ATLANTIS
Journal of the Spanish Association of Anglo-American Studies
47.2 (December 2025): 25-46
e-ISSN 1989-6840
DOI: http://doi.org/10.28914/Atlantis-2025-47.2.2

Content Words and Compound Structures in English-Written Astrophysical Research Paper Abstracts (2011-2021)

David Israel Méndez
david.mendez@ua.es
Universidad de Alicante

M. Ángeles Alcaraz
m.alcaraz.ariza@gmail.com

The present research focuses on a series of content words (nouns, adjectives and mathematical symbols) as well as compound groups retrieved from 220 research paper abstracts published from 2011 to 2021 in leading English-written astrophysics journals. Our main finding was a high lexical density due to content words appearing more than twice as often as function words, although our results were lower than those obtained in previous studies on scientific letters and research paper titles in the same field. We also found a low syntactic complexity and a predominance of adjectival compound groups over nominal ones, an outcome which clearly contradicts the results obtained in other disciplines or even in popular science astrophysics titles. From a cross-journal perspective, our results indicate that the use of complex compound structures may be intrinsically related not only to whether English is the L1 or L2 of the author(s)/researcher(s) but also to the US-American or European areas of influence which they belong to, i.e. geographical, social, political and economic backgrounds may have a crucial influence on the process of scientific communication.

Keywords: abstracts; research papers; astrophysics; content words; compound structures; cross-journal comparison

. . .

Palabras de contenido y estructuras compuestas en resúmenes de artículos de investigación escritos en inglés en astrofísica

En este artículo hemos analizado diversas palabras de contenido (sustantivos, adjetivos y símbolos matemáticos) así como grupos compuestos en 220 resúmenes de artículos de investigación publicados entre 2011 y 2021 en las revistas más prestigiosas de astrofísica escritas en inglés. Hay una elevada densidad léxica ya que el porcentaje de palabras de contenido es superior al doble que el de palabras funcionales, aunque es inferior al obtenido en estudios anteriores sobre títulos de cartas científicas y trabajos de investigación en el mismo campo. Asimismo, la complejidad sintáctica es baja y los grupos compuestos adjetivales predominan sobre los nominales, algo que contradice claramente los resultados obtenidos en otras disciplinas e incluso en los títulos de divulgación científica en astrofísica. En una comparación entre revistas, hemos observado que el uso de estructuras compuestas complejas puede estar intrínsecamente relacionado no solo con el uso de la lengua inglesa como primera o segunda lengua de los investigadores, sino también con las áreas de influencia estadounidenses o europeas a las que pertenecen, es decir, los antecedentes geográficos, sociales, políticos y económicos pueden tener una influencia crucial en el proceso de comunicación científica.

Palabras clave: resúmenes; artículos de investigación; astrofísica; palabras de contenido; estructuras compuestas; comparación entre revistas

1. Introduction

Weingart (2002) proposed that science is probably the fastest growing enterprise in western societies because they invest a large part of their resources in the production, revision and verification of knowledge. The exponential growth of science implies that it is not only important to carry out research that results in innovative advances, but also to publicise the results of the progress achieved so that science continues to move forward. The spread of these achievements can occur orally through conferences or by speaking directly to colleagues from the same branch of knowledge but, above all, it is done through written media such as specialised journals that help disseminate knowledge and "foster interaction, professional development, and the acquisition of discourse community membership" (Tankó 2017, 42). The journals selected for this work carry various types of articles, among them research papers (RPs), which have become the most often chosen format not only for the continuous training of scientists but also for the dissemination of new knowledge within the scientific and academic community all around the world (Swales 2004).

RPs are comprised of several sections, among them abstracts (and also titles) which play a fundamental role because they inform editors, reviewers and readers giving a brief outline of the methodology employed and the key findings. In other words, abstracts help editors decide whether the subject addressed in the research fits within the journal's scope and, therefore, its readers' interests, and whether to send the corresponding contribution to reviewers for their critical examination (Huckin 2001). Because of the interdisciplinary nature of research these days, along with the ever increasing information flow in the academic world and the tremendous growth in the number of periodicals published (Ventola 1994; Larsen and von Ins 2010; Barbic et al. 2015), which is thanks in large part to the rise of electronic publishing, readers use titles and abstracts to filter the existing literature (Hartley and Benjamin 1998) in order to select what to read so as to keep up-to-date with the latest advances in their field of interest. Similarly, the abstract is the only part of a paper that is published in conference proceedings (Andrade 2011) and conference organisers rely on them to accept or reject papers (Lorés-Sanz 2004).

The importance of abstracts in scientific investigation has thus provoked an impressive amount of research in nearly all disciplines (including linguistics and applied linguistics, chemistry, civil and mechanical engineering, economics and applied economics, educational sciences and technology, experimental and social sciences, literature, medicine, pharmacology, psychology and sociology).

As in every text, within the abstract we find both content and function words; the two categories of vocabulary being defined according to their usage and meaning. Function words are grammatical units that cannot be isolated from other words while content words are lexical units that can act on their own. Content words are used to express meaning and function words signal the structural relationships that words have to one another within sentences. Content words are characterised by their openness since their lexicon can incorporate new words (technical terms, neologisms or existing words that have been given new meanings, adoptions and/or adaptations of foreign words, etc.) or coin new terms through compounding processes. Languages can also remove words that have become obsolete, or nearly obsolete, as a result of the constant advancement of technological innovations, some quite recent examples being 'pager'/'beeper', 'fax', 'walkman' and 'VHS'. By contrast, function words are a closed class of elements that does not admit new created words.

Nevertheless, function and content words should be seen as forming a continuum rather than two different categories because sometimes they share characteristics. A very interesting collected volume on the grammatical behaviour of these overlapping classes is that edited by Corver and van Riemsdijk (2013 [2001]), who refer to them as 'semi-lexical' categories.

That said, the dichotomy of content/lexical categories vs. function/structure categories has been a very useful criterion that is followed in grammar, lexicography, linguistics, psycholinguistics and language acquisition. Studies on language disorders (aphasia, autism, stuttering, etc.) have also been based on content/function words (Bird et al. 2002; Yoder 2006; Alyahya et al. 2021 to name but a few).

In addition to the presence of content and function words, whose relative proportions account for the lexical density (or informativity) of written texts, their syntactic complexity and semantic richness may be studied in terms of the number of compound groups (CGs) they contain. CGs are used to express the relationships between different concepts and consist of sets of content words—without any linking function words—which are grouped together around a given nucleus, this nucleus being preceded by another or other nouns and/or adjective(s) that modify it. Since compounding is one of the most productive devices for coining new words in many languages, among them English (Plag 2018 [2003]), because of its synthesising nature (Sapir 1971 [1921]; Saussure 1974 [1916]), it is not surprising that many scholars, including Granville Hatcher (1960), Downing (1977), Bauer (2012 [1983]), Biber and Gray (2010), and Carrió-Pastor (2008), among others, have tackled the subject within different theoretical frameworks (generative, semantic, descriptive, and language acquisition, respectively). Apart from in general language, composition processes have also been examined in fields such as medicine (Salager-Meyer 1985), architecture (Soneira-Beloso 2015), engineering (Fries 2017; Komaromi and Jerković 2021), and photography (Mykytka 2020).

