it can be assumed that such distribution is plausible because in the media articles there is more items that are supposed to be explained thereby the topic elaboration bundles are proper for that purpose.

(28) *Though mysteries abound* ***when it comes to*** *neutron stars, astronomy is entering a golden era* (Corpus$_{MEDIA}$)

(29) *Linearly polarized Alfvén waves, **on the other hand**, have a time-varying magnetic field strength with an associated pressure anisotropy* <…> (Corpus$_{ACAD}$)

In addition to this, the topic clarification (TCLR) LBs seem to appear more in the media register than in the academic. Similarly, to TELA bundles, these LBs also provide additional information, although the clarification function aims more on adding an explanation which facilitates the understanding whereas elaboration function only provides supplementary block of information.

(30) *And, **as it turns out**, the discovery was not a fluke*. (Corpus$_{MEDIA}$)

(31) *The circular velocity of the gas has been initialized **taking into account the** effect of a radial pressure gradient due to the initial radial density profile.* (Corpus$_{ACAD}$)

Topic introduction (TINT) and focus (TFOC) sub-classes are more prominent in the academic register. As mentioned previously it is argued that the academic paper structure is more rigid than the structure of the media article. Examples of topic introduction and focus bundles are presented in (32)a), (33) and (32)b) respectively:

(32) a) ***In this work, we** consider the latest version of the HMCODE software*.

(32) b) ***In the following**, we briefly discuss relevant factors regarding MG membership status for each star* <…> (Corpus$_{ACAD}$)

(33) ***For the first time**, we were measuring dark matter from almost the earliest moments of the universe* (Corpus$_{MEDIA}$)

4. Conclusions

The purpose of the study was to investigate and compare academic and media publications on astrophysics by applying the lexical bundles approach. The aim of the study was to identify differences and similarities of structural and functional characteristics of LBs that are found in the two registers of English. Therefore, the analysis relied on theoretical and methodological approaches, including frequency-driven approach, lexical bundles approach (Biber et al. 2004).

The results of the analysis reveal that the media and academic registers tend to have more differences than similarities both structurally and functionally. The structural variation between the registers in general shows the tendency that the NP-based LBs are more frequent in the media register; the VP-based LBs are twice more occurrent in the academic register; and, the DC-based LBs are slightly more prominent in the academic register.

The foremost differences between the registers in terms of the NP-based LBs are the *noun phrase + at-phrase* structure which was predominantly found only in the media register. Moreover, structures of *other noun expression* and *noun phrase + post-modifier* were significantly more occurring in the media register as well. It can be assumed that the academic register contains more prepositional phrases, in particular of-phrases. However, the usage of the *comparative expressions* was similar in both registers.

Furthermore, two main differences were identified while comparing the VP-based LBs. Firstly, the structure of *verb phrase with a passive verb* is approximately as twice more frequent in the academic register; secondly, the usage of *3rd person pronoun + verb phrase* structure is roughly two times more frequent in the media register than in the academic register. In academic register the 3rd person pronoun usually functions as a dummy *it* that fills subject's syntactical place and, in the media, register it mostly functions as a referable subject.

The comparison of DC-based LBs has shown a clear difference between the preferred sub-structures of LBs in the two registers. In the media register LBs that incorporate *WH-clause fragment*, *verb + to-clause* and *bare to-clause fragments* prevail, whereas, in the academic register *that-clause fragments*, bare and with 1st person, are more apparent. The sub-structure of *1st person + that-clause* is found nearly exclusively in the academic register.

Regarding the functional variation between the two registers, it was found that the media register contains more referential LBs; whereas, in the academic register stance and discourse-organising LBs were approximately two and four times more frequently occurring respectively.

In terms of the discourse organising function, this might be due to the suggestion that the media register texts presumably are looser in the organization than the academic register texts.

