ELUA
ISSN 2171-6692
Núm. 38, 2022, págs. 241-262
https://doi.org/10.14198/ELUA.22267

A corpus-based study of 4-grams in the research article genre

Estudio basado en corpus de 4-gramas en el artículo científico

Eva Lucía Jiménez-Navarro
Universidad de Córdoba, España
lucia.jimenez@uco.es
https://orcid.org/0000-0001-9377-6921

Abstract
The analysis of phraseology in the specialized discourse of science has sparked researchers' interest in the last few decades, probably because the use of word groupings in specific registers can provide information about certain typical features of the genre. For instance, Gledhill (2009) explores colligations of tenses in scientific articles and discovers that the present tense is used for qualitative and empirical expressions, while the past tense provides quantitative and research-oriented descriptions; Pérez-Llantada (2014) investigates 4-word lexical bundles in research articles, finding that these multiword combinations express referential meaning and organize the text; finally, Jiménez-Navarro (2019) analyzes adjective + noun collocations in a corpus of scientific papers and concludes that these phraseological units convey specific meanings when used in this genre, since they represent the contents of research articles.
The aim of the current study is to contribute to the analysis of 4-grams in the language of science. To this end, two specific objectives are defined: first,

Resumen
El análisis de la fraseología en el discurso especializado de la ciencia ha despertado el interés de los/as investigadores/as en las últimas décadas, probablemente porque el uso de grupos de palabras en registros específicos puede informar de algunas características típicas del género. Por ejemplo, Gledhill (2009) explora las coligaciones de tiempos verbales en artículos científicos y descubre que el tiempo presente se usa para expresiones cualitativas y empíricas, mientras que el tiempo pasado proporciona descripciones cuantitativas y orientadas a la investigación; Pérez-Llantada (2014) investiga grupos léxicos de cuatro palabras en artículos de investigación y descubre que estas combinaciones multilexémicas expresan significado referencial y organizan el texto; finalmente, Jiménez-Navarro (2019) analiza colocaciones de adjetivo + sustantivo en un corpus de artículos científicos y concluye que estas unidades fraseológicas aportan significados específicos cuando se usan en este género,

To cite this article: Jiménez-Navarro, E. L. (2022). A corpus-based study of 4-grams in the research article genre. *ELUA*, (38), 241-262. https://doi.org/10.14198/ELUA.22267

Recibido: 16/03/2022 Aceptado: 14/05/2022

to ascertain the structure of 4-grams; second, to analyze the function they perform. The methodology was based on a corpus and entailed five major steps: (1) a specialized corpus of research articles was built, (2) a list of 4-grams was automatically extracted using the software Sketch Engine, (3) the resulting list was manually verified in order to suppress inaccurate candidates, (4) the selected units were classified depending on their structural framework, and (5) the selected units were categorized according to their function in the text. The findings show that, in terms of the first objective, the most typical 4-grams were noun phrases; and as for the second objective, the sequences examined mostly concerned the research conducted and the authorship of the texts. All in all, the 4-grams identified were structures that were specific to the genre under study but could also be used in other domains.

KEYWORDS: specialized corpus; 4-gram; function; research article; structure.

puesto que representan los contenidos del artículo de investigación.

El objetivo de este estudio es contribuir al análisis de 4-gramas en el lenguaje de la ciencia. Para ello, se han definido dos objetivos específicos: en primer lugar, establecer la estructura de estas secuencias de palabras; en segundo lugar, analizar su función. La metodología empleada se basó en corpus y conllevó cinco pasos principales: (1) la construcción de un corpus especializado de artículos científicos, (2) la extracción de una lista de 4-gramas de manera automática usando el software Sketch Engine, (3) la verificación manual de esa lista para eliminar candidatos inadecuados, (4) la clasificación de las unidades seleccionadas dependiendo de su estructura, y (5) la categorización de las unidades seleccionadas según su función en el texto. Los resultados muestran que, con respecto al primer objetivo, los 4-gramas más típicos fueron sintagmas nominales; en relación con el segundo objetivo, las secuencias examinadas trataban principalmente con la investigación llevada a cabo y la autoría de los textos. En conjunto, podemos decir que estas estructuras eran específicas del género estudiado, aunque también podrían ser usadas en otros dominios.

PALABRAS CLAVE: corpus especializado; 4-grama; función; artículo científico; estructura.

1. INTRODUCTION

According to Sinclair (2000: 197), 80% of the combinations of words are made through co-selections, as corpus evidence has demonstrated, which has relegated grammar to the background and may explain the exponentially growing interest in the field of phraseology in the last decades. Phraseology, deriving from Greek *phrasis* and *logia*, can be defined as the discipline which studies groups of words. It can be considered a dynamic one, for not only does it motivate the analysis of types of words that combine with each other, but also other aspects, such as the psychological factors behind the way we recall word combinations from memory, the steps involved in teaching and learning them in second language acquisition, their semantic status in specialized languages, and so forth.

The establishment of Phraseology as a discipline was supported by a group of Russian scholars conducting research in the 1940s. They became especially active in this field, their aim being to categorize all set expressions as well as to describe their varying degrees of fixedness. The group was led by Victor V. Vinogradov (he published his selected works in 1947) and Natalia N. Amosova (she studied the foundations of English phraseology in 1963), who prepared the ground for further work on the classification of word combinations. For

instance, Klappenbach (1968), Weinreich (1969), and Lipka (1974) followed Vinogradov's steps in the 1960s and 1970s, and other authors working in the 1980s and 1990s deserve special mention for stating more practical than theoretical purposes, namely Anthony P. Cowie (1981, 1998), Peter A. Howarth (1996), and Igor A. Mel'čuk (2006, 2012).

As to the phraseological units emerging from the categorizations of these groups of authors, these sequences were defined according to two main criteria: degree of fixedness and compositionality. As such, a continuum is conceived where free combinations (i.e., the most flexible and most transparent units) are located at one end and idiomatic expressions (i.e., the most fixed and most opaque structures) are at the other. In the middle are collocations, defined as typical combinations of two lexical items. They have normally been studied from two perspectives (Nesselhauf 2005; Granger & Paquot 2008; Gablasova *et al.* 2017): the phraseological approach, according to which they are partly stable and transparent, and the statistical approach, focused on the frequency of the co-occurrence of the two words.

Aside from that, other phraseological units have also been identified that are not based on the criteria of fixedness and compositionality. For example, *lexical bundles* are defined as "extended collocations: sequences of three or more words that show a statistical tendency to co-occur" (Biber & Conrad 1999: 183). Their importance in a language is highlighted by Biber and Barbieri (2007), who state that they "serve important discourse functions in both spoken and written texts" (p. 264). Another instance is the concept of *collostruction*, a term coined by Stefanowitsch and Gries in 2003 that refers to the combination of a lexeme and a construction, both of which being mutually attracted. Finally, a *collocational catena* is described as "an independently recurrent, multiword dependency subtree which is formally and/or contextually stereotyped" (Pęzik 2018: 96). To put it another way, several words which establish dependency relations combine to form longer structures, therefore, a collocational catena can be regarded as a cluster of several collocations.