2. Purpose

A quick look on Google Scholar will suffice to find many studies on abstracts but, to our knowledge, there is however a branch of science in which abstracts have either been underrepresented or overlooked, and it is that of astrophysics.

The scarcity of linguistic research in this domain propelled us to launch a large research project on RP abstracts (RPAs), which started with a pilot diachronic analysis of the linguistic and authorial implications in a corpus of RPAs retrieved from *Monthly Notices of the Royal Astronomical Society*, one of the leading journals in the field (Méndez and Alcaraz 2020).

In the present paper we plan to extend our pilot study to three other high-ranking international astrophysics journals written in English, namely *Astronomy and Astrophysics*, *The Astronomical Journal* and *The Astrophysical Journal*. More precisely, and in order to study the lexical density, syntactic complexity and semantic richness of astrophysical abstracts, the current research focuses on a series of content words (nouns, adjectives and mathematical symbols) as well as CGs retrieved from RPAs published within a given time (2011-2021), in the four journals mentioned above. Other content words such as adverbs, verbs and function words will be addressed only in passing but will be approached in depth in future research. Although the time period analysed is fairly short, we thought that it would be interesting to carry out a somewhat restricted 'diachronic' study as well as a cross-journal comparison in order to obtain more in-depth results.

3. Corpus

Regarding the choice of the journals, we adopted a four-step methodology and selected publications which fulfilled the following criteria:

(i) Be one of the most authoritative astrophysics journals
(ii) Publish papers on observational data and/or theoretical analyses
(iii) Have a high impact factor
(iv) Be freely accessible on-line

Following these criteria, the journals selected were *Astronomy & Astrohpysics*, *The Astronomical Journal*, *Monthly Notices of the Royal Astronomical Society* and *The Astrophysical Journal*.

Astronomy & Astrophysics (A&A), with an impact factor of 6.5 (2022), is an originally European-based journal that publishes papers on theoretical, observational and instrumental astronomy and astrophysics. Back in 2001, the words "A European Journal" were removed from the front cover of A&A because the journal was becoming increasingly global in scope. A&A is published and distributed by EDP Sciences (Édition Diffusion Presse Sciences) on behalf of the European Southern Observatory. EDP Sciences was acquired in 2019 by China Science Publishing & Media.

The Astronomical Journal (AJ), with an impact factor of 5.3 (2022), primarily publishes papers on astronomical research.

Monthly Notices of the Royal Astronomical Society (MNRAS), with an impact factor of 5.2 (2023), covers research on astronomy and astrophysics and is published on behalf of the Royal Astronomical Society. MNRAS is often preferred by astronomers from the United Kingdom and the Commonwealth.

The Astrophysical Journal (ApJ), with an impact factor of 4.9 (2022), has a more theoretical focus and publishes papers on astronomy and astrophysics.

Both AJ and ApJ are US-based and are published on behalf of the American Astronomical Society.

Following researchers such as Banks (2005), Belcher (2005), Flowerdew (2005) and Nesi (2013), among others, who highlighted the value and significance not only of conducting pilot studies but also of working with relatively small linguistic corpora, we randomly collected our RPAs from the accessible on-line version of the four selected journals. Banks (2005) also claimed that small-scale studies can act as pilot studies for future large-scale research.

Furthermore, and to have a more diversified corpus, our analysis covers a period extending from the year 2011 to the year 2021, i.e. 11 years, at the rate of five abstracts per year and journal, i.e. 55 abstracts per journal, which gives a total of 220 abstracts.

RPAs were randomly selected from the publicly accessible online editions of the four target journals. To ensure temporal balance in the corpus, we employed a systematic sampling method. In 2011, five RPAs were selected from issues published in January, March, May, July, and September. In 2012, the selection shifted to February, April, June, August, and October. In 2013, November and December replaced two of the previously sampled months, while still maintaining a total of five RPAs per year per journal. After completing this full cycle, the selection pattern was restarted.

In the case of A&A, since RPs are grouped by topic within astrophysics (cosmology, extragalactic astronomy, stellar structure, planets, etc.), we tried to choose articles which covered all of them. Moreover, it is worth mentioning that the only structured RPAs are those published in A&A in the sense that they must contain the following parts: Context, Aims, Methods, Results and Conclusions. By contrast, RPAs in AJ, ApJ and MNRAS consist of a single paragraph.

4. Methodology

After collecting our corpus, we counted the total number of words that make up each of the RPAs. The task was not so easy because as well as words separated by blank spaces—which do not present any difficulty when counting—the RPAs contain other types of words: hyphenated words, abbreviations (made up of words but also containing numbers) and complex expressions including words, numbers, chemical and/or mathematical symbols, which are more difficult to count. It therefore became obvious that we had no other option than to count the words manually because commonly-used concordance programs would not have helped us at all. As such, the compound 'spin-down' (AJ, January 2011) was counted as two words, whereas the abbreviation 'AGNs' (< 'Active Galactic Nuclei') (A&A, January 2020) was registered as three words in line with the number of its semantic components. Acronyms—i.e. strings of letters with a syllabic structure that is usually pronounced as one word and not letter-by-letter like abbreviations—such as 'CARMA' [< 'Combined Array (for) Research (in) Millimeter (wave) Astronomy'] (AJ, September 2016) and '2MASS' [< 'Two Micron All-Sky Survey'] (AJ, March 2011) were considered as one single word. A more complicated expression containing numbers, abbreviations (including common and proper names together with chemical and mathematical symbols) like the following one was computed as 16 words:

"log $M_{GC}/M_{\odot} \approx 4.50+2.17Fe/H+1.30$" (AJ, July 2019)

[log (< common *logarithm*); M_{GC} (< *Mass Globular Cluster*); / (< mathematical symbol of division); $M_{\odot}$ (< *Mass* of the *Sun*); ≈ (< mathematical symbol of 'almost equal to'); + (< mathematical symbol of 'plus'); *Fe* (< chemical symbol of iron); / (< mathematical symbol of division); *H* (< chemical symbol of hydrogen); + (< mathematical symbol of 'plus')]

Using our counting methodology, our whole sample amounts to a total of 57,972 words (see table 1), distributed as follows: A&A (17,076 words), ApJ (14,341 words), MNRAS (13,517 words) and AJ (13,038 words).

Taking into account that distinguishing between content and function words is essential when estimating abstract informativity, or lexical density, which is the first point that we wanted to address in our study, we divided the total running words into the two different categories. Moreover, we also registered the individual numbers of nouns, adjectives and mathematical symbols per year and journal.

At this point, it is important to mention that in English the border between some content and function words is not clearly defined from either a grammatical, semantic or syntactic perspective, as illustrated in the following examples that belong to three different abstracts from our corpus:

(1) "The physical structure of hot molecular cores, where forming massive stars have heated *up* (adverb) dense dust and gas, but have not yet ionized the molecules, poses a prominent challenge in the research of high-mass star formation and astrochemistry." (A&A, December 2011).