It was also found that referential LBs in the media register mostly function as *identification* and *specification of place/time deixis, quantity attributes* and *multifunctional referencing*, whereas this type LBs in the academic register would usually function as *specifications of intangible framing* and *text deixis*. Moreover, in the academic register *focus*-function LBs were almost three times more occurrent than in the media.

The stance function quite differs as well between the registers. In the media register, the *dynamic modality* expressions are mostly personal, whereas in the academic register they are impersonal. In terms of the *epistemic modality*, it was more usually apparent in the academic register. In addition to this, both, high probability and high certainty, were evident in the academic corpus, whereas in the media corpus the high certainty degree is omitted. The most noticeable difference is that in the media register the *attitudinal stance* is essentially expressing evaluation, feelings, whereas in the academic register - prediction is commonly expressed.

In the case of discourse organising function it was revealed that the *topic clarification* and *topic elaboration* functions were more occurring in the media register. However, the function of *topic introduction* and *focus* were more prominent in the academic register.

In addition to the presented findings, it is important to be aware of the size and type of the audience which is identified as one of the most important situational variable differences between media and academic registers. Media / news register focus on a wide-public, non-specialist audience, whereas academic register targets more specialist audiences. It can be concluded that the main difference between the two analysed registers is that academic papers on astrophysics follow a more rigid structure which automatically enforces a number of expressions to create the coherence and cohesion, to avoid subjectivity or misleading statements. In contrast, media articles on astrophysics have a looser structure which allows the authors to simply narrate the story without necessarily following rigid rules or style conventions as long as the story serves its purpose - to narrate, inform about what has been discovered and summarize information. In the research papers, authors usually are examining one or another case in the investigative field and such type of structure and function is convenient to use; whereas, in media articles authors do not report any scientific findings or procedures. They provide and share information about specific innovations or events by

mentioning and referencing scientific research papers together with their own evaluation about it.

Hopefully this study serves as a modest contribution to the ever-growing field of register analysis as it has focused on analysing media and academic registers. With this contrastive analysis it was endeavoured to determine the typical structural and functional features of the media article on astrophysics by comparing it with the extensively analysed academic register.

Without doubt, however, this study has its limitations such as a relatively small data sample, focus only on the English language and astrophysics subject. Future research in this field could be carried out by focusing on more or other languages to detect potential cross-linguistic specificity of language varieties. More comprehensive analyses of register would provide a better understanding of the language itself and the content that is produced by it in the certain contexts. Because every register has its own typical features, whether it is an academic research article or a media article on astrophysics or an abstract of research paper on medicine, the language that is used in that context and is shaped by the specific context (Biber and Conrad (2009) situational variables) flourishes into something distinctly unique which can be observed by the linguists.

5. References

1. Ädel, A. & B. Erman. 2012. Recurrent Word Combinations in Academic Writing by Native and Non-native Speakers of English: A Lexical Bundles Approach. *English for Specific Purposes*. Vol. 31, 81–82.
2. Anthony, L. 2023. *AntConc* (Version 4.2.0) [Computer Software].
3. Barlow, M. 2011. Corpus Linguistics and Theoretical Linguistics. *International Journal of Corpus Linguistics*. Vol. 16, 3–44.
4. Biber, D. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers.* Amsterdam: John Benjamins Publishing Company.
5. Biber, D., Johansson, S., Leech, G., Conrad, S. & E, Finegan. 1999. *Longman Grammar of Spoken and Written English.* Harlow, England: Longman.
6. Biber, D., Conrad, S. M. & Cortes, V. 2004. *'If you look at …': Lexical Bundles in University Teaching and Textbooks*. Applied Linguistics 25, 371–405.
7. Biber, D. & S. M. Conrad. 2005. The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. *Applied Linguistics Faculty Publications and Presentations.*
8. Biber, D. & F. Barbieri. 2007. Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes.* Vol. 26, 263–286. Elsevier.
9. Biber, D. & S. Conrad. 2009. *Register, Genre and Style.* New York: Cambridge University Press.
10. Biber, D. & B. Gray. 2016. *Grammatical Complexity in Academic English.* United Kingdom: Cambridge University Press.
11. Biber, D. & J. Egbert. 2018. *Register Variation Online.* United Kingdom & New York: Cambridge University Press.
12. Byrd, P. & A. Coxhead. 2010. *On the other hand:* Lexical Bundles in Academic Writing and in the Teaching of EAP. *University of Sydney Papers in TESOL*. Vol. 5.
13. Cambridge Dictionary. 2023. Accessed on 16 January 2023, available from: https://dictionary.cambridge.org/.
14. Cao, F. 2021. A Comparative Study of Lexical Bundles Across Paradigms and Disciplines. *Corpora 2021*. Vol. 16, 97–128.
15. Chen, Y. & P. Baker. 2010. Lexical Bundles in L1 and L2 Academic Writing. *Language Learning & Technology*. Vol. 14, No. 2, 30–49.

