

SAL 33/2018

Lexical Word-Class
Distributions in
Research Articles
of Four Subject
Areas

Received 01/2018
Accepted 10/2018

Lexical Word-Class Distributions in Research Articles of Four Subject Areas

Kalbos dalių pasiskirstymas keturių skirtingų mokslo sričių tyrimuose

SOCIOLINGUISTICS / SOCIOLINGVISTIKA

Deny Arnos Kwary

Airlangga University, Indonesia

Almira F. Artha

Airlangga University, Indonesia

Yuni S. Amalia

Lecturer, Airlangga University, Indonesia



<http://dx.doi.org/10.5755/j01.sal.33.0.19945>

This study aimed to determine the lexical word-class distribution in research articles of four subject areas: social sciences, health sciences, physical sciences and life sciences. A total of 5,754,560 tokens or running words were extracted from research articles published by Elsevier for examination. Results show both similarities and differences in distribution across the four subject areas. For health, physical and life sciences, the noun is the most dominant lexical word class, followed by the adjective, verb and adverb. For social sciences, the verb is more dominant than the adjective. This finding reflects that research articles in social sciences use the highest number of words and are more conversational in nature. For types of nouns, singular nouns are used more often than plural nouns in all subject areas; this usage might indicate that research articles tend to focus on a single research object. For types of verbs, research articles in health, life and physical sciences tend to prefer using past tense and past participle forms over others; this usage indicates emphasis on reporting what has been done. In contrast, social sciences research articles show more frequent use of the verbs' base form, and this usage possibly signifies arguments regarding general truths.

KEYWORDS: lexical word class, research article, social sciences, health sciences, physical sciences, life sciences, academic prose.



Introduction

Words are the building blocks of texts, and their composition likely differs from one type of text to another, particularly if texts are from different subject areas. Biber, Conrad, & Reppen (2006) assert the importance of differences in word use across subject areas. Differences are found not only in word families but also in word classes. In general, lexical word classes can be classified into four categories: nouns, verbs, adjectives and adverbs. Biber et al. (1999) have shown differences in the distribution of lexical word classes in academic prose, conversation, fiction, newspapers, press reports, official documents, conversations and prepared speeches. In academic prose, Biber et al. (1999) found that the noun was the most frequently used lexical word class, followed by the adjective, the verb and the adverb.

Academic prose can further be classified based on subject areas, and the distribution of lexical word classes in one subject area may differ from that in another. However, to the best of the author's knowledge, no study has compared the distribution of lexical word classes in different subject areas. Such research is necessary because knowing the distribution of lexical word classes can reveal each subject area's typical characteristics. In turn, teachers of academic writing can also create more focused teaching materials.

The present study uses the corpus linguistics approach to examine lexical word classes. Corpus linguistics research has played a significant role in the study of languages since it analyses vast collections of electronic text aided by computer software (Baker, 2010). Through corpus linguistics, large amounts of text data can be effectively and efficiently processed and elaborated on without any salient errors. In vocabulary research, corpus linguistics has been used to plan, control and monitor vocabulary development (Nation, 2001). In teaching English as a foreign language or teaching English to non-native speakers, corpus linguistics is often utilised to select and prepare teaching materials (Walsh, 2010). Results of studies using the corpus linguistics approach have helped teachers decide the number and kind of words they should teach their students (McCarten, 2007). Accordingly, both teachers and learners can acquire the most relevant and frequently used words during their learning process. Furthermore, results of corpus linguistics research are fundamentally important for assessing the abilities of first- and second-language learners (cf. Schmitt, Schmitt & Clapham, 2001; Cameron, 2002; Qian & Schedl, 2004).

Several studies using the corpus linguistics approach have focused on the use of academic vocabulary, for example, the Academic Word List (AWL) by Coxhead (2000); the New Academic Word List by Browne, Culligan, and Phillips (2013); and the Academic Vocabulary List by Gardner and Davies (2014). These studies focus on academic word lists formulated from academic texts in various disciplines, but they do not focus on a particular subject area or scientific field (Liu & Han, 2015). Several studies have also focused on creating word lists from particular subject fields. One example is the Medical Academic Vocabulary List that only examines words in medical texts (Lei & Liu, 2016). Another example is the Chemistry Academic Word List that provides a word list related to chemistry texts (Valipouri & Nassaji, 2013). Martinez, Beck, and Panza (2009) conducted an even more focused study to create a word list based on academic vocabulary from agricultural research articles.