In this first example, the word "up" works as an adverb and so we counted it as a content word.

(2) "Masses of isolated giant stars *up* (preposition) to now were only estimated from evolutionary tracks, which led to different results depending on the physics considered." (A&A, March 2012).

Here the word "up" works as a preposition and so we counted it as a function word.

(3) "The model parameters are degenerate, so we provide relevant information for follow-*up* (noun) observations, which are suggested in order to place further constraints on this unique Kepler object." (AJ, October 2020).

In this last example, the word "up" works as noun and we counted it as a content word.

References to chemical elements, which we have included in our counting of nouns and which are symbolically represented by one or two letters ('H' < hydrogen, 'Fe' < iron or 'He' < Helium), will be covered separately in a future investigation focusing on abbreviations.

The second goal of our research is related to syntactic complexity and semantic richness and for this we recorded the number of CGs. For the purpose of our study, we divided CGs into nominal groups and adjectival groups. Nominal compounds (CNs) are groups of words containing at least two words that function as nouns while adjectival compounds (CAs) are groups of words containing at least one word that functions as an adjective. Despite the traditional definition of "CGs" that specifies that they should not contain any function words, in our case it has been necessary to include them on some occasions because without them the registered CG would not make sense. As an example, this is the case with the compound noun 'g′-i′ versus i′ color-magnitude diagram' (A&A, April 2011) where the preposition 'versus' is necessary to form the CG. In this respect, it is interesting to note that these structures may also include collocations such as 'emission lines' (A&A, July 2011), which is a very well-known nominal compound within the astrophysics community at large. Moreover, the members of some CGs may be written either separated by a blank space or linked by a hyphen depending on the preference or style of the author(s) or journal. An example of this is the compound noun 'surface gravity' (AJ, December 2012) which also appears with a dash ('surface-gravity', AJ, March 2019). The example clearly reflects the erratic nature of the English language already pointed out by Bauer (1998) and Lieber and Štekauer (2012).

In addition to all this counting methodology, we also computed the number of nouns, adjectives, mathematical symbols, CGs, CNs and CAs per number of words for each RPA and then obtained the mean values per year and journal.

Finally, to complement our quantitative analysis and to determine whether the differences observed in the numerical variables studied were statistically significant or not, we also submitted our data to parametric Student's t-tests with a significance level of 0.05. That is, our study combined both manual and computer analyses.

5. Results and Discussion

5.1. Lexical Density (Informativity)

Table 1 summarises the number and percentage of content and function words per journal in the whole period analysed.

Table 1. Number and percentages of content and function words per journal

Journal	No. of Words	No. of Content Words	No. of Function Words
A&A	17,076	11,280 (66.06%)	5,796 (33.94%)
ApJ	14,341	9,773 (68.15%)	4,568 (31.85%)
MNRAS	13,517	9,144 (67.65%)	4,373 (32.35%)
AJ	13,038	8,845 (67.84%)	4,193 (32.16%)
Total	57,972	39,042 (67.35%)	18,930 (32.65%)

As can clearly be seen, RPAs have a high lexical density since the proportion of content words is more than double that of function words. This should not be surprising as an abstract is no more than a short version of an article and its main function is to give readers a complete, yet concise, understanding of the research and findings. From a cross-journal point of view, we can see that ApJ has the highest lexical density and A&A the lowest, with AJ and MNRAS occupying intermediate positions in the scale of lexical density.

Continuing with informativity, it can be added that in terms of diachronicity, the whole period studied presents erratic behaviour, the year 2020 having the highest informativity (68.95%) and the year 2019 the lowest (65.31%). Nevertheless, it must be taken into account that these data only provide a small glimpse into the overall scenario due to the short period analysed.

Focusing specifically on the content words that make up our sample, their breakdown by category is shown in Figure 1.

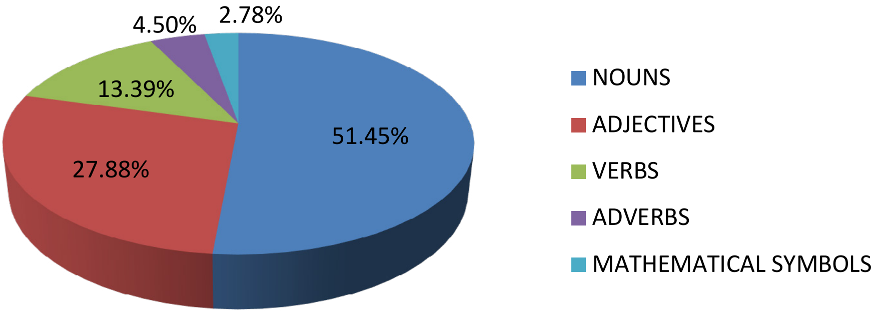

FIGURE 1. Breakdown of content words in the corpus

Figure 1 shows that the highest percentage of content words corresponds to nouns and the lowest to mathematical symbols, which is not surprising as the greater information load is usually provided by the former. From a cross-journal standpoint, the highest percentage of nouns out of the total number of content words is found in ApJ (52.22%) and the lowest in AJ (50.66%) while A&A and MNRAS show similar percentages (51.49% and 51.35%, respectively). With reference to adjectives, their percentages are as follows: MNRAS (28.18%), AJ and ApJ (28.15% each) and A&A (27.19%). Conversely, A&A contains the highest percentage of verbs (14.39%) and ApJ the lowest (12.30%), with MNRAS and AJ having 13.75% and 12.96%, respectively. The reason why A&A RPAs contain more verbs may be due to their format of being structured into variously-focused paragraphs. From a diachronic point of view, the highest percentage of verbs is obtained in 2018 (14.44%) and the lowest in 2020 (12.46%). As for adverbs, AJ contains the highest percentage (4.92%), followed by A&A (4.83%), MNRAS (4.22%) and ApJ (4%). In terms of diachronicity, the highest percentage of adverbs is found in 2021 (5.53%) and the lowest in 2018 (3.59%).

In relation to the mathematical symbols retrieved in the four journals, the percentages of the total number of content words, in decreasing order, are: ApJ (3.33%), AJ (3.31%), MNRAS (2.51%) and A&A (2.10%). From a diachronic standpoint, the year 2013 has the highest percentage of mathematical symbols (4.31%) and the year 2020 the lowest (2%).

If we compare the actual lexical density of the RPAs analysed in the present study with that of the titles of astrophysical scientific letters (SLs) and RPs found in one of our previous articles (Méndez and Alcaraz 2017), we can say that RPAs show a lower level of informativity: 67.35% in RPAs compared to 74.36% in SL titles (SLTs) and 76.02% in RP titles (RPTs). This comes as no surprise since titles, of either SLs or RPs, are more condensed than RPAs and therefore need to carry a higher information load. On the other hand, the lexical density obtained in our study is very similar to that observed in the titles of popular science astrophysics papers (67.7%) (Méndez and Alcaraz 2015). Nevertheless, we have to take into account that the titles analysed in the 2017 article covered the period 2000-2015 and did not include the journal AJ, which does not publish SLs, whereas the 2015 corpus included all the titles related to astrophysical matters published in *Scientific American Magazine* between 1990 and 2014.