Regarding the type of combination explored in this paper, an *n-gram* can be defined as "a recurrent string of uninterrupted word-forms" (Stubbs 2007: 90); the letter 'n' defines the number of elements involved. In this work, 4-grams are to be examined, which means that the focus will be on uninterrupted sequences of four words. Many scholars have acknowledged that n-gram is another name for lexical bundle (e.g., Allen 2010; Granger 2014; Kwary *et al.* 2017; Biel 2018). This type of multiword expression has usually been tackled through a frequency-based approach (Biber *et al.* 1999; Wray 2008) and has played a crucial role in linguistic production (Pawley & Syder 1983; Thomson 2017). For instance, n-grams are said to be a marker of proficient language use and native-like language competence, to contribute to natural language use and greater fluency, and to facilitate language production. These may be some of the reasons why n-grams have become a key issue in specialized languages, such as tourism (Fuster-Márquez 2014), academic writing (Shin *et al.* 2018), and economics (Aimenova *et al.* 2019).

This study seeks to contribute to the analysis of 4-grams in the specialized language of science, more specifically, in the research article genre. The underlying hypothesis being that these phraseological units can help to structure the text and convey specific nuances of meaning when used in a concrete text type. In order to test this hypothesis, two specific aims will be pursued: (1) to explore the structural frameworks of these sequences of words, and (2) to examine their function in the text. A compelling reason for focusing on 4-grams and not 3- or 5-grams is that 4-grams often subsume 3-grams (Pérez-Llantada 2014: 86).

Furthermore, 4-grams are far more frequent than 5-grams and present a clearer range of structures and functions (Hyland 2008: 44). The organization of the paper is as follows. I start with a broad overview of studies on word sequences in research articles (Section 2, Background). Then, I describe the methodology employed to achieve my objectives (Section 3, Methodology). After that, I present the main findings of the study (Section 4, Results) and compare them to those obtained in previous works (Section 5, Discussion). Finally, I summarize the conclusions drawn as well as provide some possible research lines for the future (Section 6, Conclusions and Further Research).

2. BACKGROUND

The phraseology that characterizes the scientific genre has attracted the attention of linguists for many years, and a wide variety of word combinations have been analyzed, such as collocations (Luzón Marco 2000; Jiménez-Navarro 2019), colligations (Gledhill 2009; Menon & Mukundan 2010), and lexical bundles (Pérez-Llantada 2014). These studies have supported Biber and Barbieri's (2007) idea that each register has a distinct set of phraseological units which are "associated with the typical communicative purposes of that register" (p. 265). To exemplify, Luzón Marco (2000) finds that the linguistic conventions of a genre determine the specific collocates chosen as well as their meaning in the text, the latter being related to measure, quantification, probability, or cause/result. On the other hand, Menon and Mukundan (2010) point out that some elements need to be interpreted beyond their literal sense as they convey a more specific and/or extended meaning in the scientific field. In this section, I will describe in more detail some of the investigations into the role of n-grams[1] that have been carried out in several disciplines.

First, Jalali *et al.*'s (2015) aim is threefold: (1) to retrieve the most frequent lexical bundles from COMRA, the Corpus of Medical Research Articles; (2) to explore the forms of these multiword units; and (3) to discover their function in the texts. Their corpus contains 790 articles published between 2009 and 2011 covering 33 subject areas (including oncology, psychiatry, surgery, public health, and health policy), resulting in around 2.4 million words. In their study, they only focus on 4-word sequences and extract 102 different items of this type, highlighting that just 0.3% of the whole corpus consists of lexical bundles. The results show that almost half (i.e., 44.5%) of these lexical bundles selected are prepositional phrases, and the most common function performed is related to the text itself. For example, more than 40% of these units are used to express transition and result, among others. The authors' conclusion is that "the context establishes the function of bundles" (p. 61).

Second, Jalilifar *et al.* (2017) try to identify common core lexical bundles across three disciplines, namely, arts and humanities, sciences, and social sciences. Their 6-million-word corpus contains over 1,300 research articles published between 2010 and 2015. Unlike Jalali *et al.* (2015), these authors consider 3-grams, 4-grams, and 5-grams, extracting separate lists from each discipline and then collapsing them into one single list which includes all types of n-grams retrieved from every discipline. They set a minimum frequency of three co-occurrences for a particular lexical bundle to be regarded as a core one. They extract a total of 11,670 bundles, 661 of them being core and 3-word sequences

1 They are referred to as lexical bundles in the studies described here.

being the most frequent (85%). In their opinion, these results reveal that lexical bundles are highly frequent and significant in academic writing. As for their function in the text, they are mostly text-oriented (58.5%), which means that they are used to focus the reader on the text. In addition, these authors address the extent to which these core bundles overlap with the bundles from a corpus of applied linguistics and discover that 593 (out of 661) sequences are central to this discipline too.

And, finally, Jalilifar and Ghoreishi (2018) make use of the previous study to find the proportion of general and discipline-specific formulaic sequences in a corpus of research articles from the field of applied linguistics. They retrieve sequences of 3, 4, and 5 words, setting a minimum frequency threshold of 10 tokens per million words and a dispersion criterion of five texts (i.e., the lexical bundles should occur in at least five different texts from the corpus). Their results show that general sequences account for 53% of all the extractions, and discipline-specific bundles for 47%. In these authors' opinion, these findings demonstrate that multiword units have clear pedagogical implications for academic writing students, in that they should be aware of the existence and relevance of these word sequences in order to incorporate them into their written production. Furthermore, as to function, again text-oriented sequences are most common in both groups of the formulaic sequences under study, that is, 58.5% in the general context and 55.2% in the discipline-specific context.

In short, the main idea suggested by the previous studies is that the most recurrent n-grams found in research articles across different disciplines (i.e., medicine, arts and humanities, sciences, social sciences, and applied linguistics) are concerned with the organization of the text and its meaning. This is in consonance with the findings of other research into the language of science, such as Luzón Marco (2000) and Pérez-Llantada (2014). The current paper aims to learn more about the structure and the function performed by 4-grams in, specifically, scientific research articles. The next section addresses the methodological steps followed in order to achieve this goal.

3. METHODOLOGY

Without a doubt, the composition of the corpus is an essential element in corpus-based and corpus-driven studies, since the results obtained will depend solely on its characteristics and, of course, on the researchers' skills in obtaining and evaluating these results. For this reason, how it is compiled is of supreme importance. The construction of UCOSCIENCOR, the specialized corpus used in the current research, will comprise the first part of this section. After that, I will explain the extraction of 4-grams, their classification according to their structural framework, and their categorization based on their function in the text.

3.1. Compilation of UCOSCIENCOR

UCOSCIENCOR is a specialized corpus of over three million words built for the linguistic analysis of the English language used by non-native speakers who are part of the Universidad de Córdoba (UCO, Spain). It contains more than 600 publicly available research articles dealing with the disciplines of medicine, veterinary science, and genetics, and published in high-impact international journals (such as *Nature Reviews Immunology, The Lancet, Nature Reviews Molecular Cell Biology*, *Vaccine*) from 1980 to 2020 by leading scholars. Given

the specificity of the corpus, the texts were manually selected first and then automatically annotated using the corpus management software Sketch Engine. The characteristics of the texts selected (e.g., authorship, place of publication, year of publication, length) and the large size of the corpus make UCOSCIENCOR representative of the genre under study.[2] Table 1 provides a summary of the main features of UCOSCIENCOR.