16. Chen, Y. & P. Baker. 2014. Investigating Criterial Discourse Features across Second Language Development: Lexical Bundles in Rated Learner Essays, CEFR B1, B2 and C1. *Applied Linguistics.* 849–880.
17. Chi-square Test Calculator. (2023, May 20). Retrieved from https://www.socscistatistics.com/tests/chisquare2/default2.aspx.
18. Connor, U. M. & A. I. Moreno. (2005). Tertium Comparationis: A Vital Component in Contrastive Research Methodology. In, *Directions in Applied Linguistics: Essays in Honor of Robert B. Kaplan*, Clevedon, Bruthiaux P., Atkinson D., Eggington W. G., Grabe W., & V. Ramanathan (eds), pp. 153–164. England: Multilingual Matters.
19. Cortes, V. 2004. Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for Specific Purposes*. Vol. 24, 397–423.
20. Cortes, V. 2015. Situating Lexical Bundles in the Formulaic Language Spectrum. Origins and Functional Analysis Developments, in: Cortes, V. & Csomay, E. (Eds) *Corpus-based Research in Applied Linguistics: Studies in Honor of Doug Biber*. Amsterdam: John Benjamins Publishing Company, pp. 197–216.
21. Crawford, W. J. & E. Cosmay. 2016. *Doing Corpus linguistics,* New York and London: Routledge.
22. Crystal, D. 1964. A Liturgical Language in a Linguistic Perspective. *New Blackfriars.* Vol. 46. No. 534. 148–156.
23. Dontcheva-Navratilova, O. 2012. Lexical Bundles in Academic Texts by Non-native Speakers. *Brno Studies in English*. Vol. 38, No. 2.
24. Farvardin, M. T., Afghari, A. & M. Koosha. 2012. Analysis of Four-Word Lexical Bundles in Physics Research Articles. *Advances in Digital Multimedia.* Vol. 1, No. 3.
25. Ferguson, C. A. 1983. Sports Announcer Talk: Syntactic Aspects of Register Variation. *Language in Society.* Vol. 12, No. 2. 153–172.
26. Grabowski, L. 2013. Register Variation across English Pharmaceutical Texts: A Corpus Driven Study of Keywords, Lexical Bundles and Phrase Frames in Patient Information Leaflets and Summaries of Product Characteristics. *Social and Behavioral Sciences.* Vol. 95, 391 – 401.
27. Grabowski, L. 2015. Keywords and Lexical Bundles Within English Pharmaceutical Discourse: A Corpus-driven Description. English for Specific Purpose. *English for Specific Purposes.* Vol. 38, 23–33.
28. Güngör, F. & H. H. Uysal. 2016. A Comparative Analysis of Lexical Bundles Used by Native and Non-native Scholars. *English Language Teaching.* Vol. 9, No. 6.