A number of studies have focused on the use of lexical words in different subject areas. However, no study has been conducted on the distribution of lexical word classes across different subject areas. Biber et al. (1999) initiated a similar study, but they focused on differences in the distribution of lexical word classes in different text genres. Their study can be expanded by focusing on a particular genre, that is, academic prose, which can be further categorised into several subject areas. One source of academic prose is the text

from research articles, which can be classified into four subject areas: social sciences, health sciences, physical sciences, and life sciences. These categories are based on broad classifications of subject areas currently used in the Scopus database. Consequently, there are two main objectives of the present study. The first is to determine the distributions of lexical word classes in research articles of the four subject areas. The second is to compare the types of nouns and verbs used in each subject area. These objectives will be achieved by using the corpus linguistics approach.

Method

This study focuses on the lexical word-class distribution of typical word types across four broad subject areas: social, health, physical and life sciences. The classification is based on Scopus, the largest database of academic literature (Elsevier, 2016). Scopus has indexed a number of publishers, of whom Elsevier is the publisher with the largest data share. In this study, therefore, data were extracted from research articles published in Elsevier journals indexed at Scopus. Elsevier also categorises subject areas into four. Two are exactly the same as those used at Scopus, i.e., health sciences and life sciences. The other two are quite similar, i.e., Elsevier uses the terms 'social sciences and humanities' and 'physical sciences and engineering', whereas Scopus uses the terms 'social sciences' and 'physical sciences'.

In selecting research articles, four criteria were considered. First, only open access research articles were selected. Open access means that research articles can be accessed and downloaded without charge. This was necessary so that collected data was copyright free. Second, research articles must be from journals with a 5-year impact factor, meaning research articles need to be published by journals established for more than 5 years and whose articles have also been cited in other research articles for more than 5 years. Third, the research articles must be written in English, the language used in most international publications. Finally, only research articles published during the last 5 years (2011–2015) were chosen to ensure up-to-date data. Table 1 summarises the data of the corpus collected for this study.

Table 1
Description of the
collected corpus

Sub-corpus	Number of Running Words	Number of Research Articles Selected	Sub-Disciplines based on Elsevier's classifications
<i>Health sciences</i>	1,094,205	246	(1) Medicine and Dentistry, (2) Nursing and Health Professions, (3) Pharmacology, Toxicology and Pharmaceutical Science, (4) Veterinary Science and Veterinary Medicine
<i>Life sciences</i>	987,279	161	(1) Agricultural and Biological Sciences, (2) Biochemistry, Genetics and Molecular Biology, (3) Environmental Science, (4) Immunology and Microbiology, (5) Neuroscience
<i>Physical sciences</i>	2,632,817	366	(1) Chemical Engineering, (2) Chemistry, (3) Computer Science, (4) Earth and Planetary Sciences, (5) Energy, (6) Engineering, (7) Materials Science, (8) Mathematics, (9) Physics and Astronomy
<i>Social sciences</i>	1,040,259	122	(1) Arts and Humanities, (2) Business, Management and Accounting, (3) Decision Sciences, (4) Economics, Econometrics and Finance, (5) Psychology, (6) Social Sciences
Total	5,754,560	895	

As shown in Table 1, each of the four subject areas contains sub-disciplines that follow Elsevier's categorisation, which became this study's data source. The representation of each sub-discipline and each journal within each sub-discipline was ensured in the study's corpus. This means that the corpus comprises research articles from all journals in these various sub-disciplines. The corpus was placed in the website corpus.kwary.net for ease of access in data analysis (Kwary 2018). This website can be freely accessed by anyone interested in conducting research on words used in research articles.