Number of words	3,009,296
Type of corpus	Specialized / electronic
Size	Medium
Mode	Written
Language	English
Domain	Science
Subdomain	Medicine / Veterinary Science / Genetics
Genre	Research article
Text length	Complete texts
Purpose	Linguistic analysis
Communicative situation	Specialized
Publication date	1980-2020
Source of texts	High-impact international journals
Publishers	Non-native UCO academic researchers

Table 1. Defining characteristics of UCOSCIENCOR.

3.2. Extraction of 4-grams

After having compiled my specialized corpus, the 'N-grams' function available at Sketch Engine was used to extract 4-word sequences typical of the corpus. Following Jalali *et al.* (2015), the frequency threshold set was 20 tokens per million words, which means that only those 4-grams occurring at least 60 times were retrieved. Using a minimum frequency criterion helps to identify groupings of words characteristic of the language under analysis and to avoid arbitrary uses of language. Apart from this, another criterion established was that no distinction was made between lowercase and uppercase versions of the same sequence.

This step produced a list of 251 items with a total frequency of 33,784 tokens in the corpus. However, manual work was needed to evaluate the results and 70 items were discarded for the following reasons:

(1) They were linked to journal names, for instance, *j am soc nephrol, am j clin nutr*;

(2) They were part of an institution's name, such as *reina sofia university hospital, instituto maimónides de investigación*;

(3) They were abbreviations of words, for example, *t i c l, e i n f;*

2 According to Vargas Sierra (2006: 11) and Corpas Pastor and Seghiri Domínguez (2010: 124), specialized texts are denser than general language texts when it comes to terminology, consequently, a specialized corpus does not need to contain as many millions of words as it would be desirable for a reference corpus so as to be representative of a specific domain.

(4) They were irrelevant to the focus of this study, such as *elsevier b.v. all rights, lists available at sciencedirect.*

The final list of 4-grams amounted to 181 items.

3.3. Structural categorization of 4-grams

The next step in this research was the categorization of the selected 4-grams on the basis of their structure following Biber *et al.* (1999: 96-106), who describe five types of phrase, namely: (1) noun phrases, that is, phrases with a noun as head (e.g., *her gold **watch**, any printed **material** discovered*); (2) verb phrases, which contain a lexical verb or primary verb as head or main verb, either alone or accompanied by one or more auxiliaries (e.g., *had been **making**, should have **said***); (3) adjective phrases, with an adjective as head, optionally accompanied by modifiers (e.g., ***slow** to respond, **guilty** of a serious crime*); (4) adverb phrases, which contain an adverb as head, optionally accompanied by modifiers (e.g., *quite **melodiously**, much more **quickly** than envisaged*); and (5) prepositional phrases, consisting of a preposition and a complement, most typically in the form of a noun phrase (e.g., ***in** the morning, **to** Sue*).

3.4. Functional categorization of 4-grams

The final step of my methodology was the categorization of the 4-grams selected according to their function in the text. To do so, I followed Hyland (2008: 49), who suggests three broad categories loosely based on Halliday's (1994) linguistic macrofunctions: (1) research-oriented, which serves an ideational function; (2) text-oriented, that is, phrases concerned with textual functions; and (3) participant-oriented, expressing interpersonal meanings. More detail about this taxonomy is presented below:

(1) Research-oriented sequences. They help writers to structure the information.
   a. Location, which indicates time and place (e.g., *at the beginning of, at the same time, in the present study*).
   b. Procedure, concerning methods and processes (e.g., *the use of the, the role of the, the purpose of the*).
   c. Quantification, related to quantities (e.g., *the magnitude of the, a wide range of, one of the most*).
   d. Description, used to present facts (e.g., *the structure of the, the size of the, were no significant differences*).
   e. Topic, connected with the field of research (e.g., *in the Hong Kong, the currency board system, risk of cardiovascular disease*).

(2) Text-oriented sequences. These are concerned with the organization of the text and the meaning of its elements as a message or argument.
   a. Transition signals, which establish additive or contrastive links between elements (e.g., *on the other hand, in addition to the, in contrast to the*).
   b. Resultative signals, which mark inferential or causative relations between elements (e.g., *as a result of, it was found that, these results suggest that*).
   c. Structuring signals, defined as text-reflexive markers which organize stretches of

discourse or direct the reader elsewhere in text (e.g., *in the present study, in the next section, as shown in fig.*).

d. Framing signals, which situate arguments through specifying limiting conditions (e.g., *with respect to the, on the basis of, with the exception of*).

(3) Participant-oriented sequences. These are focused on the writer or the reader of the text (Hyland 2005):

a. Stance features, which convey the writer's attitudes and evaluations (e.g., *are likely to be, may be due to, it is possible that*).

b. Engagement features, addressing readers directly (e.g., *it should be noted, as can be seen, it must be highlighted*).

Additionally, I decided to create a third subcategory of participant-oriented sequences that I called authorship-related, which concerns details linked to the authors of the publications and the publications themselves (e.g., *the author of this, profiles for this publication, would like to thank*), given the relevance of such phrases in UCOSCIENCOR, as I will explain in Section 4.2.

4. RESULTS

In total, 181 items were manually selected and analyzed both structurally and functionally in this work. The full list of 4-grams is provided in Appendix 1, although Table 2 shows the top ten items. They are organized in terms of their frequency in the corpus and are accompanied by their frequency per million (i.e., their relative frequency), the number of texts in which they occur, and the relative document frequency (i.e., the percentage of texts that contain the item).

Rank	4-gram	Frequency	Frequency per million	Document frequency	Relative document frequency
1	*in the present study*	376	86.64	177	29.45
2	*the authors of this*	375	86.41	72	11.98
3	*of the authors of*	372	86.41	372	61.89
4	*of this publication are*	361	83.86	361	60.06
5	*some of the authors*	360	83.62	360	59.90
6	*the user has requested*	360	83.62	360	59.90
7	*page was uploaded by*	360	83.62	360	59.90
8	*and author profiles for*	360	83.62	360	59.90
9	*following this page was*	360	83.62	360	59.90
10	*user has requested enhancement*	360	83.62	360	59.90

Table 2. Top ten 4-grams extracted from UCOSCIENCOR.

As can be seen, the most frequent 4-gram is *in the present study*, with 376 tokens occurring in 177 different texts, which makes a corpus distribution of 29.45%. Other recurrent and more highly distributed sequences are *of the authors of* (372 tokens, 61.89%) and *of this publication are* (361 tokens, 60.06%), along with the rest of 4-grams included in the table,

which occur 360 times in 360 texts (59.9%). However, the case of *the authors of this* must be highlighted, given that it is highly recurrent (375 tokens) but unevenly distributed (11.98%).

At the other end of the list, there are nine 4-grams that occur only 60 times, which was the minimum frequency threshold set in the extraction step, namely: *was significantly higher in, that the presence of, study was to evaluate, seroprevalence of toxoplasma gondii, the area under the, and the number of, was supported by the, not included in the*, and *was found to be*. With respect to their distribution in the corpus, *was supported by the* and *not included in the* show the highest distribution (9.65% and 9.15%, respectively), followed by *study was to evaluate* and *and the number of* (8.48% and 8.15%, respectively). After that, the most widely distributed 4-grams are: *was significantly higher in* (7.15%), *that the presence of* (6.48%), *the area under the* (6.32%), and *was found to be* (6.32%). Finally, the least distributed sequence is *seroprevalence of toxoplasma gondii* (2.66%), which may be due to the fact that it is so much more specific than the others and therefore found in a lower number of texts (16 out of 601 texts).