5.2. Nouns

Coming back to the most important content words, i.e. nouns, figure 2 shows the cross-journal mean number of nouns per number of words in all the RPAs.

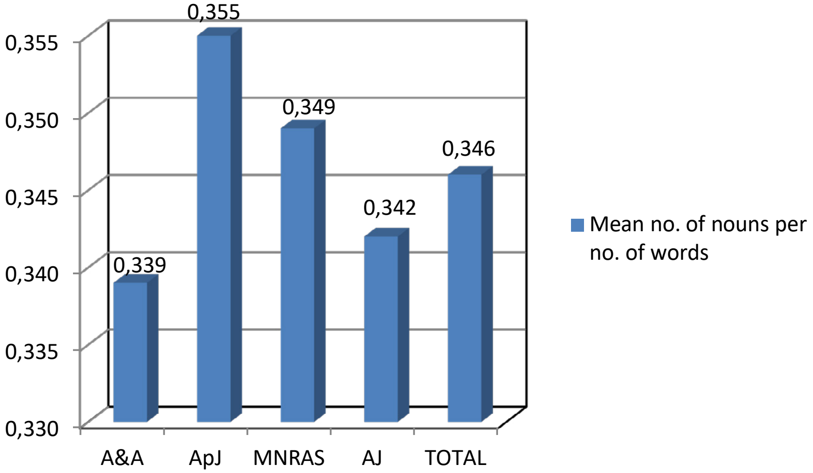

FIGURE 2. Cross-journal mean number of nouns per number of words in RPAs

As we can see in figure 2, the journal with the highest mean number of nouns per number of words is ApJ, while A&A has the lowest, although there are no statistically significant differences between the journals studied. Diachronically speaking, the highest mean number of nouns per number words is reached in 2011 and 2016 (both 0.353) while the lowest corresponds to 2013 (0.334). The differences, which barely account for 5% of variation, are also not statistically significant.

5.3. Adjectives

In figure 3, the cross-journal mean number of adjectives per number of words in all the RPAs is shown.

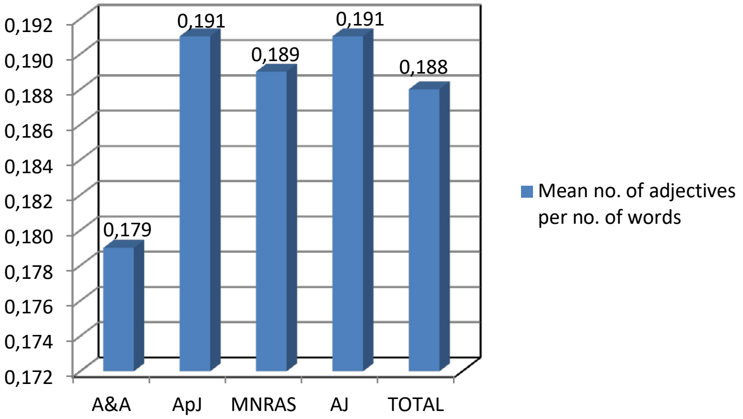

FIGURE 3. Cross-journal mean number of adjectives per number of words in RPAs

The highest mean number of adjectives per number of words is found in ApJ and AJ while A&A has the lowest. Similar to the case of nouns, there are no statistically significant differences between the four journals included in our corpus.

In terms of diachronicity, the highest mean number of adjectives per number of words is reached in 2020 (0.206) and the lowest in 2018 (0.174). The differences in the period under study are considerably higher than in the case of nouns as they amount to more than 15%. Nevertheless, the values are not statistically significant except for the time spans 2018-2020 (p=0.007) and 2019 (0.176)-2020 (p=0.023).

5.4. Mathematical Symbols

Figure 4 depicts the cross-journal mean number of mathematical symbols per number of words in all the RPAs.

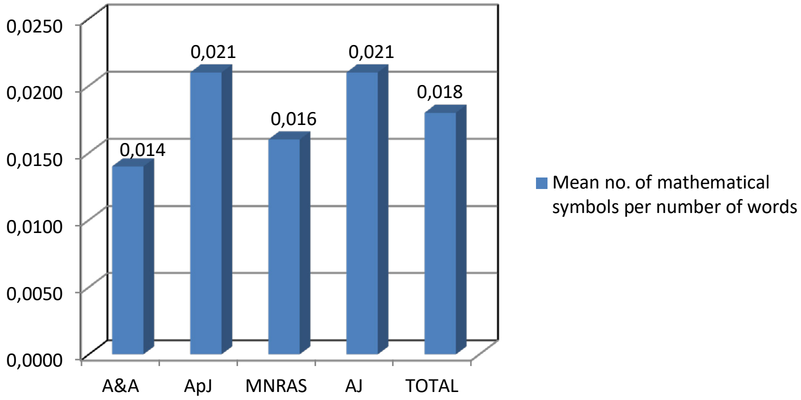

FIGURE 4. Cross-journal mean number of mathematical symbols per number of words in RPAs

As in the case of adjectives, ApJ and AJ contain the highest mean numbers of mathematical symbols per number of words while A&A and MNRAS show the lowest, with the differences between AJ and A&A being statistically significant (p=0.037). From a diachronic standpoint, differences are important since they can increase by up to 60%. The highest mean number of mathematical symbols per number of words is reached in the year 2013 (0.030) and the lowest corresponds to the year 2020 (0.013), the difference between the two periods being statistically significant (p=0.008). Another statistical difference is also shown between the years 2013 and 2021 (0.014) (p=0.014).

Table 2 presents the results of the cross-journal analysis of the number and variants of mathematical symbols found in the corpus under study.

TABLE 2. Cross-journal analysis of the number and variants of mathematical symbols

Journal	No. of math. symbols	No. of variants
A&A	237	21
ApJ	326	20
MNRAS	231	19
AJ	293	20
Total	1087	27

As is clearly seen in table 2, there is a greater total number of mathematical symbols used in ApJ, while the higher number of variants is found in A&A. On the other hand, MNRAS contains both the lowest number of mathematical symbols and fewest variants.

Table 3 shows the number and percentages of the most common mathematical symbols found across the whole sample.

TABLE 3. Number and percentages of the 10 most common mathematical symbols

Math. symbols	No. of appearances
-	166 (15.27%)
~	150 (13.80%)
/	126 (11.59%)
=	122 (11.22%)
+	87 (8%)
%	73 (6.72%)
±	73 (6.72%)
<	71 (6.53%)
x	41 (3.77%)
>	31 (2.85%)

From a cross-journal point of view, the "minus" (-) sign occupies the first place in ApJ and MNRAS while the "similarity" (~) sign makes most appearances in A&A and AJ.