To determine the classification of academic words in each sub-corpus, the software AntWordProfiler (Anthony, 2014) was used to classify all words in a text into the general service list (GSL), the AWL and a list of words not in the GSL and AWL. This study used AntWordProfiler to classify texts in this corpus into words in the GSL and words not in the GSL. Words in the GSL are general words frequently found in a general text (Nation, 2001). Since the focus was academic text in research articles, general words needed to be excluded from the data, an automatic function of AntWordProfiler, which also provides the details of words' range and frequency. The following list explains the steps involved in collecting data using AntWordProfiler.

1. The files of each sub-corpus in AntWordProfiler are opened.
2. The GLS stop lists, i.e., the lists of general words are chosen. This means that general words were automatically deleted, leaving only academic words, technical words and other low frequency words. The list produced from this step was still substantially long, so more words needed to be deleted from the list. The words' range and frequency were the next criteria.
3. For the range, only words that occur in at least two-thirds of the number of sub-disciplines were selected, ensuring the selection of only the words used in most of the sub-disciplines. For example, health sciences has four sub-disciplines (see Table 1); thus, a word's range is a minimum of three (four times two-thirds equals 2.76, rounded up to three). Social sciences has six sub-disciplines (see Table 1); thus, a word's range must be a minimum of four (six times two-thirds equals four) for selection.
4. After selecting words based on range, words were selected based on their frequencies. In this case, only words with a frequency of more than 10% of the total number of research articles in the particular subject area were selected. For example, in health sciences, as the corpus had 246 articles (see Table 1), the words selected had a minimum frequency of 25 (i.e., $246 \times 10\% = 24.6$, rounded to 25).
5. The word list produced from the fourth step was quite long. The next technique allocated a score to each word. The score was calculated as follows: $(\text{range} \times 10) + (\text{frequency})$. All words were then sorted based on their scores. Finally, the top 1,000 words were selected from each sub-corpus.
6. From the top 1,000 words from each subject area, words found in more than one subject area were excluded because this study focused on typical academic words, i.e., words used in only one particular subject area. The number of academic words used only in social sciences, health sciences, life sciences and physical sciences was 329, 323, 273 and 314, respectively.

After obtaining typical word types for each subject area, the words were then run using CLAWS Tagger for English, available at <http://ucrel.lancs.ac.uk/claws>. This tagger automatically determines the word class of each word type. However, since a word type may have more than one word class, some results from the tagger required further analysis to determine their correct word classes by examining how the words were used in context.

Results and Discussion

Distribution of Lexical Word Classes across the Four Subject Areas

As stated in the Methodology section, data collection resulted in the following numbers of typical academic words: 329 for social sciences, 323 for health sciences, 273 for life sciences and 314 for physical sciences. Since these numbers are unequal, percentages needed to be recalculated. The new percentages were used in the discussion on the distribution of word classes (Table 2).

Table 2
Distribution of lexical word classes across four subject areas

No.	Word Class	Social Science		Health Science		Life Science		Physical Science	
		Number	%	Number	%	Number	%	Number	%
1	Adjective	60	18.24%	90	27.86%	52	19.04%	79	23.16%
2	Adverb	7	2.13%	2	0.62%	6	2.21%	7	2.23%
3	Noun	193	58.66%	222	68.73%	178	65.2%	188	59.9%
4	Verb	69	20.97%	9	2.79%	37	13.55%	40	12.73%
Totals		329	100%	323	100%	273	100%	314	100%

As Table 2 shows, the most frequently used lexical word class across the four subject areas is the noun. According to Biber et al. (1999), nouns represent a high density of information. Radford (2009) mentioned that nouns have the semantic property of denoting entities, meaning that the high frequency of nouns in research articles indicates that the articles tend to present dense information and denote a large number of entities. If each subject area is examined, the percentage of nouns is found to be the highest in health sciences (68.73%), followed by life sciences (65.2%), physical sciences (59.9%) and social sciences (58.66%).

Based on the percentages in Table 2, in terms of noun usage, research articles in health sciences were close to those in life sciences and social sciences were close to physical sciences. However, for adjective use, health sciences were close to physical sciences, while social sciences were close to life sciences. For verbs and adverbs use, health sciences were significantly lower than the other three.