4.1. Structural features of 4-grams in UCOSCIENCOR

The first aim pursued in this paper was the structural classification of the 181 4-grams selected. As I have previously mentioned, Biber *et al.*'s (1999: 96-106) work inspired this step, thus, the 181 sequences were divided into five different groups: (1) noun phrases, (2) verb phrases, (3) adjective phrases, (4) adverb phrases, and (5) prepositional phrases. Table 3 shows this classification, which is organized in terms of the number of 4-grams found in each type of phrase.

Type of phrase	Number of types	Number of tokens	% in the corpus	Examples
Noun phrase	80	10,627	45.6	*some of the authors, the results of the, the present study was*
Prepositional phrase	52	6,938	29.8	*at the end of, in the presence of, for the treatment of*
Verb phrase	43	4,904	21.1	*also working on these, are shown in table, were included in the*
Adjective phrase	4	576	2.5	*it is possible that, was significantly higher in*
Adverb phrase	2	243	1	*as well as the, as well as in*
Total	**181**	**23,288**	**100**	

Table 3. Structural classification of the 4-grams selected.

Table 3 reveals that the largest structural category of 4-grams is noun phrases, with 80 distinct types (i.e., distinct phrases regardless of their repetition in the corpus), which accounts for almost half (45.6%) of the total number of sequences. The number of tokens found in the corpus is 10,627, and some of the types of phrases included in this group consist of a noun phrase plus a preposition. In terms of frequency, the combination noun phrase + *of* (e.g.,

the end of the, the effect of the) is the most recurrent (38.75%), followed at a considerable distance (7.5%) by the pattern noun phrase + *in* (e.g., *an important role in, a significant increase in*), noun phrase + *with* (6.25%, e.g., *patients with ankylosing spondylitis, patients with rheumatoid arthritis*), and noun phrase + *for* (5%, e.g., *anova for repeated measures, profiles for this publication*).

The second largest category is prepositional phrases (29.8%). In total, 52 4-grams are identified, with an overall frequency of 6,938 tokens in the corpus. The most typical preposition introducing a phrase is *in* (34.62%, e.g., *in advanced colorectal cancer, in the treatment of*), followed by *of* (19.23%, e.g., *of this publication are, of the mediterranean diet*) and *for* (11.54%, e.g., *for the management of, for the first time*). Some of the most recurrent phrases in UCOSCIENCOR are prepositional, such as *in the present study* (376 tokens, the highest frequency in the corpus), *of the authors of* (372 tokens), and *of this publication are* (361 tokens).

Verb phrase is the third more recurrent category in the corpus (21.1%). In total, 43 different 4-grams are found consisting of 4,904 tokens. Various subtypes are observed, the most common (37.2%) being the combination passive verb + prepositional phrase (e.g., *was approved by the, has been associated with*). After that, 16.28% of the recurrences belong to two types, namely, verb *to be* + adjective/noun (e.g., *be due to the, is one of the*) and main verb + prepositional phrase (e.g., *distributed under the terms, included in the study*). Next, in terms of quantity, 9.3% of the verb phrases is represented by the combination main verb + noun phrase (e.g., *plays an important role, has requested enhancement of*), and 6.97% belongs to two types, anticipatory *it* + passive (e.g., *it has been shown, it has been suggested*) and modal verb phrase (e.g., *could be due to, would like to thank*). The passive forms achieve the highest overall frequency with 1,870 tokens.

Finally, adjective phrases and adverb phrases are the least common types of sequence in UCOSCIENCOR. As for the former, only 4 distinct 4-grams (2.5%), with 576 tokens, are included, examples being *it is important to* and *it is possible that*. Regarding adverb phrases, only 2 different 4-grams (1%), with a total frequency of 243 tokens, are identified: *as well as the* and *as well as in*.

4.2. Functions of 4-grams in UCOSCIENCOR

The second aim stated in this work was the functional classification of the 181 4-grams selected. To this end (as previously mentioned), I followed Hyland (2008: 49), who proposes three large categories: research-oriented, text-oriented, and participant-oriented sequences. Each of them presents several subcategories. Additionally, I decided to create a new subcategory in the participant-oriented group, that is, authorship-related. Table 4 shows the results of this classification and includes the number of 4-grams, their overall frequency in UCOSCIENCOR, and the percentage of the categories and subcategories in the corpus.

Category/Subcategory	Number of types	Number of tokens	% in the corpus
Research-oriented	109	10,715	59.9
Location	15	2,431	8.2
Procedure	29	2,777	16

Category/Subcategory	Number of types	Number of tokens	% in the corpus
Quantification	13	1,302	7.2
Description	27	2,146	14.8
Topic	25	2,059	13.7
Text-oriented	33	3,572	18
Transition signals	4	617	2.2
Resultative signals	14	1,191	7.6
Structuring signals	1	194	0.6
Framing signals	14	1,570	7.6
Participant-oriented	39	9,001	22.1
Stance features	11	805	6.7
Engagement features	0	0	0
Authorship-related	28	8,196	15.4
Total	**181**	**23,288**	**100**

Table 4. Functional classification of the 4-grams selected.

As the table makes clear, the largest amount of 4-grams (59.9%) belong to the research-oriented category, specifically, 109 items. The sequences included in this group describe various features of the research presented in the texts of UCOSCIENCOR: (1) 29 items (16%) address the methods and purposes of the research so are in the procedure subcategory (e.g., *aim of this study, play an important role, for the treatment of*); (2) 27 items (14.8%) represent qualities of the studies within the description subcategory (e.g., *an increased risk of, no significant differences in, the influence of the*); (3) 25 sequences (13.7%) are to be found in the topic subcategory, given that they are related to the field of research (e.g., *ankylosing spondylitis functional index, microbiology and infectious diseases, patients with rheumatoid arthritis*); (4) 15 sequences (8.2%) belong to the location subcategory and indicate time and place (e.g., *after consumption of the, at the same time, on these related projects*); and (5) the least recurrent subcategory is quantification with 13 items (7.2%), which describe amounts and numbers (e.g., *a large number of, a wide range of, the rest of the*). In spite of the fact that both the description and topic subcategories are represented by a much larger number of 4-grams than the location subcategory (27 and 25 items, respectively, vs. 15 items), it should be highlighted that, if overall frequency of the sequences is considered, the location subcategory is more recurrent than the others, having 2,431 tokens in total, while the description subcategory has 2,146 tokens and the topic subcategory, 2,059 tokens.

Secondly, 39 items (22.1%) are in the participant-oriented category. Despite having previously identified three subgroups, only two of them are in fact evoked by the 4-grams selected. On the one hand, the newly created authorship-related subcategory contains 28 sequences (15.4%) that concern details related to the authors of the publications and the publications themselves (e.g., *authors of this publication, for personal use only, this work was supported*). Needless to say, this new subcategory is essential, containing as it does 8,196 tokens in total. On the other hand, the stance features subcategory is less commonly found and includes 11 items (6.7%), which convey the writers' attitudes and evaluations (e.g., *has been associated with, it is important to, was found to be*). In this work, the engagement features subcategory is not evoked by any of the 4-grams selected.