5.5. Syntactic Complexity and Semantic Richness

Focusing henceforth on compound structures, figure 5 plots the mean numbers of CGs, CNs and CAs per number of words in the whole sample.

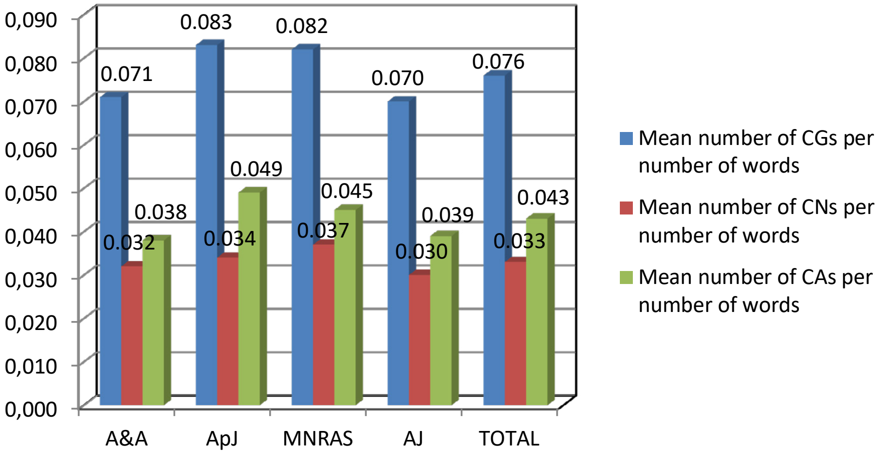

FIGURE 5. Cross- journal mean number of CGs, CNs and CAs per number of words

It is worth noting, as shown in figure 5, the high number of compound structures present in the RPAs. On average, there is approximately one CG for each 13 words, with one CN per 30 words and one CA per 23 words, indicating that CAs are more numerous than CNs across the whole corpus. This same trend is evident for each journal, particularly ApJ.

Below are some examples of the simplest and most complex GCs found in our corpus:

- 2-word CN: "light curves" (AJ, September 2021)
- 9-word CN: "Joint Light-curve Analysis" (JLA) data set. (APJ, August 2016)
- 3-word CA: "stellar metallicity gradient" (A&A, October 2014)
- 39-word CA: "3d4s, 3d^2, 4s^2, 3d4p, 4s4p, 3d5s, 3d4d, 3d5p, 4p^2 and 3d4f configurations" (MNRAS, August 2012)

(In the final example, the letters s, p, d, and f designate the shape of the orbital, which is a consequence of the magnitude of the electron's angular momentum, resulting from its angular motion; the numbers refer to the different levels occupied and the number of electrons in each orbital.)

Figure 5 demonstrates that ApJ has the highest mean number of CGs per number of words, closely followed by MNRAS. On the other hand, AJ and A&A show the lowest such values (0.070 and 0.071, respectively). Statistically significant differences are found between AJ and ApJ (0.083) (p=0.0001), AJ and MNRAS (0.082) (p=0.025), A&A and ApJ (p=0.0007) and A&A and MNRAS (p=0.012). Diachronically speaking, an irregular pattern can be observed, with the maximum value reached in 2020 (0.082) and the minimum in 2013 (0.069) (p=0.014).

As for the mean number of CNs per number of words, the highest value is in MNRAS (0.037) and the lowest one is found in AJ (0.030), the difference being statistically significant (p=0.036). From a diachronic standpoint, we also observe an erratic pattern. The year 2013 (0.024), the year with the lowest value of this indicator, presents statistically significant differences with respect to the years 2011 (0.032) (p=0.05), 2012 (0.035) (p=0.011), 2014 (0.036) (p=0.006), 2015 (0.033) (p=0.04), 2016 (0.035) (p=0.005), 2018 (0.037) (p=0.008) and 2019 (0.041) (p=0.007), 2019 being the year with the maximum value.

In relation to the mean number of CAs per number of words, behaviour is similar to that of the total number of CGs per number of words. ApJ (0.049) has the maximum value, again closely followed by MNRAS (0.045), with AJ and A&A (0.039 and 0.038, respectively) showing the lowest values. Statistically significant differences are again evident between AJ and ApJ (p=0.004), AJ and MNRAS (p=0.05), A&A and ApJ (p=0.0008) and A&A and MNRAS (p=0.014). Diachronically speaking, an irregular pattern is observed, the maximum value being reached in the year 2020 (0.051) and the minimum in the year 2019 (0.036) (p=0.022).

Figure 6 provides the results of a percentage analysis of different types of CGs used in each journal.

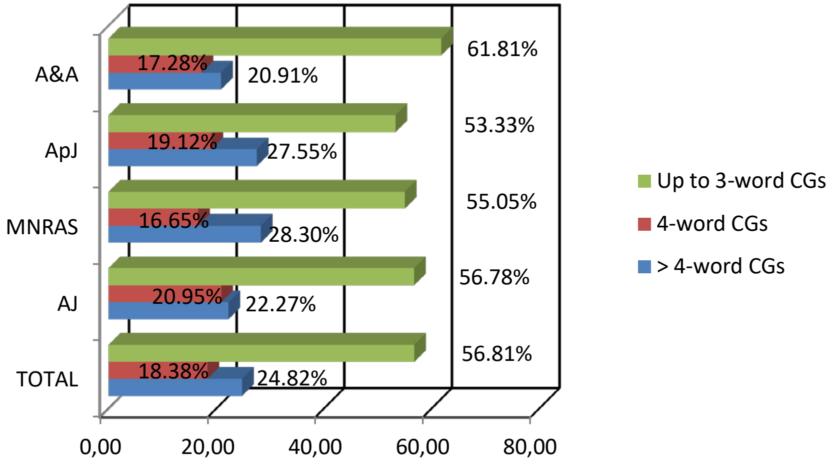

FIGURE 6. Size distribution of CGs in RPAs per journal

As figure 6 clearly illustrates, A&A has the highest percentage of up to 3-word CGs, i.e. the simplest ones, and the lowest percentage of more than 4-word CGs, i.e. the most complex ones. Conversely, ApJ exhibits the lowest percentage of up to 3-word CGs, whereas MNRAS has the highest percentage of more than 4-word CGs. As for 4-word CGs, the highest percentage corresponds to AJ and the lowest to MNRAS.

Focusing specifically on CNs, table 4 provides a more detailed analysis of the number and percentages of the most common types across journals.