Considering the use of nouns, health sciences research articles were inferred to contain a higher density of information and denote more entities than articles from other subject areas. On the contrary, research articles in social sciences were found to contain the least density of information and entities in comparison with those in other subject areas. This finding possibly indicates research articles in health sciences present more focused information and entities, whereas those in social sciences use more explanation to present information. Data in Table 1 reveal that the average number of words in health sciences research articles is lowest, i.e., 4,448 words (1,094,205/246), while the average for social sciences is highest, i.e., 8,527 words (1,040,259/122). These data confirm the conclusion that research articles in health sciences are more straightforward, whereas those in social sciences are wordier.

For the second most frequent word class, differences appeared between one subject area and the rest, particularly between research articles in social sciences and those in the other three subject areas. For social sciences, the verb was the second most frequent word class, but for the other three subject areas, the adjective held the second position.

The high frequency of adjectives in health, life and physical sciences shows that nouns in these subject areas are usually modified by adjectives. In fact, frequent adjective use can be considered a characteristic of an academic text. The results of this study seem to coincide

with those of Biber et al. (1999), who found the adjective to be the second most frequent word class in academic prose. Since research articles can be classified as academic prose, it is reasonable that the adjective holds the second position in the present corpus of research articles.

From the four subject areas, as shown in Table 2, health sciences had the highest percentage of adjectives, i.e., 27.86%. Data show that research articles in health sciences contain unique adjectives that are also typical words in this subject area. Some examples are *renal*, *arterial*, *endothelial*, *gestational*, *myocardial* and *ocular*. In addition, frequent adjective use also shows that nouns in health sciences research articles are often modified to make them more specific. For example, the noun *kidney* has several adjectives as adjacent collocates or in left-neighbour occurrences (see Fig. 1).

Left Neighbours Cooccurrences
chronic (31), acute (17), the (19), of (4), nhl (4), human (4), and (4), with (3), colon (3), rectum (3), bladder (3), brain (2), men (2), or (2), leukaemia (2), each (2), right (2), lung (1), by (1), ne (1), liver (1), abnormal (1), metanephric (1), for (1), excluded (1), chronic (1), in (1), testis (1), increases (1), congenital (1)

Fig. 1

Adjacent collocates of the noun *kidney*

Fig. 1 shows that *kidney* can be modified by the following adjectives: *chronic*, *acute*, *human*, *right*, *abnormal*, *metanephric* and *congenital*. This is only one example of adjectives' ubiquitous use in health sciences, illustrating how health sciences articles use more adjectives than other subject area articles.

In social sciences, however, the verb holds second position as the most frequent word class used. As shown in Table 2, typical word types in the social sciences are 20.97% verbs, but only 18.24% adjectives. That social sciences uses a lesser proportion of adjectives than other subject areas indicates that social sciences research articles modify nouns relatively infrequently. In comparing this result with that of Biber et al. (1999), the distribution of word classes in social sciences research articles is found to be more similar to that of conversation than that of academic prose. Biber et al. (1999) found that in conversation, the verb is the second most frequent word class, while in academic prose, the adjective is the second most frequent, indicating that social sciences research articles are relatively more conversational. The high frequency of verbs in social sciences research articles can also be interpreted as a sign that those research articles contain longer explanations. According to the *Oxford Learner's Dictionary of Academic English* (2014), a verb is defined as 'a word or group of words that expresses an action (such as eat), an event (such as happen) or a state (such as exist)'. In addition, Radford (2009) stated that verbs have the semantic property of denoting actions or events, thus indicating that social sciences research articles tend to contain more explanations about actions and events. Since a verb also comprises the main part of a sentence's predicate, the greater number of verbs may indicate a greater number of sentences. These greater numbers can again be related to the average number of words (8,527) in social sciences research articles being highest, in comparison with the average number of words in the remaining subject areas.

