ANNALS OF THE POLISH ASSOCIATION
OF AGRICULTURAL AND AGRIBUSINESS ECONOMISTS

received: 24.06.2019
acceptance: 09.08.2019
published: 20.09.2019

Annals PAAAE • 2019 • Vol. XXI • No. (3)

JEL codes: A31, B49, C38

DOI: 10.5604/01.3001.0013.3448

ISTVÁN TAKÁCS, KATALIN TAKÁCS-GYÖRGY

Óbuda University, Hungary

MAIN FOCUSES OF ENGLISH PAPERS OF ANNALS (PAAAE) DURING THE LAST TEN YEARS

Key words: text mining, R language programme, agriculture, publication, terms

ABSTRACT. The aim of the research is to analyze English language scientific articles that have been published in the scientific journal entitled Annals PAAAE from 2009 to 2018 by using tools of text mining and text analysis, measuring how the frequency of key terms has changed, and to conclude the probable change of research focus of authors based on it. The subjects of the research are 393 English language articles that have been issued in the volumes of Annals PAAAE (Polish title is Roczniki Naukowe SERiA) from 2009 to 2018 (from Vol. 11 to Vol. 20), which are ready to download in pdf format on the website of the publishing house. Text mining and text analysis were conducted using own self-adapted routines of samples written in the R language programme. At the hierarchic cluster analysis, the Ward D method was used, and the association of frequency of terms was measured using Pearson's correlation. It was found that the weight of research concerning the context of the product-production-market has increased, but there was low usage of the term "competitiveness". What was of key significance was that in the case of clustering articles by author nationality, the role of the origin of the author has decreased as a determinant factor, which could suggest the uniformity of terminology as well as similarity of research topics and the researcher's interests, which can initialize cooperation and common research. The growth of the number of articles elaborated on by international teams is also a projection of this tendency.

INTRODUCTION

In 2017, the annually elected president of the Southern Agricultural Economics Association, Lisa House, chose the topic of her presidential address to be, "Choose your words wisely". [House 2017] Using tools of text data mining, the articles of the Journal of Agricultural and Applied Economics and the American Journal of Agricultural Economics were analyzed and compared. Her research question was: What can we learn from investigating word usage in our journals? The final conclusion was: "Word analysis is far from perfect … however, it does give us food for thought".

According to Ian H. Witten, text mining or text data mining has become an accepted method of exploring deeper content and connections of written material [Witten 2004]. Text mining is increasingly more commonly used and more and more scientific papers are published discussing the results obtained from this method. The software available for text analysis (e.g. Atlas) and volunteer developers of R open source programming

language developed for self-conscious researchers offer more and more solutions for so-called lay-users (see for example I.H. Witten [2004], Graham Williams [2016], Yanchang Zhao [2013]).

The reason for this research is the fact that one of or the most important scientific journals of Polish agricultural economists is Annals PAAAE (Roczniki Naukowe SERiA), published in English from 2019 onwards. Since this modification was introduced, an era of the journal came to its end in 2018. It has since become international and is one of the major journals of agricultural economics in CEE countries in the last 15 years mainly due to the appearance of Hungarian authors (but in addition to Polish and Hungarian authors, there were authors from ten other countries during the period under review).

The main focus of the paper was to show how the subjects of English papers have changed in the last ten years, due to the fact that first only Polish papers were published in the journal. Due to such limitations, the paper does not discuss the changes of topics of Polish papers, so the results are not representative or valid for all articles, but cover all English papers. Native spoken Polish authors should analyze that issue. It is expected that the papers will deal more with international aspects of questions regarding agricultural economic issues. The discussion of the formerly published articles by foreign and Polish authors, also helps to discover and compare tendencies in international processes.

Since 2009, five years have passed since the countries of the region joined the European Union. Research connected to European support programmes were focused on by new Member States at that time but, simultaneously, the financial and economic crisis that took place was already showing signs and its impact is proven by the papers [Takács, Baranyai 2009].

One of the focal points was the development of rural areas, as part of questions and relations regarding tourism and the countryside. The following topics were discussed and published: the employment issues of tourism [Balińska 2009], the role of tourism in rural development [Brelik 2009], as well as connections between rural areas and urban development [Staszewska 2009]. Reducing regional differences is a basic objective in the EU, however, differences in the development of rural areas can be mitigated by different development policies. This topic appeared in year-2009 papers [Koreleski 2009]. Of course, the topics that had previously been characteristic did not disappear by 2018, like rural tourism, but new contexts also appeared [Wojcieszak, Zawadka 2018]. At the same time, researchers started to examine the questions of the new European Union budget cycle [Wieliczko 2018].

This paper analyses what changes can be identified by investigating the expressions of papers published in the last decade.

The research questions were as follows: Does the profile created with tools of text mining meet the declared scope of the journal? Does the considerable share of foreign authors have an impact on the profile as the publication culture is different in different countries? And do the possible differences in the use of terminology have an impact on the set of words of papers beyond its narrow professional content?