TABLE 4. Distribution of CNs across journals

Journal	A&A	ApJ	MNRAS	AJ	TOTAL
No. of 2-word CNs	366 (65.83%)	304 (63.60%)	294 (60.62%)	239 (60.82%)	1203 (62.92%)
No. of 3-word CNs	107 (19.24%)	108 (22.59%)	117 (24.12%)	74 (18.83%)	406 (21.23%)
No. of 4-word CNs	56 (10.07%)	37 (7.74%)	41 (8.45%)	48 (12.21%)	182 (9.52%)
No. of > 4-word CNs	27 (4.86%)	29 (6.07%)	33 (6.81%)	32 (8.14%)	121 (6.33%)
Total no. of CNs	556 (100%)	478 (100%)	485 (100%)	393 (100%)	1,912 (100%)

Table 4 highlights that the largest number of CNs is found in A&A, followed by MNRAS, then ApJ and finally AJ. The 2-word structure tops the frequency scale, far ahead of CNs composed of 3, 4 and more than 4 words. The percentage of 2-word CNs is highest in A&A and lowest in MNRAS. In contrast, A&A contains the lowest percentage of more than 4-word CNs and AJ the highest. In relation to 3-word CNs, the highest percentage corresponds to MNRAS while the lowest is for AJ, while for 4-word CNs, the reverse is true: the maximum percentage corresponds to AJ and the minimum to MNRAS. If we take the sample as a whole, the more complex the CN structures, the lower their presence, a fact that is also observed in each individual journal.

Moving to look at CAs, table 5 shows the number and percentages of the most common types across journals.

TABLE 5. Distribution of CAs across journals

Journal	A&A	ApJ	MNRAS	AJ	TOTAL
No. of 3-word CAs	278 (42.18%)	221 (31.17%)	194 (31.60%)	202 (39.30%)	895 (35.86%)
No. of 4-word CAs	154 (23.37%)	190 (26.80%)	142 (23.13%)	142 (27.63%)	628 (25.16%)
No. of 5-word CAs	103 (15.63%)	102 (14.39%)	101 (16.45%)	79 (15.37%)	385 (15.43%)
No. of 6-word CAs	51 (7.74%)	60 (8.46%)	54 (8.79%)	52 (10.12%)	217 (8.69%)
No. of > 6-word CAs	73 (11.08%)	136 (19.18%)	123 (20.03%)	39 (7.58%)	371 (14.86%)
Total no. of CAs	659 (100%)	709 (100%)	614 (100%)	514 (100%)	2,496 (100%)

Contrary to the case for CNs, ApJ contains the highest number of CAs, followed by A&A, MNRAS and AJ, respectively. With reference to 3-word CAs, i.e. the simplest ones, they are most frequent in A&A and least frequent in ApJ. MNRAS contains the lowest percentage of 4- word CAs but the highest of 5-word CAs, the lowest percentage of 5-word CAs being found in ApJ. The lowest percentage of 6-word CAs is found in A&A, while AJ shows the highest percentages of both 4-word and 6-word CAs. Finally, the highest percentage of more than 6-word CAs, i.e. the most complex adjectival structures, is in MNRAS and the lowest in A&A. Although the situation with CNs is mirrored in as much as the percentage of examples found decreases with the increasing complexity of the CA up to and including 6-word CAs for all journals, when it comes to more than 6-word CAs, for all journals except AJ the percentage increases compared to the 6-word category.

With reference to the mean number of CGs, CNs and CAs per number of words (figure 5), the values are lower in the RPAs studied here than in SLTs and RPTs studied previously (Méndez and Alcaraz 2017). This is most notable in terms of the mean number of CGs and CAs per number of words and, though to a lesser extent, in the mean number of CNs per number words.

Another interesting point to note is that the CG values between A&A, ApJ and MNRAS are positively correlated in SLTs and RPAs: A&A (0.104), MNRAS (0.110), and ApJ (0.111) SLTs, and A&A (0.071), MNRAS (0.082), and ApJ (0.083) RPAs. However, the CG values are negatively correlated in RPTs and RPAs: ApJ (0.110), MNRAS (0.122), and A&A (0.134) RPTs, and A&A (0.071), ApJ (0.073), and MNRAS (0.082) RPAs.

As for the CN values between A&A, ApJ and MNRAS, they are negatively correlated in SLTs and RPAs: ApJ (0.037), MNRAS (0.045), and A&A (0.047) SLTs, and MNRAS (0.030), A&A (0.032), and ApJ (0.034) RPAs. Interestingly, ApJ shows the lowest value in SLTs and the highest one in RPAs.

As in our comparison with SLTs and RPTs, the percentage of CGs with up to three words in RPAs (56.81%) closely matches that in SLTs (55.08%) but exceeds the proportion observed in RPTs (52.33%). CGs consisting of more than four words, however, exhibit a somewhat different pattern: RPAs (24.82%), SLTs (23.65%), and RPTs (29.65%).

With reference to 2-word CNs (see table 4), their percentage is higher in RPAs (62.92%) than in SLTs (58.51%) and RPTs (54.24%) More than 4-word CNs display a different pattern: 6.33% in RPAs, 5.81% in SLTs and 9.32% in RPTs. As regards to 3-word CAs (see table 5), their percentages are very similar across all three text types: 35.86% for RPAs, 35.73% for SLTs and 36.28% for RPTs, whilst CAs composed of more than six words dominate in RPAs (14.86%) compared to RPTs (12.39%) and SLTs (11.57%). The predominance of shorter CGs over longer ones in all the text types studied—which is in line with the findings of Entralgo et al. (2015) in their study of scientific article titles in natural sciences—is, without doubt, due to the aim of informing readers in the clearest and most accurate way in order to conform to the principles of informativity and economy (Bush-Lauer 2000). Moreover, the predominance of longer and more sophisticated CGs might involve a lack of attention, or even a rejection, on the readers' part, especially if they are non-native English speakers due to the effort likely needed to fully decode the CGs.

From a cross-journal perspective (see Table 4), and based on the comparison with previous findings for SLTs and RPTs, A&A shows the highest percentage of two-word CNs in both RPAs (65.83%) and RPTs (64.58%), while ApJ leads in SLTs (61.97%). The highest percentage of more than 4-word CNs is found in AJ RPAs (8.14%) and the lowest one in A&A RPAs (4.86%), whilst A&A has the highest number in SLTs (7.78%) and ApJ in RPTs (12.12%).

Regarding 3-word CAs (see table 5), A&A has the highest percentage in both RPAs (42.18%) and RPTs (40.74%) but ApJ has the highest rate in SLTs (39.58%), although it has the lowest for RPAs (31.17%). The first position when referring to more than 6-word CAs is occupied by the RPAs in MNRAS (20.03%), while for SLTs the highest percentage is found in A&A (15.52%) and ApJ has the highest percentage in RPTs (16.42%), though it has the lowest in SLTs (9.03%).

The consistent predominance of CAs over CNs in RPAs, SLTs, and RPTs, along with the high lexical density, may be attributed to the scientificity and informativity required in astrophysical discourse. In this regard, our findings do not coincide with those of Algeo (1999), who found that CNs are the largest group of compounds in general English, or with those of Mykytka (2020), who found that the nominal pattern is the most common one in the field of photography. Neither do they agree with the predominance of CNs found in our previous articles on popular science astrophysics titles (Méndez and Alcaraz 2015) and on MNRAS RPAs (Méndez and Alcaraz 2020), except for Block D, which corresponds to the 2018 data included in the 2020 paper. We think the discrepancy may be a consequence, not only of the different classifications of CGs used in the studies mentioned but also of the periods of time studied in each. Moreover, the distribution of CNs and CAs does not greatly differ (always smaller than 10%) from that found in Méndez and Alcaraz (2020), the percentages being even more similar to those for Block D.