29. Güngör, F. & H. H. Uysal. 2020. Lexical Bundle Use and Crosslinguistic Influence in Academic Texts. *Lingua*. 242.
30. Halliday, M. A. K. & R. Hasan. 1976. *Cohesion in English.* London and New York: Routledge.
31. Heller, V. & M. Morek. 2015. Academic Discourse as Situated Practice. *Linguistics and Education.* Vol. 31. 174–186.
32. Herbel-Eisenmann, B., Wagner, D. & V. Cortes. 2010. Lexical Bundle Analysis in Mathematics Classroom Discourse: the Significance of Stance. *Educ Stud Math.* Vol. 75, 23–42.
33. Hernández, P. S. 2013. Lexical Bundles in Three Oral Corpora of University Students. *Nordic Journal of English Studies*. Vol. 13, 187–209.
34. Hong, A. L. & T. K. Hua. 2015. Specificity in English for Academic Purposes (EAP): A Corpus Analysis of Lexical Bundles in Academic Writing. *The Southeast Asian Journal of English Language Studies.* Vol. 24, 82 – 94.
35. Hyland, K. 2008. As can be seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes*. Vol. 27, 4–21.
36. Hyland, K. & P. Tse. 2009. Academic Lexis and Disciplinary Practice: Corpus Evidence for Specificity. *International Journal of English Studies.* Vol. 9, 111–129.
37. Hussain, G., Zahra, T. & A. Abbas. 2021. Discourse Functions of Lexical Bundles in Pakistani Chemistry and Physics Textbooks. Journal of Language Studies. *Journal of Language Studies*. Vol 21.
38. Leech, G. 1992. Corpora and Theories of Linguistic Performance. In, *Trends in Linguistics Studies and Monographs 83,* Winter, W. (ed). Berlin & New York: Mouton de Gruyter.
39. Li, X. & L., Li. 2015. Characteristics of English for Science and Technology. *International Conference on Humanities and Social Science Research*. Atlantis Press.
40. Nesi, H. & H. Basturkmen. 2015. Lexical bundles and discourse signaling in academic lecturers. *International Journal of Corpus Linguistics.* Vol. 11, 283–304.
41. Noor, M. & B. Anwar. 2020. Templates Developed for Writing Introduction Section for Research Scholars: Lexical Bundle Approach in Post-Covid-19 Situation. *Pakistan Journal of Social Sciences*. Vol. 40, No. 3, 1299–1315.
42. Parkinson, J. 2013. English for Science and Technology. In, *The Handbook of English for Specific Purposes,* Paltridge, B. & S. Starfield (eds). Boston: John Wiley& Sons.

43. Reid, Thomas Bertram. 1956. *Linguistics, Structuralism and Philology*. Archivum Linguisticum 8. 28–37.
44. Rujirawan, L. M. 2021. *Structural and Functional Analyses of Four-Word Lexical Bundles from Articles in the Forbes Website*. Master's thesis. Language Institute Thammasat University.
45. Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Oxford, England: Oxford University Press.
46. Tarone, E., Dwyer, S., Gillete, S. &V. Icke. 1981. On the Use of the Passive in Two Astrophysics Journal Papers. *The ESP Journal.* Vol. 1, No. 2.
47. Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work.* Amsterdam / Philadelphia: John Benjamins Publishing Company.
48. Vilkaitė, L. 2016. Formulaic language is not all the same: comparing the frequency of idiomatic phrases, collocations, lexical bundles, and phrasal verbs. *Taikomoji Kalbotyra* 2016 (8).
49. Yin, X. & S. Li. 2021. Lexical Bundles as an Intradisciplinary and Interdisciplinary Mark: A Corpus-based Study of Research Articles from Business, Biology, and Applied linguistics. *Applied Corpus Linguistics.* Vol. 1.
50. Wray, A. & M. R. Perkins. 2000. The Functions of Formulaic Language: an Integrated Model. *Language & Communication.* Vol. 20, 1–28.
51. Xu, M. & T. Sun. 2022. Structural Analysis of Lexical Bundles in English News on Health from China and UK Newspapers. *International Sociology, Economics, Education and Humanities Conference.*