The goals of the research were to examine the increasing tendency of English language papers appearing in the Annals of PAAAE, due to the fact that in the first issues mainly Polish papers were published in the journal; to show how the subjects of them changed in

the last ten years; to characterize the ten years of study; to identify the change of research focus in the papers by evaluating the changes of the frequency of terms used; to analyse the keywords and to conduct a cluster identification of papers by the usage of terms and the characterization of clusters.

RESEARCH MATERIAL AND METHODS

The English language articles published in the volumes of the journal entitled the Annals Polish Association of Agricultural and Agribusiness Economists (PAAAE) (The Polish name is Roczniki Naukowe SERiA) from 2009 to 2018 (from Vol. 11 to Vol. 20) are the data resource (the analyzed texts) of the research. The articles could be downloaded in pdf format from the website of the publishing house [see https://rnseria.com/resources/html/archives]. During the 10 years, 393 English articles were published, of which one article in 2013 and a further 24 articles in 2015 were not uploaded into the repository – only their abstracts were. Due to these facts, during the next steps of the research, 40 articles were analyzed in 2013 and only 8 articles in 2015. Because of a high share of lacking articles, volume 2015 was disregarded from some analyses.

Text mining and text analysis were conducted by adapting sample routines in the open source programme language R published by G. Williams [2016] and Y. Zhao [2013] and using the manuals of the R packages (see [https://cran.r-project.org/]).

The steps of the process were 1) Preparation of data: reading pdf files and creating the text corpus 2) Transformation of the corpus for analysis preparation: converting the text to lower case, removing numbers and punctuation, removing stop words, stemming

Table 1. Frequency of origin countries of English language articles in Annals PAAAE from Vol. 11 to Vol. 20

Origin of authors	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Sum of origin	
	piece										piece	%
Sum of years	26	38	50	55	41	42	32	40	25	44	393	100.0
Poland	15	15	32	30	20	26	24	19	9	29	219	55.7
Poland & others				1		1	2	2		3	9	2.3
Hungary	9	17	16	22	18	14	4	17	13	10	140	35.6
Hungary & others									1		1	0.3
US	1	1		1	2						5	1.3
US & others			1				1	1	1		4	1.0
Other countries	1	5	1	1	1	1	1	1	1	2	15	3.8

Source: own study

and identifying synonyms 3) Creating the Text Document Matrix and the Document Text Matrix 4) Exploring the Text Document Matrix: frequency of terms, visualization of the distribution of term frequencies 5) Identifying frequent items and counting associations (correlations) 6) Design of the word cloud 7) Cluster analysis of terms 8) Cluster analysis of articles 9) Design of the correlation plots of terms (Remark: the process of analysis was realized by a batch file including 283 rows of commands in R. Because of the limited number of pages, this article is not able to present all results, but the elaborated tools are also a kind of novelty that could be used for similar research later).

During analyses, terms were used which cropped up at least 50 times in all articles in a single year.

One step of the transformation was stemming words involving cleaning tags of words, e.g. the "development", the "developing", the "developed", transforming it to "develop" and its modified number of frequency containing each term with the root "develop".

A further methodological choice was to apply the Ward2 method during hierarchic cluster analysis.

During the next phase of the research, a comparative analysis of annual results was carried out, and the change of the specific frequency of terms was analyzed identifying the impact of the nationality of the author on the frequency of terms with statistical confirmation. It was then analyzed by Pearson's correlation and linear regression models using SPSS.

RESEARCH RESULTS

Having analyzed the share of nationalities in English language articles, it can be identified that the share of Polish authors has changed and is between 36-66%, while the share of Hungarian authors has changed from 23 to 52% in different years (Table 2). The average of articles written by Hungarians was about four-fifths of Polish origin articles, but in 2017 the share of articles signed by Hungarians exceeded the share of articles published by Polish authors. It is evident that the new publishing policy introduced from 2019 (i.e. all articles will be published in English) would result in fundamental change.

Assuming that the frequency of terms related to the authors' emphasized message,n prevalence (frequency) per article was examined in the articles published in the ten year volumes in the following steps (Table 3). From the set of terms that remained after the

Table 2. Change of share of origin countries of English language articles in the Annals PAAAE, Vol. 11 to Vol. 20

Denomination	Means	SD	Min	Max	Year of min	Year of max
Share of Polish authors	0.53	0.11	0.36	0.66	2017	2018
Share of Hungarian authors	0.38	0.09	0.23	0.52	2018	2017
Rate of Hungarian to Polish	0.79	0.35	0.34	1.44	2018	2017

Source: own study

manipulation of texts, nearly 100 to 250 terms met the criteria of the occurrence of at least 50 times in the articles of a volume. The terms (e.g. "agriculture", "farming", "rural") after the unification of terms with related meanings characteristically reflect the professional orientation of the journal as well as the founder organization.