5.6. Cross-Journal Final Remarks

Although A&A RPAs contain the highest number of words, especially content words, they display the lowest lexical density and the lowest mean numbers of nouns, adjectives, mathematical symbols and CAs per number of words. Conversely, they have the highest percentage of verbs and mathematical symbol variants. Moreover, they show the highest percentages of up to 3-word CGs, 2-word CNs and 3-word CAs. By contrast, they yield the lowest percentages of more than 4-word CGs and CNs and of 6-word CAs. In other words, RPAs in A&A are the longest in spite of their lower semantic informativity and complexity. This can probably be explained in terms of a geographic scenario, where researchers publishing in this journal may be mostly English L2 writers who do not have the same mastery of the English language as native English (L1) speakers, who usually have a greater repertoire of linguistic knowledge (Ruan 2018; Xue and Ge 2021). Moreover, the dramatic increase in the impact factor of A&A—from 4.8 in 2017-2018 to 6.5 in 2021-2022—could also be explained in terms of the increasing contributions from English L2 writers from countries beyond Europe.

AJ RPAs are the shortest, and also have the lowest percentage of nouns as well as the lowest mean numbers of CGs and CNs per number of words. They also exhibit the second lowest mean number of CAs per number of words. Moreover, they have the highest percentage of adverbs and the second highest number of mathematical symbols (after ApJ) and the highest mean numbers of adjectives and mathematical symbols per number of words (shared with ApJ). These results fit with a scenario of English L1 researchers that are mostly concerned with technical aspects and therefore do not require either a high load of semantic information or a large number of complex structures to express their findings.

ApJ RPAs have the second highest number of words, together with the highest lexical density and the highest mean numbers of nouns, adjectives and mathematical symbols (the last two shared with AJ RPAs) per number of words. Furthermore, they yield the highest mean numbers of CGs and CAs per number of words but the lowest percentages of verbs and adverbs and of up to 3-word CGs. In addition, they show a high percentage of more than 4-word CGs (the second highest after MNRAS RPAs) and the lowest percentages of 4-word CNs and 3- and 6-word CAs.

MNRAS RPAs have the highest percentage of adjectives and the lowest number of mathematical symbols and variants. They contain the highest mean number of CNs per number of words and the second highest mean numbers of nouns, CGs and CAs (after ApJ) per number of words. They also yield the lowest percentage of 4-word CGs together with the highest percentage of more than 4-word CGs as well as the second highest percentage of more than 4-word CNs (after AJ) and the highest percentage of more than 6-word CAs. Additionally, the mean numbers of CGs, CNs and CAs per number of words are similar, although slightly higher than those observed in Block D (Méndez and Alcaraz 2020). The results observed in RPAs in ApJ and MNRAS are in accordance with a geographic scenario of English L1 researchers who know how to use more complex structures together with a higher load of informativity when writing their RPAs. It is also worth mentioning that the greater values of the mean numbers of adjectives per number of words (see figure 3) found in MNRAS, AJ and ApJ RPAs in comparison with A&A RPAs could imply that English L1 researchers usually give a more accurate description of the facts, adjectives being words that modify or describe nouns in the sense that they provide more specific information about them.

Furthermore, the highest mean numbers of mathematical symbols per number of words being found in AJ and ApJ RPAs would suggest that US-based astrophysicists tend to use a more mathematically-oriented and more direct language to arrive clearly and explicitly to the point. In contrast, the lower presence of mathematical symbols in the RPAs published in the two European-based journals (A&A and MNRAS) may suggest a more narrative discourse, which could mean that readers who are not so familiar with mathematical issues would have less difficulty in understanding what is being communicated.

6. Conclusions and Implications

First of all, it is worth mentioning that from a diachronic standpoint, we have found an erratic pattern for all the variables analysed, with no notable statistically significant differences between the different years of the sample. Studying a longer period of time would probably be needed—as in Méndez and Alcaraz (2020)—in order to obtain more conclusive results.

Although our study was based on a series of content words (nouns, adjectives, mathematical symbols) as well as CGs registered in a limited sample of abstracts published in the four principal English-written astrophysical journals (A&A, AJ, MNRAS and ApJ) during a short period of time (2011-2021), it has shed some light on in-depth ethnological issues, i.e. on the influence of geographical, social and political contexts when communicating the results of astrophysical research. In other words, it is likely that astrophysicists express their ideas and conclusions in a very different way if they are English L1 or L2 writers, but also if they belong to the US-American or European areas of influence. To corroborate the assertions made here, it would be interesting to conduct analyses of larger corpora retrieved from longer time sets.

If we take into account that novice researchers would probably begin to approach science mainly through the reading of abstracts, the study of the possible influence of the characteristics of the language used when writing them, and especially the content words they contain, could become a key strategy when analysing the evolution of astrophysics itself. Moreover, expanding this type of research to other scientific disciplines and to other genres such as RPs, reviews, editorials, etc., which constitute a more comprehensive and accurate communication of knowledge, could illustrate the challenging and thought-provoking idea of a possible correlation between how science is conveyed and its advancement over time.

Works cited

Algeo, John. 1999. *Fifty Years among the New Words: A Dictionary of Neologisms, 1941-1991*. Cambridge: Cambridge UP.

Alyahya, Reem S. W., Paul Conroy, Ajay D. Halai and Matthew A. Lambon Ralph. 2021. "An Efficient, Accurate and Clinically-applicable Index of Content Word Fluency in Aphasia." *Aphasiology* 36 (8): 921-39.

Andrade, Chittaranjan. 2011. "How to Write a Good Abstract for a Scientific Paper or Conference Presentation." *Indian Journal of Psychiatry* 53 (2): 172-75.

Banks, David. 2005. "The Case of Perrin and Thompson: An Example of the Use of a Mini Corpus." *English for Specific Purposes* 24 (2): 201-213.

Barbic, Skye, Karen Roberts, Zachary Durisko and Cheolsoon Lee. 2015. "Readability Assessment of Psychiatry Journals." *European Science Editing* 41 (1): 3-11.

Bauer, Laurie. (1983) 2012. *English Word-Formation*. Cambridge: Cambridge UP.

—. 1998. "When Is a Sequence of Two Nouns a Compound in English?" *English Language & Linguistics* 2 (1): 65-86.

Belcher, Diane. 2005. "Editorial." *English for Specific Purposes* 14: 119-121.

Biber, Douglas and Bethany Gray. 2010. "Challenging Stereotypes about Academic Writing: Complexity, Elaboration, Explicitness." *Journal of English for Academic Purposes* 9 (1): 2-20.

Bird, Helen, Sue Franklin and David Howard. 2002. "'Little Words'—Not Really: Function and Content Words in Normal and Aphasic Speech." *Journal of Neurolinguistics* 15 (3-5): 209-37.