6. Summary in Lithuanian

Tyrime analizuojamas akademinių ir žiniasklaidos straipsnių apie astrofiziką kalbinės raiškos sudėtingumas. Pagrindiniai šio tyrimo uždaviniai buvo nustatyti ir išnagrinėti registrų skirtumus pagal struktūrinius ir funkcinius pasikartojančių žodžių sekų, tai yra leksinių samplaikų, požymius. Atsižvelgiant į tyrimo tikslus, buvo sudaryti du akademinių ir žiniasklaidos straipsnių apie astrofiziką tekstynai. Tyrimas buvo atliktas taikant leksinių samplaikų tyrimo prieigą (Biber et al. 2004), kuri apibrėžia leksinių samplaikų generavimo iš tekstyno ir analizės būdus.

Tyrimo rezultatai atskleidė, kad žiniasklaidos ir akademiniai registrai tiek struktūriškai, tiek funkciškai turi daugiau skirtumų nei panašumų. Struktūriniai registrų skirtumai apskritai rodo tendenciją, kad žiniasklaidos registre dažniau pasitaiko daiktavardinės leksinės samplaikos; akademiniame registre dvigubai dažniau pasitaiko veiksmažodinių leksinių samplaikų; akademiniame registre šiek tiek dažniau pasitaiko šalutinio sakinio leksinių samplaikų. Kalbant apie funkcinius registrų skirtumus, nustatyta, kad: žiniasklaidos registre daugiau referentinių leksinių samplaikų; o akademiniame registre autoriaus požiūrio raiškos ir diskursą organizuojančių leksinių smaplaikų pasitaiko maždaug du ir keturis kartus dažniau atitinkamai. Be to, kiekviename analizuotame registre nustatyta po vieną vyraujančią išskirtinę struktūrą. Akademiniame registre tai buvo 1-ojo asmens + *that*-šalutinio sakinio fragmentas, o žiniasklaidos registre – daiktavardinis junginys + *at*-prielinksninis junginys.

Be pateiktų išvadų, svarbu atkreipti dėmesį į auditorijos dydį ir tipą, kuris pažymėtinas kaip vienas iš svarbiausių žiniasklaidos ir akademinių registrų skirtumų; žiniasklaidos / naujienų registras orientuojasi į plačią, nespecializuotą auditoriją, o akademinis registras - į labiau specializuotą skaitytoją. Galima daryti išvadą, kad pagrindinis skirtumas tarp abiejų analizuotų registrų yra tas, kad akademinis straipsnis apie astrofiziką turi griežtesnę struktūrą, kuri iš dalies paaiškina, kodėl čia nustatyta daugiau diskursą organizuojančių samplaikų, tuo tarpu žiniasklaidos straipsniai apie astrofiziką turi laisvesnę struktūrą, kuri leidžia kalbai tiesiog pasakoti kaip istoriją be jokių griežtų apribojimų, jei tik ji tarnauja savo tikslui - pasakoti, informuoti apie tai, kas atrasta, ir apibendrinti informaciją. Moksliniuose straipsniuose autoriai paprastai nagrinėja vieną ar kitą tiriamąjį atvejį, todėl tokio tipo struktūrą ir funkciją patogu naudoti; tuo tarpu žiniasklaidos straipsniuose autoriai mokslinių tyrimų neaptaria. Jie pateikia ir dalijasi informacija apie konkrečias naujoves ar įvykius, paminėdami ir nurodydami mokslinių tyrimų darbus kartu su savo vertinimu apie tai.

7. Appendices

7.1. The additional information about the sources from which the data was collected

	Link	Publisher / Eds
Astronomy Magazine	https://astronomy.com/	US: David Eicher, Kalmbach Publishing
Quanta Magazine	https://www.quantamagazine.org/	US: Thomas Lin, Simons Foundation
The Guardian	https://www.theguardian.com/uk?INTCMP=CE_UK	UK: Katharine Viner: The Observer
The astrophysical Journal	https://iopscience.iop.org/journal/0004-637X	UK: Oxford University Press
The Astrophysical Journal Supplement	https://iopscience.iop.org/journal/0067-0049	US: American Astronomical Society (IOP Publishing)
The Monthly Notices of the Royal Astronomical Society	https://academic.oup.com/mnras	US: American Astronomical Society (IOP Publishing)

In addition, the SCImago Journal Rank (SJR) indicator is a measure of the scientific influence of scholarly journals that accounts for both the number of citations received by a journal and the importance or prestige of the journals where the citations come from.