In comparing this study's results with those of Paquot (2010), it is found that the word class distribution of social sciences research articles is quite similar to that of the Academic Keyword List (AKL). Paquot (2010) showed that of the 930 words in the AKL, 38.17% are nouns, 25.05% are verbs and 19.35% are adjectives. This means that the top three word

classes, in order of frequency, are nouns, verbs and adjectives; these percentages are similar to the distribution of typical words in social sciences research articles. That the distribution of AKL is similar to that of social sciences but differs from health, life and physical sciences might be because the AKL corpus, as Paquot (2010) mentioned, is skewed towards the humanities and social sciences.

The higher frequency of verbs in social sciences research articles can be related to a verb being mainly used for narration, indicating that social sciences research articles narrate more than health, life and physical sciences articles.

The least frequent word class for all subject areas, shown in Table 1, is the adverb. Indeed, the adverb, which usually modifies a verb, is the least frequent word class in other genres as well (cf. Biber et al., 1999). This present study shows that adverbs rarely modify verbs in research articles, indicating that verbs are used more directly, without any modifiers, because the authors of these articles generally want to communicate information directly.

Distribution of Noun Categories across the Four Subject areas

Nouns are the most frequently used word class in all genres and, as shown in this study, in all subject areas. The result of the CLAWS tagger for nouns in this study showed three categories of nouns: NN, NN1 and NN2. Based on the information from UCREL CLAWS6 Tagset (<http://ucrel.lancs.ac.uk/claws6tags.html>), NN denotes a common noun, neutral for number (e.g. sheep, cod, headquarters), NN1 is the tag for a singular common noun (e.g. book, girl) and NN2 tags a plural common noun (e.g., books, girls). Table 3 displays distribution of noun categories across the four subject areas.

Table 3
Distribution of noun categories across the four subject areas

No.	Types of Nouns	Social Science		Health Science		Life Science		Physical Science	
		Number	%	Number	%	Number	%	Number	%
1	NN	0	0	0	0	1	0.56	0	0
2	NN1	116	60.1	174	78.38	128	71.91	139	73.94
3	NN2	77	39.9	48	21.62	49	27.53	49	26.06
Totals		193	100	222	100	178	100	188	100

As Table 3 shows, NN1 is the most frequently used noun type in all subject areas, indicating that all research articles tend to focus on singular nouns or single objects of research. However, if percentages are compared, social sciences is the odd one out of these four subject areas. NN1 usage in research articles in health, life and physical sciences is more than 70%, whereas its usage in social sciences is only 60.1%. NN2 usage in health, life and physical sciences is less than 30%, while its usage in social sciences reaches 39.9%.

The percentage of NN1 usage is the highest in health sciences at 78.38%, while social sciences shows the lowest usage at 60.1%. Percentages are reversed for NN2; social sciences has the highest, at 39.9%, and health sciences is lowest at only 21.62%. This could mean that research articles in health sciences are more concerned with singular nouns than those in social sciences. However, plural nouns are more common in social sciences research articles than in health sciences research articles. Examples of plural nouns in social sciences research articles are *firms* and *transactions*. Using the Academic Article Concordancer (corpus.kwary.net), 622 occurrences were noted for *firms* but only 435 for *firm* and 66 occurrences were noted for *transactions* but only 49 for *transaction*. *Artery* and *syndrome* are examples of singular nouns in health sciences research articles. Data in the Academic Article Concordancer show 221 occurrences for *artery* but only 34 for *arteries* and 121 occurrences for *syndrome* but only 17 for *syndromes*.

Singular and plural noun usage in health sciences and social sciences can also be compared in terms of the frequency of words used in both subject areas. One example is the singular noun *informant* and its plural *informants*. In health sciences research articles, *informant* is used three times, but *informants* is used only once; thus, the singular noun occupies 75%. In social sciences, the singular noun *informant* is used 12 times, while its plural is used 15 times, with the singular noun occupying only 44% (Table 3).

The high percentage of NN1 in health sciences research articles shows that research in health sciences tends to focus more on single nouns or single objects of research. In contrast, the high percentage of NN2 in research articles in social sciences indicates that research in this field tends to focus more on plural nouns or more than one object of research in comparison with the other three subject areas. In general, however, social sciences still uses NN1 more often than NN2 (i.e., 60.1% against 39.9%).