Busch-Lauer, Ines. 2000. "Titles of English and German Research Papers in Medicine and Linguistics Theses and Research Articles." In Trosborg 2000, 77-94.

Carrió-Pastor, María Luisa. 2008. "English Complex Noun Phrase Interpretation by Spanish Learners." *Revista Espanola de Linguistica Aplicada* 27-44.

Čmejrková, Světla, František Daneš and Eva Havlovát, eds. 1994. *Writing vs. Speaking. Language, Text, Discourse, Communication*. Tübingen: Gunter Narr.

Corver, Norbert and Henk van Riemsdij. (2001) 2013. *Semi-lexical Categories: the Function of Content Words and the Content of Function Words*. Berlin: Mouton de Gruyter.

DOWNING, Pamela. 1977. "On the Creation and Use of English Compound Nouns." *Language* 53 (4): 810-42.

ENTRALGO, Johanna, Françoise Salager-Meyer and Marianela Luzardo Briceño. 2015. "¿Cuán gramaticalmente complejos son los títulos de los artículos científicos en las ciencias naturales?" *Revista de Lenguas para Fines Específicos* 21 (2): 70-97.

FLOWERDEW, Lynne 2005. "An Integration of Corpus-based and Genre-based Approach to Text Analysis in EP/ESP: Countering Criticisms against Corpus-based Methodologies." *English for Specific Purposes* 25 (3): 321-332.

FRIES, Marie-Hélène. 2017. "Teaching Compound Nouns in ESP: Insights from Cognitive Semantics." In Sarré and Whyte 2017, 93-109.

GRANVILLE HATCHER. Anna. 1960. "An Introduction to the Analysis of English Noun Compounds." *Word* 16: 356-73.

HARTLEY. James and Michelle Benjamin. 1998. "An Evaluation of Structured Abstracts in Journals Published by the British Psychological Society." *British Journal of Educational Psychology* 68: 443-56.

HEWINGS, Martin, ed. 2001. *Academic Writing in Context: Implications and Applicatio*ns. Birmingham: UBP.

HUCKIN, Thomas. 2001. "Abstracting from Abstracts." In Hewings 2001, 93-103.

KOMAROMI, Bojana and Jelena Jerković. 2021. "Variation in the Translation Patterns of English 'noun + noun' Compounds in ESP: The Case of Engineering Students." *English Language and Literature Teaching* 18 (2): 167-184(230).

LARSEN, Peder Olesen and Markus von Ins. 2010. "The Rate of Growth in Scientific Publication and the Decline in Coverage Provided by Science Citation Index." *Scientometrics* 84 (3): 575-603.

LIEBER, Rochelle and Pavold Štekauer, eds. 2012. *The Oxford Handbook of Compounding*. Oxford: Oxford UP.

LORÉS-SANZ, Rosa. 2004. "On RA Abstracts: From Rhetorical Structure to Thematic Organization." *English for Specific Purposes* 23 (3): 280-302.

MÉNDEZ, David I. and M. Ángeles Alcaraz. 2015. "Astrophysics Titles in Scientific American Magazine (1990-2014): Linguistic and Discourse Practices". *International Journal of Applied Linguistics and English Literature* 4 (6): 39-51.

—. 2017. "Titles of Scientific Letters and Research Papers in Astrophysics: A Comparative Study of Some Linguistic Aspects and their Relationship with Collaboration Issues". *Advances in Language and Literary Studies* 8 (5): 128-139.

—. 2020. "Research Paper Abstracts in *Monthly Notices of The Royal Astronomical Society* (1943-2018). A Diachronic Approach Focusing on Linguistic and Authorial Implications". *English Text Construction* 13 (1): 62-83.

MYKYTKA, Iryna. 2020. "Noun Compounds in Photography." *Atlantis* 42 (2): 72-98.

NESI, Hilary. 2013. "ESP and Corpus Studies." In Paltridge and Starfield 2013, 408-424.

PALTRIGE, Brian and Sue Starfield, eds. 2013. *The Handbook of English for Specific Purposes.* London: Wiley-Blackwell.

Plag, Ingo. (2003) 2018. *Word-Formation in English*. Cambridge: Cambridge UP.

Ruan, Zhoulin. 2018. "Structural Compression in Academic Writing: An English-Chinese Comparison Study of Complex Noun Phrases in Research Article Abstracts." *Journal of English for Academic Purposes* (36): 37-47.

Salager-Meyer, Françoise. 1985. "Syntax and Semantics of Compound Nominal Phrases in Medical English Literature: A Comparative Study with Spanish." *English for Specific Purposes* 95: 6-12.

Sapir, Edward. (1921) 1971. *Language.* London: Rupert Hart-Davis.

Sarré, Cédric and Shona Whyte, eds. 2017. *New Developments in ESP Teaching and Learning Research*. Voillans: Research Publishing.

Saussure, Ferdinand de. (1916) 1974. *Curso de Lingüística General.* Buenos Aires: Losada.

Soneira-Beloso, Begoña. 2015. *A Lexical Description of English for Architecture: A Corpus Based Approach*. Bern: Peter Lang.

Swales, John. 2004. *Research Genres. Exploration and Application*. New York: Cambridge UP.

Tankó, Gyula. 2017. "Literary Research Article Abstracts: An Analysis of Rhetorical Moves and their Linguistic Realizations." *Journal of English for Academic Purposes* 27: 42-55.

Trosborg, Anna, ed. 2000. A*nalysing Professional Genres*. Amsterdam and Philadelphia: John Benjamins.

Ventola, Eija. 1994. "Abstracts as an Object of Linguistic Study". In Čmejrková et al. 1994, 333-352.

Xue, Qing and Tianshuang Ge. 2021. "A Corpus-based Study on Phrasal Complexity in Computer Science Abstracts of Novice and Advanced Writers." *Open Journal of Modern Linguistics* 11 (05): 808-822.

Yoder, Paul. 2006. "Predicting Lexical Density Growth Rate in Young Children with Autism Spectrum Disorders." *American Journal of Speech-Language Pathology* 15 (4): 378-388.

Weingart, Peter. 2002. "The Moment of Truth for Science: The Consequences of the 'Knowledge Society' for Society and Science." *EMBO Reports* 3: 703-706.

Received 07 July 2023 Revised version accepted 29 February 2024

David Israel Méndez holds a Ph.D. in astrophysics from La Laguna University (Canary Islands, Spain) and he is a lecturer in the Department of Physics, Engineering Systems and Signal Theory at the Polytechnic University College of the University of Alicante (Spain). He teaches undergraduate and master courses, both in English and Spanish, in fundamental engineering physics, optics and acoustics. He has published numerous research articles on astrophysics, optics, non-linear oscillations, information science and English linguistic matters in leading international journals.

M. Ángeles Alcaraz is now retired but was a lecturer in the Department of English Studies at the University of Alicante (Spain). She is the author of numerous research articles, conference papers and other publications on the linguistic and pragmatico-rhetorical analysis of English-, French- and Spanish-written specialised discourses (medicine and astrophysics).