Finally, 33 sequences (18%) are included in the text-oriented category, where four subcategories are considered: (1) framing signals, which contain 14 sequences (7.6%) used to limit conditions (e.g., *in relation to the, in the absence of, in the case of*); (2) resultative signals, also 14 items (7.6%) although with a lower frequency in the corpus (1,191 vs. 1,570 for framing signals), which mark inferential or causative relations between elements (e.g., *as a consequence of, as a result of, studies have shown that*); (3) transition signals, with only 4 sequences (2.2%), are used to establish additive or contrastive links between elements (e.g., *in addition to the, on the other hand, as well as the*); and (4) structuring signals, with only 1 item (0.6%, i.e., *are shown in table*) which directs the reader elsewhere in the text.

4.3. Interconnection between structures and functions of 4-grams in UCOSCIENCOR

After having examined the most common structural frameworks found in UCOSCIENCOR and the functions they perform, the interconnection between these two central aspects cannot be disregarded. Thus, Table 5 illustrates the type of structure used to serve each of the functions; the number of 4-grams representing the distinct categories and their percentage in UCOSCIENCOR are shown.

	Research-oriented		Participant-oriented		Text-oriented	
Noun phrase	64	58.8%	14	35.8%	2	6.1%
Prepositional phrase	30	27.5%	4	10.3%	18	54.4%
Verb phrase	14	12.8%	18	46.2%	11	33.4%
Adjective phrase	1	0.9%	3	7.7%	0	0%
Adverb phrase	0	0%	0	0%	2	6.1%
Total	**109**	**100%**	**39**	**100%**	**33**	**100%**

Table 5. Structural frameworks used in the functional categorization.

Table 5 shows that each major functional category is mostly characterized by a different specific structural framework. Noun phrases are the most typical structure when it comes to research-oriented 4-grams (e.g., *statistically significant differences were*, *the influence of the*), with 64 structures of this type identified that account for 58.8% of the structures used in this function. Verb phrases are common to represent participant-oriented sequences (e.g., *be explained by the, be due to the*), 18 of them comprising 46.2% of the structures identified. A coherent explanation for this may be that it is specifically the use of certain verbs that makes possible the conveyance of the writer's attitude (e.g., *would like to thank*) or the carrying out of their work (e.g., *was supported by the*).

And finally, 18 prepositional phrases describe text-oriented 4-grams (e.g., *in the pathogenesis of, under the terms of*), making up 54.4% of the structures found in this function. Without a doubt, prepositions are very much needed to organize the text by introducing additions (e.g., *in addition to the*), contrasts (e.g., *on the other hand*), marking relations between elements (e.g., *as a consequence of*), or framing arguments (e.g., *in relation to the*), among other functions.

5. DISCUSSION

In the first place, in terms of the structural frameworks identified in this work, it must be pointed out that these results coincide with those obtained by Jalali *et al.* (2015) as regards the two most frequent types of phrases, that is, noun phrases and prepositional phrases, although the order is reversed in the two works. Noun phrases are the most typical combinations in UCOSCIENCOR (45.6%) with prepositional phrases in second place (29.8%), while prepositional phrases are the most typical sequences in COMRA (44.5%) followed by noun phrases (20.42%). In both studies verb phrase is in third place, and recurrence is similar in both works (21.1% in UCOSCIENCOR and 20.59% in COMRA).

Taking this into consideration, the results yielded in the current research are in consonance with the long-held assumption that nominal complexes have always been present in the language of science, and that complex noun phrases characterize the specialized language of science in English (Nagy 2014: 265). For example, Halliday (1993) performs a comprehensive analysis of works which support the idea that nominal structures have always been a distinctive feature of scientific texts (e.g., Chaucer 1391; Dalton 1827; Maxwell 1881). More recently, other authors have also focused on multiword units arranged around a noun, such as Luzón Marco (2000) and Pérez-Llantada (2014). What is more, noun phrases have also been found to be common in scientific texts written in Spanish (Soto Vergara & Zenteno Bustamante 2004), which is the mother tongue of the authors of the research articles included in UCOSCIENCOR, this being another plausible reason for the more varied repertoire of noun phrases compared to COMRA (i.e., the influence of the authors' first language).

On the other hand, Jalali *et al.* (2015) emphasize that prepositional phrases are used to identify a particular time period or location, although they do not specify why this is the case nor mention the most frequent prepositions. One explanation for the prevalence of prepositional phrases in COMRA may be the nature of the texts, since they may be more concerned with reporting these two aspects of the research conducted (i.e., time and location) or including time references to previous research (Swales 1990: 144, as cited in Gledhill 2000: 36). Additionally, Jalali *et al.* (2015) compare their work to previous studies and conclude that some prepositional phrases typical of their corpus were not extracted in other works, such as *in the present study* (this is in fact the most frequent 4-gram extracted from UCOSCIENCOR)*, at the time of, of the present study, in the control group, in the current study.* Nevertheless, the results of both studies (i.e., Jalali *et al.* 2015 and the present research) are in line with Chen and Baker's (2010) findings that these two groups of phrases, nominal and prepositional, are frequently used in expert writing.

In second place, with respect to the functions of the 4-grams analyzed in this study, significant differences must be noted when they are compared to other works. As has been previously stated, research-oriented sequences are the most common (59.9%) in UCOSCIENCOR, followed by participant-oriented sequences (22.1%) and text-oriented sequences (18%). In contrast, text-oriented sequences are the most recurrent in the work of Salazar (2014), Jalali *et al.* (2015), Jalilifar *et al.* (2017), and Jalilifar and Ghoreishi (2018); to clarify, they account for 48%, 42.5%, 58.5%, and 58.51%, respectively, in the corpora analyzed by these authors. Jalali *et al.* (2015) justify their finding of the extensive use of text-oriented sequences by arguing that scholars employ them in order to show a sophisticated approach to language and demonstrate their mastery of it. Added to that, this type of sequence helps

to organize their discourse in the way that readers can better understand it, which is fully achieved thanks to bundles that make connections, present contrastive information, frame arguments, and so on.

However, the fact that research-oriented sequences are so prevalent in UCOSCIENCOR accords with Parvizi's findings in a study focusing on education (2011, as cited in Jalali *et al*. 2015: 65) that they outweigh all other functional types of bundles. Moreover, when it comes to the language of science, Jalilifar *et al.* (2017) suggest that this difference may also be due to the methods used in distinct types of science, that is, soft and hard, given that "writers in hard sciences [are required] to use less evaluative patterns of argument which are mostly realized by text-oriented bundles" (p. 11).

Finally, a common feature among all disciplines (i.e., science, applied linguistics, etc.) is that participant-oriented sequences have the lowest representation of all categories. It might be expected given that these bundles involve the reader or the writer of the texts, but in these fields the emotive or conative functions of language (Jakobson 1960) do not normally emerge, since the texts are referential and are more interested in transmitting accurate and precise information rather than focusing on the addresser or addressee of the message.

6. CONCLUSIONS AND FURTHER RESEARCH

The current study has delved into the use of 4-grams by academic researchers from the UCO in their scientific papers. After having compiled a specialized corpus, I automatically extracted a list of potential 4-grams and then manually discarded those which did not meet the eligibility criteria. In total, 181 items were selected and examined in order to achieve the two main objectives of this work.