Distribution of Verb Categories across the Four Subject Areas

In addition to nouns, verbs are also further categorised in this study using the CLAWS tagger into the following categories: VV0, VVI, VVG, VVN, VVD and VVZ. The explanation of these categories is based on information from UCREL CLAWS6 Tagset (<http://ucrel.lancs.ac.uk/claws6tags.html>). Details are as follows:

- _ VV0 = base form of lexical verb (e.g., give, work)
- _ VVI = infinitive (e.g., to give, to work)
- _ VVG = -ing participle of lexical verb (e.g., giving, working)
- _ VVD = past tense of lexical verb (e.g., gave, worked)
- _ VVN = past participle of lexical verb (e.g., given, worked)
- _ VVZ = -s form of lexical verb (e.g., gives, works)

This study is concerned with typical words in each subject area. Thus, the verb categories mentioned above needed to be simplified, particularly because the forms of verbs were the same in the study's corpus. Therefore, data for VV0 and VVI were combined into one (VV0/VVI) on the basis of their similar forms. Additionally, the calculation for VVD and VVN were also combined as VVD/VVN because some verbs have the same forms for both categories. Table 4 shows the distribution of verb categories across the four subject areas.

No.	Types of Verbs	Social Science		Health Science		Life Science		Physical Science	
		Number	%	Number	%	Number	%	Number	%
1	VV0/VVI	24	34.78	1	11.11	5	13.51	8	20
2	VVG	19	27.54	1	11.11	4	10.81	9	22.5
3	VVD/VVN	17	24.64	7	77.78	24	64.87	19	47.5
4	VVZ	9	13.04	0	0	4	10.81	4	10
Totals		69	100	9	100	37	100	40	100

Table 4

Distribution of verb categories across the four subject areas

As seen in Table 4, the distribution of verb categories in health sciences is quite similar to that of life sciences. VV0/VVI and VVG usage in these two subject areas is approximately 11%. The use of VVD/VVN in health sciences and life sciences is also similar, i.e. more than 50% when compared to the other two subject areas, in which VVD/VVN use is lower than 50%. These percentages indicate a close relationship between health sciences and life sciences and are supported by Rothwell's (2002) statement that health sciences and life sciences are similar because they both focus on the study of living organisms.

Based on the calculation in Table 4, the type of verbs most commonly used in all subject areas, except social sciences, is VVD/VVN (i.e. past tense and past participle). VVD and VVN are found at 47.5% in physical sciences, 64.87% in life sciences and 77.78% in health sciences, indicating frequent and common use of past tense and the past participle. Winkler and Metherell (2008) state that in research articles, the past tense indicates study or experimental results that have taken place in the past. From these three subject areas, health sciences has the highest percentage of VVD/VVN: 77.78% means that more than 7 out of 10 verbs used in health sciences research articles are in the past tense or past participle tense. One example is the verbs *undergo* and *underwent*. Based on concordance lines at corpus.kwary.net, the past form (*underwent*) is used 152 times, whereas its base form (*undergo*) is used only 58 times in health sciences research articles. The same verb (*undergo*) is used only three times in its past form (i.e. *underwent*) and four times in its base form (*undergo*) in social sciences research articles.

For social sciences, VVD/VVN use is lower than VV0/VVI and VVG use. As shown in Table 2, percentages for VV0/VVI, VVG and VVD/VVN uses are 34.78, 27.54 and 24.64, respectively. The frequent use of VV0/VVI in social sciences research articles tends to present universally relevant arguments, making them true then and now (Winkler & Metherell 2008, 113). One example is the use of *write* and *wrote*. Based on concordance lines at corpus.kwary.net, the base form (*write*) is used 58 times, whereas its past form (*wrote*) is used only 17 times in social sciences research articles. In contrast, research articles in health sciences use *write* only five times and *wrote* nine times.

Conclusion

The present study has shown similarities and differences in the distribution of the lexical word classes of typical academic words in social, health, life and physical sciences. In all four subject areas, the noun is the most frequent word class, while the adverb is the least frequent. The noun is particularly high in health sciences research articles, indicating that health sciences research articles present the highest density of information among the four subject areas.