Q1 to Q4 refer to journal ranking quartiles within a subdiscipline using the SJR citation index. Thus, a first quartile journal (i.e., Q1) has an SJR in the top 25% of journals for at least one of its classified subdisciplines. The academic journals that were chosen for the corpora compilation are in the first rank Q1.

7.2. Full lists of the analysed LBs in the academic and media registers

NO	Academic	NormF	Media	NormF
1	as a function of	447	at the university of	431
2	in the case of	151	after the big bang	267
3	with respect to the	138	of the milky way	263
4	a function of the	120	at the center of	231
5	on the other hand	116	in the early universe	164
6	by a factor of	113	james webb space telescope	156
7	the line of sight	112	an astronomer at the	152
8	as well as the	103	the hubble space telescope	148
9	per cent of the	98	times the mass of	136
10	in this paper we	95	the end of the	124
11	we find that the	94	the mass of the	116
12	in this work we	90	of a black hole	116
13	of the magnetic field	81	an astrophysicist at the	112
14	is consistent with the	79	of the black hole	112
15	in this section we	75	the cosmic microwave background	112
16	we note that the	71	the event horizon telescope	112
17	the evolution of the	68	in the astrophysical journal	108
18	the presence of a	65	of the solar system	108
19	is shown in fig	64	one of the most	104
20	the fact that the	64	a supermassive black hole	104
21	are shown in fig	56	black hole at the	104

22	the mass of the	55	astronomer at the university	96
23	in the context of	54	mass of the sun	96
24	the effect of the	54	in the milky way	96
25	the position of the	54	center of the milky	96
26	the size of the	53	light years from earth	92
27	an order of magnitude	51	of the university of	92
28	in this case the	50	hole at the center	92
29	the shape of the	50	million years after the	92
30	of the order of	48	for the first time	88
31	the total number of	48	of gas and dust	88
32	on the order of	47	the european space agency	84
33	the origin of the	47	in the solar system	80
34	the properties of the	47	the speed of light	80
35	we assume that the	46	the large magellanic cloud	80
36	the location of the	45	when the universe was	80
37	along the line of	45	over the course of	76
38	as shown in fig	45	the university of California	76
39	can be used to	45	goddard space flight center	76
40	in the absence of	44	the size of a	76
41	is due to the	44	at the heart of	72
42	can be found in	43	of the early universe	72
43	in terms of the	43	into a black hole	72
44	of the x ray	43	the center of our	72
45	are listed in table	42	million light years away	68
46	in addition to the	42	at the end of	68

47	the value of the	42	the supermassive black hole	68
48	to that of the	41	the milky way and	68
49	the ratio of the	40	in our solar system	64
50	in the presence of	40	it is hard to	64
51	the magnetic field is	40	the university of Arizona	64
52	in the vicinity of	39	of our solar system	60
53	shown in fig the	39	end of the universe	60
54	the amplitude of the	39	some of the most	56
55	the magnetic field strength	39	at the same time	56
56	a function of time	37	expansion of the universe	56
57	in the form of	37	supermassive black hole at	56
58	to account for the	37	the fabric of space	52
59	as a result of	37	theory of general relativity	52
60	can be seen in	37	times more massive than	52
61	the difference between the	37	about light years away	52
62	are consistent with the	36	is one of the	52
63	in the range of	36	mass of our sun	52
64	the results of the	36	we do not know	52
65	it is important to	35	of the big bang	52
66	the center of the	35	the big bang the	52
67	the majority of the	35	the black hole at	52
68	as can be seen	34	of the universe is	52
69	at the end of	34	a member of the	48
70	for each of the	34	fabric of space time	48
71	in fig we show	34	of the universe the	48
72	are shown in figure	33	on the other hand	44
73	the distribution of the	33	the size of the	44
74	to the magnetic field	33	billion light years away	44
75	we found that the	33	end of its life	44