First of all, and based on Biber *et al*.'s (1999: 96-106), I explored the structural frameworks of these sequences and discovered that noun phrases were the most typical structure (45.6%), followed by prepositional phrases (29.8%) and verb phrases (21.1%). The least common structures were adjective phrases (2.5%) and adverb phrases (1%). These results validate the widespread assumption that scientific language is characterized by complex nominals, albeit they are not in line with Jalali *et al*.'s (2015) findings, where prepositional phrases account for almost half (44.5%) of the multiword units analyzed. Nevertheless, it is important to realize that these two types of phrases are also said to be frequent in expert writing, which is the category where academic scholars may fit in.

After that, these sequences were classified according to the function they perform in the corpus, as also done by Hyland (2008: 49). I found that the majority of 4-grams (59.9%) were used to express information related to the research described in the scientific papers itself. Furthermore, these word combinations mostly concerned the authors of the articles, their publications, or their evaluations (22.1%), as well as the texts themselves (18%). At this point, it must be emphasized that these results do not conform to those obtained in Salazar (2014), Jalali *et al.* (2015), Jalilifar *et al*. (2017), or Jalilifar and Ghoreishi (2018), where text-oriented sequences were the most recurrent, which may be explained on the basis that scholars are likely to incorporate such bundles into their writing so as to make their discourse more elaborate.

As for the relation between these two aims, different structures were more frequently used depending on the type of function fulfilled. As such, noun phrases mainly represented

research-oriented 4-grams (58.8%), verb phrases were prominently identified with participant-oriented 4-grams (46.2%), and finally text-oriented 4-grams were typically described by prepositional phrases (54.4%). Without a doubt, the headword of these sequences contributed to the function served in the texts, as complex nominals are usually the most specialized multiword units (which explains their prevalence in the research-oriented category), while verbs are used to express evaluations and opinions (participant-oriented category), and prepositions frequently introduce phrases that are useful in organizing the text (text-oriented category).

My goals have therefore been successfully achieved and this paper has contributed to the description of multiword units (more specifically, 4-grams) used in the scientific research article genre. Moreover, I can confirm my hypothesis, that is, some of these phraseological units help to structure and organize the text (i.e., the text-oriented sequences) and acquire specific nuances of meaning in this particular context. To put it another way, despite knowing that they could also be employed in other domains, as some of the works mentioned above acknowledge (e.g., Pérez-Llantada 2014; Jalilifar & Ghoreishi 2018), it is important to realize that the research-oriented combinations analyzed, which were the most recurrent type, dealt with various specific aspects of the research presented in the corpus, such as detailed descriptions, procedures, and locations.

Future studies should focus on shorter and/or longer n-grams (e.g., 3-/5-grams) in the same corpus as well as in other languages, such as Spanish. Aside from that, these phraseological units may be explored separately in the different sections of the research article (e.g., abstract, methods, discussion) and the results compared afterwards. Furthermore, the same methodological steps could be employed to examine other specialized languages, such as academic writing, law, or the environment, and see whether the type and function of multiword structures coincide with those found here.

REFERENCES

Aimenova, M., A. Ospanova, A. Rakhimova, A. Sarsembayeva & Z. Mazhit. (2019). Phraseological terminology in the English economic discourse. *Xlinguae,* 12(1), 228-238. https://doi.org/10.18355/XL.2019.12.01.18

Allen, D. (2010). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education: KJEE*, 1, 105-127.

Amosova, N. N. (1963). *Osnovy anglijskoj frazeologii*. Izdatelstvo Leningradskogo Universyteta.

Biber, D. & F. Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes,* 26, 263-286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D. & S. Conrad. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181-189). Rodopi.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. (1999). *The Longman grammar of spoken and written English*. Longman.

Biel, Ł. (2018). Lexical bundles in EU law: The impact of translation process on the patterning of legal language. In S. Goźdź-Roszkowski & G. Pontrandolfo (Eds.), *Phraseology in legal and institutional settings: A corpus-based interdisciplinary perspective* (pp. 11-26). Routledge. https://doi.org/10.4324/9781315445724

Chaucer, G. (1391). *A treatise on the astrolabe*. Early English Text Society Edition.

Chen, Y. & P. Baker. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology,* 14(2), 30-49.

Corpas Pastor, G. & M. Seghiri Domínguez. (2010). Size matters: A quantitative approach to corpus representativeness. In R. Rabadán, M. Fernández López & T. Guzmán González (Eds.), *Lengua, traducción, recepción en honor de Julio César Santoyo* (pp. 111-145). Universidad de León, Área de Publicaciones.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics,* 2(3), 223-235. https://doi.org/10.1093/applin/II.3.223

Cowie, A. P. (1998). Phraseological dictionaries: Some east-west comparisons. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 209-228). Clarendon Press.

Dalton, J. (1827). *A new system of chemical philosophy*. George Wilson.

Fuster-Márquez, A. (2014). Lexical bundles and phrase frames in the language of hotel websites. *English Text Construction,* 7(1), 84-121.
https://doi.org/10.1075/etc.7.1.04fus

Gablasova, D., V. Brezina & T. McEnery. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning,* 67(S1), 155-179. https://doi.org/10.1111/lang.12225

Gledhill, C. J. (2000). *Collocations in science writing*. Gunter Narr Verlag Tübingen.

Gledhill, C. J. (2009). Colligation and the cohesive function of present and past tense in the scientific research article. In D. Banks (Ed.), *Les temps et les textes de spécialité* (pp. 65-84). L'Harmattan.

Granger, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast,* 14, 58-72. https://doi.org/10.1075/bct.87.04gra

Granger, S. & M. Paquot. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 27-49). John Benjamins Publishing Company. https://doi.org/10.1075/z.139.07gra

Halliday. M. A. K. (1993). On the language of physical science. In M. A. K. Halliday & J. R. Martin (Eds.), *Writing science: Literacy and discursive power* (pp. 54-68). The Falmer Press.

Halliday, M. A. K. (1994). *Functions of language. 2nd edition*. Arnold.

Howarth, P. A. (1996). *Phraseology in English academic writing. Some implications for language learning and dictionary making*. Niemeyer. https://doi.org/10.1515/9783110937923

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies,* 7(2), 173-191. https://doi.org/10.1177/1461445605050365

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics,* 18(1), 41-62. https://doi.org/10.1111/j.1473-4192.2008.00178.x

Jakobson, R. (1960). Linguistics and poetics. In T. A. Sebeok (Ed.), *Style in language* (pp. 350-377). The Massachusetts Institute of Technology.

Jalali, Z. S., M. R. Moini & M. A. Arani. (2015). Structural and functional analysis of lexical bundles in medical research articles: A corpus-based study. *International Journal of Information Science and Management,* 13(1), 51-69.

Jalilifar, A. & S. M. Ghoreishi. (2018). *From the perspective of*: Functional analysis of formulaic sequences in applied linguistics research articles. *International Journal of English Studies,* 18(2), 161-186. https://doi.org/10.6018/ijes/2018/2/310351

Jalilifar, A., S. M. Ghoreishi & S. A. E. Roodband. (2017). Developing an inventory of core lexical bundles in English research articles: A cross-disciplinary corpus-based study. *Journal of World Languages,* 3(3), 182-203. https://doi.org/10.1080/21698252.2017.1301279

Jiménez-Navarro, E. L. (2019). Nominal collocations in scientific English: A frame-semantic approach. In G. Corpas Pastor & R. Mitkov (Eds.), *Computational and corpus-based phraseology. Third International Conference, Europhras 2019. Malaga, Spain, September 25-27, 2019. Proceedings* (pp. 187-199). Springer. https://doi.org/10.1007/978-3-030-30135-4_14

Klappenbach, R. (1968). Probleme der Phraseologie. *Wissenschaftliche Zeitschrift der Karl-Marx-Universität Leipzig,* 17(2), 221-227.