The second and third highest distributions of lexical word classes also reveal differences between the subject areas. In social sciences, the second most used form is the verb and the adjective is the third most used. In the other three subject areas, however, the adjective is the second most used, and the verb is the third most used. This positioning shows that the typical academic words in health, life and physical sciences research articles are akin to academic prose, while typical academic words in social sciences research articles are close to conversation. In addition, as concluded from the use of verbs, research articles in social sciences use longer sentences and longer explanations than do those in the other subject areas.

This study also shows that singular nouns are used more often than plural nouns in all subject areas. Health sciences research articles show the highest percentage of singular nouns among the four subject areas. Research articles in social sciences use plural nouns more frequently than those in the other subject areas. For verbs, the past tense and past participle forms are more commonly found in health, life and physical sciences, while the base form appears more often in social sciences. This situation might be due to the fact that health sciences, life sciences, and physical sciences focus mostly on experiments that have been conducted in the past, while social sciences tend to focus on arguments expected to be true regardless of time.

1. Anthony, L., 2014. AntWordProfiler (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from <http://www.laurenceanthony.net/>.
2. Baker, P., 2010. Corpus Methods in Linguistics. In: *Research Methods in Linguistics*, (ed.) Litosseliti, L. London: Continuum.
3. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R., 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
4. Biber, D., Conrad S., & Reppen, R., 2006. *Corpus Linguistics; Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
5. Browne, C., Culligan, B., & Phillips, J., 2013. A New Academic Word List: The most Important Words for Understanding Academic Text. Retrieved from <http://www.newacademicwordlist.org/>.
6. Cameron, L., 2002. Measuring Vocabulary Size in English as an Additional Language. In: *Language Teaching Research*, 6(2), pp. 145–173. <https://doi.org/10.1191/1362168802lr103oa>.
7. Coxhead, A., 2000. A New Academic Word List. In: *TESOL Quarterly*, 34(2), pp. 213–238. <https://doi.org/10.2307/3587951>.
8. Elsevier, 2016. Elsevier. Retrieved from www.elsevier.com/solutions/scopus/content.
9. Gardner, D., & Davies, M., 2014. A New Academic Vocabulary List. In: *Applied Linguistics*, 35(3), pp. 305–327. <https://doi.org/10.1093/applin/amt015>.
10. Kwary, D. A., 2018. A Corpus and a Concordancer of Academic Journal Articles. In: *Data in Brief*, 16, pp. 94–100.
11. Lei, L., & Liu, D., 2016. A New Medical Academic Word List: A Corpus-Based Study with Enhanced Methodology. In: *Journal of English for Academic Purposes*, 22, pp. 42–53. <https://doi.org/10.1016/j.jeap.2016.01.008>.
12. Liu, J., & Han, L., 2015. A Corpus-Based Environmental Academic Word List Building and Its Validity Test. In: *English for Specific Purposes*, 39, pp. 1–11. <https://doi.org/10.1016/j.esp.2015.03.001>.
13. McCarten, J., 2007. *Teaching Vocabulary; Lessons from the Corpus, Lessons for the Classroom*. New York: Cambridge University Press.
14. Martinez, I. A., Beck, S. C., & Panza, C. B., 2009. Academic Vocabulary in Agriculture Research Articles: A Corpus-Based Study. In: *English for Specific Purposes*, 28(3), pp. 183–198. <https://doi.org/10.1016/j.esp.2009.04.003>.
15. *Oxford Learner's Dictionary of Academic English*, 2014. Oxford: Oxford University Press.
16. Nation, I. S. P., 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139524759>.
17. Paquot, M., 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. New York: Continuum.
18. Qian, D. D., & Schedl, M., 2004. Evaluation of an In-Depth Vocabulary Knowledge Measure for Assessing Reading Performance. In: *Language Testing*, 21(1), pp. 28–52. <https://doi.org/10.1191/0265532204lt273oa>.
19. Radford, A., 2009. *An Introduction to English Sentence Structure*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511800924>.
20. Rayson, P., 2016. Free CLAWS WWW Tagger Lancaster University [Computer Software]. Retrieved from <http://ucrel.lancs.ac.uk/claws/trial.html>.
21. Rothwell, N., 2002. *Who Wants to Be a Scientist?: Choosing Science as a Career*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511754999>.
22. Schmitt, N., Schmitt, D., & Clapham, C., 2001. Developing and Exploring the Behaviour of Two New Versions of the Vocabulary Levels Test. In: *Language Testing*, 18(1), pp. 55–88. <https://doi.org/10.1177/026553220101800103>.
23. Valipouri, L., & Nassaji, H., 2013. A Corpus-Based Study of Academic Vocabulary in Chemistry Research Articles. In: *Journal of English for Academic Purposes*, 12(4), pp. 248–263. <https://doi.org/10.1016/j.jeap.2013.07.001>.