76	the end of the	33	the rest of the	44
77	the nature of the	33	in the night sky	44
78	the peak of the	33	matter in the universe	44
79	as described in section	32	our understanding of the	44
80	the surface of the	32	some light years away	44
81	in agreement with the	32	university of california Berkeley	44
82	it is clear that	32	cosmologist at the university	44
83	the case of the	32	the nobel prize in	44
84	the national science foundation	32	of the sun and	40
85	the parameters of the	32	our own milky way	40
86	we see that the	32	our own solar system	40
87	a wide range of	31	the center of a	40
88	has made use of	31	who was not involved	40
89	is the number of	31	to learn more about	40
90	at the same time	31	max planck institute for	40
91	the speed of light	31	a black hole and	40
92	to the fact that	31	a theoretical astrophysicist at	36
93	is assumed to be	30	are some of the	36
94	that there is a	30	california institute of technology	36
95	the direction of the	30	like the milky way	36
96	the structure of the	30	some of the first	36
97	a result of the	29	the max planck institute	36
98	it is possible that	29	the next few years	36
99	shown in figure the	29	as well as the	36
100	the magnitude of the	29	at the centers of	36
101	the presence of the	29	average earth sun distance	36
102	is shown in the	28	of supermassive black holes	36
103	similar to that of	28	the history of the	36

104	taking into account the	28	the milky way galaxy	36
105	the anonymous referee for	28	times stronger than earth	36
106	the case of a	28	a cosmologist at the	36
107	the other hand the	28	a few million years	36
108	are shown in the	28	from the big bang	36
109	in the x ray	28	image of a black	36
110	is shown in figure	28	the milky way is	36
111	it should be noted	28	the shape of the	36
112	which is consistent with	28	the university of Chicago	36
113	at a distance of	27	thousands of light years	36
114	for the first time	27	university of california santa	36
115	in section we describe	27	what is going on	36
116	in the following we	27	in the middle of	36
117	is given by the	27	the inner solar system	36
118	should be noted that	27	scientist at the university	36
119	the formation of the	27	the black hole in	36
120	the strength of the	27	the dark energy survey	36
121	this is consistent with	27	at a distance of	32
122	is based on the	27	have been able to	32
123	is expected to be	27	in a paper published	32
124	the rest of the	27	in the coming years	32
125	the width of the	27	in the form of	32
126	a large number of	26	in this case the	32
127	as a result the	26	is not the only	32
128	in good agreement with	26	lead author of the	32
129	is related to the	26	located about # light	24
130	is the same as	26	more massive than the	32
131	it is possible to	26	objects in the universe	32
132	the same as in	26	of a massive star	32
133	we show that the	26	smithsonian center for astrophysics	32

134	the slope of the	26	the sun and the	32
135	are the same as	25	the universe is expanding	32
136	can be explained by	24	to the end of	32
137	is in agreement with	24	was not involved in	32
138	thank the anonymous referee	24	and a member of	32
139	the velocity of the	24	billion years after the	32
140	this means that the	24	light years away the	32
141	a consequence of the	24	may be able to	32
142	can be written as	24	of the universe and	32
143	is operated by the	24	one of the first	32
144	the impact of the	24	the astrophysical journal letters	32
145			the centers of galaxies	32
146			the milky way in	32
147			the supermassive black holes	32
148			the surface of the	32
149			understanding of the universe	32
150			we do not have	32
151			with the naked eye	32
152			a fraction of a	32
153			a physicist at the	32
154			black holes at the	32
155			space telescope science institute	32
156			the middle of the	32
157			times that of the	32
158			around a black hole	32
159			into the black hole	32