Kwary, D. A., D. Ratri & A. F. Artha. (2017). Lexical bundles in journal articles across academic disciplines. *Indonesian Journal of Applied Linguistics,* 7(1), 132-140. https://doi.org/10.17509/ijal.v7i1.6866

Lipka, L. (1974). Probleme der Analyse englischer Idioms aus struktureller und generativer Sicht. *Linguistik und Didaktik,* 20, 274-285.

Luzón Marco, M. J. (2000). Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes,* 19, 63-86. https://doi.org/10.1016/S0889-4906(98)00013-1

Maxwell, J. C. (1881). *An elementary treatise on electricity*. Clarendon.

Mel'čuk, I. A. (2006). Explanatory combinatorial dictionary. In G. Sica (Ed.), *Open problems in linguistics and lexicography* (pp. 225-355). Polimetrica.

Mel'čuk, I. A. (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology,* 3(1), 31-56. https://doi.org/10.1515/phras-2012-0003

Menon, S. & J. Mukundan. (2010). Analyzing collocational patterns of semi-technical words in science textbooks. *Pertanika Journal of Social Sciences & Humanities,* 18(2), 241-258.

Nagy, I. K. (2014). English for special purposes: Specialized languages and problems of terminology. *Acta Universitatis Sapientiae, Philologica,* 6(2), 261-273. https://doi.org/10.1515/ausp-2015-0018

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins Publishing Company. https://doi.org/10.1075/scl.14

Parvizi, N. (2011). *Identification of discipline-specific lexical bundles in education*. Unpublished master's thesis, University of Kashan, Kashan, Iran.

Pawley, A. & F. H. Syder. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). Longman.

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes,* 14, 84-94. https://doi.org/10.1016/j.jeap.2014.01.002

Pęzik, P. (2018). *Facets of prefabrication. Perspectives on modelling and detecting phraseological units*. Łódź University Press.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching*. John Benjamins Publishing Company.

Shin, Y. K., V. Cortes & I. W. Yoo. (2018). Using lexical bundles as a tool to analyze definite article use in L2 academic writing: An exploratory study. *Journal of Second Language Writing,* 39, 29-41. https://doi.org/10.1016/j.jslw.2017.09.004

Sinclair, J. (2000). Lexical grammar. *Naujoji Metodologija,* 24, 191-203.

Sketch Engine. [Computer software]. Lexical Computing Limited.

Soto Vergara, G. & C. Zenteno Bustamante. (2004). Los sintagmas nominales en textos científicos escritos en español. *ELUA (Estudios de Lingüística. Universidad de Alicante),* 18, 275-292. https://doi.org/10.14198/ELUA2004.18.14

Stefanowitsch, A. & S. T. Gries. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics,* 8(2), 209-243. https://doi.org/10.1075/ijcl.8.2.03ste

Stubbs, M. (2007). An example of frequent English phraseology: Distributions, structures and functions. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 87-105). Rodopi. https://doi.org/10.1163/9789401204347_007

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Thomson, H. (2017). Building speaking fluency with multiword expressions. *TESL Canada Journal,* 34(3), 26-53. https://doi.org/10.18806/tesl.v34i3.1272

Vargas Sierra, C. (2006). Diseño de un corpus especializado con fines terminográficos: El corpus de la piedra natural. *Debate Terminológico,* 2(7), 1-20.

Vinogradov, V. V. (1947). *Izbrannye trudy. Leksikologija i leksikografija*. Nauka.
Weinreich, U. (1969). Problems in the analysis of idioms. In J. Puhvel (Ed.), *Substance and structure of language* (pp. 23-81). University of California Press. https://doi.org/10.1525/9780520316218-003
Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford University Press.

APPENDIX 1

Rank	4-gram	Frequency	Frequency per million	Document frequency	Relative document frequency
1	*in the present study*	376	86.64	177	29.45
2	*the authors of this*	375	86.41	72	11.98
3	*of the authors of*	372	86.41	372	61.89
4	*of this publication are*	361	83.86	361	60.06
5	*some of the authors*	360	83.62	360	59.90
6	*the user has requested*	360	83.62	360	59.90
7	*page was uploaded by*	360	83.62	360	59.90
8	*and author profiles for*	360	83.62	360	59.90
9	*following this page was*	360	83.62	360	59.90
10	*user has requested enhancement*	360	83.62	360	59.90
11	*all content following this*	360	83.62	360	59.90
12	*this page was uploaded*	360	83.62	360	59.90
13	*content following this page*	360	83.62	360	59.90
14	*author profiles for this*	360	83.62	360	59.90
15	*authors of this publication*	360	83.62	360	59.90
16	*enhancement of the downloaded*	360	83.62	360	59.90
17	*profiles for this publication*	360	83.62	360	59.90
18	*on these related projects*	360	83.62	360	59.90
19	*of the downloaded file*	360	83.62	360	59.90
20	*for this publication at*	360	83.62	360	59.90
21	*working on these related*	360	83.62	360	59.90
22	*are also working on*	360	83.62	360	59.90
23	*has requested enhancement of*	360	83.62	360	59.90
24	*requested enhancement of the*	360	83.62	360	59.90
25	*also working on these*	360	83.62	360	59.90
26	*at the end of*	318	73.27	120	19.96
27	*on the other hand*	312	71.89	175	29.10
28	*the end of the*	239	63.41	120	19.96
29	*patients with ankylosing spondylitis*	201	46.31	34	5.65
30	*in the presence of*	200	46.08	89	14.80
31	*are shown in table*	194	45.06	142	23.61
32	*in the case of*	183	42.16	96	15.97
33	*as well as the*	170	39.48	127	21.13
34	*in the absence of*	155	35.71	94	15.63

Rank	4-gram	Frequency	Frequency per million	Document frequency	Relative document frequency
35	*the results of the*	151	35.07	92	15.30
36	*of the present study*	151	35.07	101	16.80
37	*an important role in*	149	34.33	100	16.63
38	*risk factors associated with*	147	33.87	52	8.65
39	*on the basis of*	147	33.87	93	15.47
40	*for the treatment of*	145	33.41	59	9.81
41	*of this study was*	143	32.95	122	20.29
42	*in the development of*	141	32.49	93	15.47
43	*in the treatment of*	137	31.82	54	8.98
44	*has been shown to*	136	31.59	101	16.80
45	*in patients with ankylosing*	134	31.12	31	5.15
46	*this study was to*	133	30.89	114	18.96
47	*this work was supported*	129	29.72	128	21.28
48	*an increase in the*	126	29.26	79	13.14
49	*work was supported by*	124	28.80	123	20.46
50	*the effect of the*	123	28.56	87	14.47
51	*was approved by the*	122	28.33	109	18.13
52	*the aim of this*	121	28.10	108	17.97
53	*ankylosing spondylitis disease activity*	119	27.63	27	4.49
54	*for the management of*	118	27.41	48	7.98
55	*as a result of*	117	27.17	78	12.97
56	*with respect to the*	117	27.17	71	11.81
57	*the presence of the*	113	26.24	65	10.81
58	*at the time of*	113	26.24	69	11.48
59	*the present study was*	112	26.01	89	14.80
60	*studies have shown that*	112	26.01	90	14.96
61	*were included in the*	112	26.01	86	14.30
62	*included in the study*	111	25.78	73	12.14
63	*the total number of*	105	24.38	70	11.64
64	*has been associated with*	104	24.15	66	10.98
65	*one of the most*	103	23.91	87	14.47
66	*aim of this study*	95	22.05	85	14.14
67	*is one of the*	95	22.05	79	13.14
68	*this is the first*	94	21.83	68	11.31
69	*area under the curve*	93	21.60	48	7.98
70	*there were no significant*	93	21.60	56	9.31
71	*for personal use only*	92	21.36	10	1.66
72	*patients with metastatic colorectal*	91	21.13	23	3.82
73	*with metastatic colorectal cancer*	91	21.13	24	3.99
74	*it is important to*	91	21.13	61	10.14
75	*these results suggest that*	90	20.90	59	9.81
76	*a systematic review and*	89	20.67	26	4.32