References

24. Walsh, S., 2010. What Features of Spoken and Written Corpora Can Be Exploited in Creating Language Teaching Materials and Syllabuses? In: *The Routledge Handbook of Corpus Linguistics*, (eds.) O'Keeffe, A., & McCarthy, M. pp. 333–344. New York: Routledge. <https://doi.org/10.4324/9780203856949.ch24>.
25. Winkler, A., & Metherell, J. R., 2008. *Writing the Research Paper: A Handbook*. Boston: Cengage Learning.

Santrauka

Deny Arnos Kwary, Almira F. Artha, Yuni S. Amalia. Kalbos dalių pasiskirstymas keturių skirtingų mokslo sričių tyrimuose

Šio tyrimo tikslas buvo nustatyti kalbos dalių pasiskirstymą socialinių, sveikatos, fizinių ir gamtos mokslų straipsniuose. Iš „Elsevier“ duomenų bazėje paskelbtų mokslinių straipsnių buvo išrinkta ir ištirta 5 754 560 žodžių formų. Rezultatai parodė kalbos dalių pasiskirstymo panašumus ir skirtumus keturiuose skirtinguose tyrimų tematikose. Sveikatos, fizinių ir gamtos mokslų srityse daiktavardis yra labiausiai dominuojanti kalbos dalis, po jo seka būdvardis, veiksmažodis ir prievieksmis. Socialiniuose moksluose vyraujanti kalbos dalis yra veiksmažodis, šiek tiek rečiau sutinkami būdvardžiai. Tai rodo, kad socialinių mokslų srities straipsniai pasižymi didesniu žodžių skaičiumi ir yra rašomi laisvesniu, beveik šnekamosios kalbos, stiliumi. Kalbant apie daiktavardžius, visose mokslo srityse dažniausiai yra vartojama daiktavardžių vienaskaita, rodanti, kad moksliniai tyrimai dažniausiai orientuojasi į vieną tiriamąjį objektą. Veiksmažodžiai sveikatos, gamtos ir fizinių mokslų straipsniuose dažniau vartojami būtojo laiko formomis. Taip pabrėžiami jau atlikti tam tikros mokslo srities darbai ar tyrima. Tuo tarpu socialinių mokslų straipsniuose dažniau vartojama pagrindinė veiksmažodžio forma – bendratis. Taip argumentuojamos bendrosios tiesos.

About the Authors

Deny Arnos Kwary

Dr., Head of English Department, Airlangga University, Indonesia

Research interests

lexicography, corpus linguistics, English language teaching

Address

Dharmawangsa Dalam, Surabaya 60286, Indonesia

E-mail

d.a.kwary@fib.unair.ac.id

Almira F. Artha

Research fellow, Airlangga University, Indonesia

Research interests

corpus linguistics, English language teaching, discourse analysis

Address

Dharmawangsa Dalam, Surabaya 60286, Indonesia

E-mail

airafidel@gmail.com

Yuni S. Amalia

Dr., lecturer, Airlangga University, Indonesia

Research interests

English language teaching, curriculum development, educational linguistics

Address

Dharmawangsa Dalam, Surabaya 60286, Indonesia

E-mail

sarihunt262@gmail.com