160	and the end of	28
161	at johns hopkins university	28
162	at the california institute	28
163	at the max planck	28
164	au is the average	28
165	beginning and the end	28
166	beginning to the end	28
167	gas and dust that	28
168	harvard smithsonian center for	28
169	in front of the	28
170	institute for radio astronomy	28
171	interferometer gravitational wave observatory	28
172	involved in the research	28
173	is the average earth	28
174	laser interferometer gravitational wave	28
175	massive than the sun	28
176	million light years from	28
177	of the universe but	28
178	one of the biggest	28
179	our solar system and	28
180	over the next few	28
181	stars in our galaxy	28
182	story comes from our	28
183	the beginning and the	28
184	the beginning to the	28

185	the formation of the	28
186	to figure out how	28
187	when it comes to	28
188	a black hole is	28
189	a theoretical physicist at	28
190	be able to see	28
191	black hole in the	28
192	by far the most	28
193	clouds of gas and	28
194	fraction of a second	28
195	from the black hole	28
196	in the northern hemisphere	28
197	it is possible that	28
198	known as the hubble	28
199	light years away in	28
200	might be able to	28
201	million to million years	28
202	of matter in the	28
203	the big bang it	28
204	the discovery of the	28
205	the international space station	28
206	the matter in the	28
207	version of this article	28
208	visible to the naked	28
209	a black hole it	28
210	a time when the	28
211	and the hubble heritage	28
212	as part of the	28
213	black holes in the	28
214	i do not think	28
215	nature of dark matter	28

216	physicist at the university	28
217	planetary scientist at the	28
218	stellar mass black hole	28
219	tens of thousands of	28
220	that of the sun	28
221	the first stars and	28
222	the nature of dark	28
223	the space telescope science	28
224	the surface of a	28
225	a black hole the	24
226	a neutron star or	24
227	a wide range of	24
228	about million light years	24
229	about # percent of	24
230	age of the universe	24
231	as it turns out	24
232	at just the right	24
233	by the end of	24
234	degrees fahrenheit degrees Celsius	24
235	do not know what	24
236	even light can escape	24
237	for thousands of years	24
238	from the center of	24
239	galaxies like the milky	24
240	galaxy million light years	24
241	hundreds of thousands of	24
242	i think it is	24
243	in the case of	24
244	in the fabric of	24

245	in the universe and	24
246	in this composite image	24
247	it is important to	24
248	lambda cold dark matter	24
249	light years away and	24
250	million years ago when	24
251	more about pulsars and	24
252	more massive than our	24
253	must see cosmic objects	24
254	of the supermassive black	24
255	of two neutron stars	24
256	our milky way galaxy	24
257	supermassive black holes are	24
258	supermassive black holes in	24
259	the age of the	24
260	the big bang this	24
261	the black hole is	24
262	the laser interferometer gravitational	24
263	the light of the	24
264	the remnant of a	24
265	the tip of the	24
266	the top of the	24
267	this is the first	24
268	to come up with	24
269	with the hubble space	24
270	would be able to	24
271	years away in the	24

272	a black hole that	24
273	and the university of	24
274	are a lot of	24
275	astronomers were able to	24
276	author of the study	24
277	be able to detect	24
278	can be used to	24
279	cloud of gas and	24
280	from the accretion disk	24
281	going to be a	24
282	history of the universe	24
283	horizon telescope eht collaboration	24
284	is known as the	24
285	of hydrogen and helium	24
286	of the universe in	24
287	on a collision course	24
288	on the surface of	24
289	some billion years ago	24
290	the big bang when	24
291	the black hole and	24
292	the core of a	24
293	the depths of space	24
294	the harvard smithsonian center	24
295	the location of the	24
296	the mass of a	24
297	the next generation of	24
298	the result of a	24
299	the university of Tokyo	24
300	the very large array	24

301	there are a few	24
302	wide field infrared survey	24
303	will be able to	24
304	x ray nasa cxc	24