Rank	4-gram	Frequency	Frequency per million	Document frequency	Relative document frequency
77	*systematic review and meta-analysis*	89	20.67	26	4.32
78	*in accordance with the*	89	20.67	74	12.29
79	*was associated with a*	87	20.20	53	8.81
80	*be due to the*	87	20.20	68	11.31
81	*a significant increase in*	85	19.74	50	8.31
82	*bath ankylosing spondylitis disease*	85	19.74	25	4.15
83	*spondylitis disease activity index*	83	19.28	25	4.15
84	*the bath ankylosing spondylitis*	83	19.28	27	4.48
85	*in patients with chronic*	83	19.28	34	5.65
86	*peripheral blood mononuclear cells*	82	19.04	36	5.99
87	*in the pathogenesis of*	82	19.04	43	7.15
88	*were no significant differences*	82	19.04	52	8.65
89	*been shown to be*	81	18.81	64	10.64
90	*the aim of the*	80	18.57	70	11.63
91	*no significant differences were*	80	18.57	59	9.81
92	*of coronary heart disease*	80	18.57	43	7.15
93	*it has been shown*	79	18.35	59	9.81
94	*was carried out in*	79	18.35	58	9.65
95	*play an important role*	79	18.35	65	10.81
96	*in the number of*	78	18.11	60	9.98
97	*it has been suggested*	78	18.11	52	8.65
98	*study was carried out*	77	17.88	69	11.48
99	*anova for repeated measures*	76	17.65	32	5.32
100	*of the metabolic syndrome*	76	17.65	26	4.32
101	*after consumption of the*	76	17.65	18	2.99
102	*distributed under the terms*	76	17.65	75	12.47
103	*no significant differences in*	75	17.42	51	8.48
104	*has been suggested that*	75	17.42	51	8.48
105	*an increased risk of*	74	17.18	32	5.32
106	*and reproduction in any*	74	17.18	72	11.98
107	*important role in the*	74	17.18	64	10.64
108	*and the risk of*	74	17.18	39	6.48
109	*the end of each*	74	17.18	37	6.15
110	*to the presence of*	74	17.18	51	8.48
111	*reproduction in any medium*	73	16.95	72	11.98
112	*as well as in*	73	16.95	64	10.64
113	*test was used to*	73	16.95	65	10.81
114	*is associated with a*	73	16.95	52	8.65
115	*to be associated with*	73	16.95	48	7.98
116	*efficacy and safety of*	72	16.72	35	5.81
117	*a wide range of*	71	16.49	61	10.14
118	*microbiology and infectious diseases*	71	16.49	10	1.66

Rank	4-gram	Frequency	Frequency per million	Document frequency	Relative document frequency
119	*in advanced colorectal cancer*	70	16.26	18	2.99
120	*significant differences were found*	70	16.26	55	9.15
121	*to the development of*	70	16.26	48	7.98
122	*has been shown that*	70	16.26	52	8.65
123	*be related to the*	70	16.26	59	9.81
124	*it has been reported*	70	16.26	52	8.65
125	*the fact that the*	69	16.02	56	9.31
126	*the rest of the*	69	16.02	48	7.98
127	*the terms of the*	69	16.02	67	11.14
128	*with the use of*	69	16.02	37	6.15
129	*plays an important role*	68	15.79	48	7.98
130	*under the terms of*	68	15.79	67	11.14
131	*aim of the present*	68	15.79	59	9.81
132	*could be due to*	68	15.79	55	9.15
133	*patients with rheumatoid arthritis*	67	15.56	32	5.32
134	*and the presence of*	67	15.56	54	8.98
135	*for the first time*	67	15.56	55	9.15
136	*informed consent was obtained*	67	15.56	65	10.81
137	*to the fact that*	67	15.56	57	9.48
138	*mediterranean diet supplemented with*	66	15.33	16	2.66
139	*a large number of*	66	15.33	49	8.15
140	*of the mediterranean diet*	66	15.33	25	4.15
141	*of the university of*	66	15.33	46	7.65
142	*would like to thank*	66	15.33	63	10.48
143	*it is possible that*	65	15.09	52	8.65
144	*renal excretion of urates*	65	15.09	7	1.16
145	*in the regulation of*	65	15.09	50	8.31
146	*be explained by the*	65	15.09	56	9.31
147	*was used for the*	65	15.09	61	10.14
148	*the presence of a*	64	14.86	51	8.48
149	*of virgin olive oil*	64	14.86	18	2.99
150	*consent was obtained from*	64	14.86	62	10.31
151	*by the presence of*	64	14.86	52	8.65
152	*as a consequence of*	64	14.86	51	8.48
153	*has been reported that*	63	14.63	46	7.65
154	*the start of the*	63	14.63	37	6.15
155	*in chronic kidney disease*	63	14.63	24	3.99
156	*in any of the*	63	14.63	44	7.32
157	*has been reported in*	63	14.63	52	8.65
158	*the influence of the*	63	14.63	53	8.81
159	*in addition to the*	62	14.40	54	8.98
160	*for the diagnosis of*	62	14.40	34	5.65

Rank	4-gram	Frequency	Frequency per million	Document frequency	Relative document frequency
161	*received in revised form*	62	14.40	62	10.31
162	*statistically significant differences were*	62	14.40	42	6.98
163	*in relation to the*	62	14.40	45	7.48
164	*on behalf of the*	62	14.40	54	8.98
165	*in the expression of*	62	14.40	30	4.99
166	*used in this study*	62	14.40	52	8.65
167	*risk of cardiovascular disease*	61	14.16	31	5.15
168	*bath ankylosing spondylitis functional*	61	14.16	28	4.65
169	*ankylosing spondylitis functional index*	61	14.16	28	4.65
170	*were observed in the*	61	14.16	49	8.15
171	*at the same time*	61	14.16	45	7.48
172	*may be due to*	61	14.16	54	8.98
173	*was significantly higher in*	60	13.93	43	7.15
174	*that the presence of*	60	13.93	39	6.48
175	*study was to evaluate*	60	13.93	51	8.48
176	*seroprevalence of toxoplasma gondii*	60	13.93	16	2.66
177	*the area under the*	60	13.93	38	6.32
178	*and the number of*	60	13.93	49	8.15
179	*was supported by the*	60	13.93	58	9.65
180	*not included in the*	60	13.93	55	9.15
181	*was found to be*	60	13.93	38	6.32