Concerning the change of focus of the articles, the last column of Table 3 can provide information. The Slope of ranks depicts negative value signs for the growth of frequency of a term and the average change of rank, too, while in the column of the Slope of specific frequency positive values show the average growth of the frequency of terms. In this case the term "agriculture" is not examined. When it came to other terms like the "production", the "product", the "cost" and the "food", the ranks of relative prevalence getting better indicate the growing importance of production issues, while a decreasing trend of the prevalence of the "European Union" indicates that time since accession has passed and the practice of CAP is stabilized more or less without any novelty.

Table 3. Change of frequency of the most mentioned terms in English language articles of the Annals PAAAE from Vol. 11 to Vol. 20

Terms in order to average frequency	Average frequency	Slope of spec. freq.	Means of ranks	Modus of ranks	SD of ranks	Best rank	Worst rank	Year of		Slope of ranks
								best rank	worst rank	
Agriculture/ agricultural	12.15	0.35	1.3	1	0.7	1	3	more	2011	-0.08
Farm/farming	9.93	-0.18	3.1	3	2.0	1	8	2011	2016	0.13
Develop	9.12	-0.20	4.6	2	3.9	1	14	2012	2013	0.15
Production	8.84	0.31	4.4	2	2.2	2	8	more	2011	-0.33
Area	6.92	-0.20	8.0	7	3.4	4	15	2009	2013	0.26
Poland/Polish	6.81	0.06	7.9	8	3.4	4	15	more	2012	-0.38
Market	6.14	-0.10	10.8	NA	4.3	5	18	2010	2012	0.16
Economic/ economy	6.04	-0.35	12.3	5	7.3	5	27	more	2017	1.73
Rural	5.67	-0.18	14.2	9	8.5	4	28	2016	2018	1.42
Product	5.18	0.10	14.2	11	4.5	9	23	2018	2012	-0.72
Food	5.12	0.06	19.3	NA	14.2	6	49	2014	2012	-1,36
Cost	4.48	0.33	33.2	NA	42.2	6	138	2018	2013	-3.25
Hungary/ Hungarian	3.70	0.19	36.0	NA	36.0	6	117	2013	2009	-5,55
Farmer	3.63	0.05	30.1	NA	5.1	21	38	2011	2012	-0,18
Land	3.48	-0.21	50.3	53	46.5	3	154	2013	2011	-3,69
System	3.39	-0.09	37.0	NA	20.0	9	76	2014	2010	-0.63
EU/Europe/ European	2.09	-0.52	82.2	50	84.7	3	277	2010	2011	-7.78

Source: own study

Table 4. Correlation of the frequency of some terms and the origin country of authors

Denomination	Share of Polish	Share of Hungarian	Rate of Hungarian to Polish	Change
Share of Polish	1	-0.940**	-0.982**	-0.736*
Share of Hungarian	-0.940**	1	0.954**	0.702*
Rate of Hungarian to Polish	-0.982**	0.954**	1	0.831**
Change	-0.736*	0.702*	0.831**	1

Remark: ** Sig 0.99, * Sig 0.95
Source: own study

The terms "farming" and "economic" have been losers and in case of terms which are usually connected to previous terms e.g. "develop", the "market", the "rural", the change was the same.

The previous assumption is confirmed by association (i.e. correlation) analysis, too. Among the terms with an association coefficient over 0.65, in case of the term the "rural" the association of it is 0.88 with e.g. "agriculture", "attractives", "authentic", "culture", "ecotourism", "infrastructure", "leisure", "revitalise", "tourism", "touristic", "tradition", "visitors", "weekend", 0.86 is with the "agritourism", "heritage", 0.78, 0,77, and 0.75 are with the "develop", the "undevelop" and the "develop strategies", furthermore 0.75 is with the "holiday", "promotes", "marketing campaign". It could be highlighted that there are terms which were characteristic in only a part of years e.g. farm commodity (0.77, 2014), environment (0.75, 2009).

If the connection between the nationality of the authors and the frequency of a term was analyzed (Table 4) then, in case of the Polish authors, the term "change" had a strong negative significant correlation with the share of the nationality, but in case of Hungarians this correlation was strong positive significant, while in case of other examined terms this kind of connection could not be identified as strong connections, only mid-strong or weak connections existed, but were not significant in any case.

Examining the correlation between terms, a high or moderately strong significant correlation was provable in several cases. (Table 5) An interesting outcome was that there is a strong, positive correlation between the terms "Poland/Polish" and "product" in the context of frequency of mentioning. The frequency of mentioning "Hungary/Hungarian" terms has a weak but negative correlation with most terms. In particular, the term "develop" has a moderately strong negative correlation that does not seem favorable and is a thoughtful outcome for the authors of this paper.

As a next step, the cluster analysis of terms was carried out. The analysis was done by the Ward. D method [Murtagh, Legendre 2014], and by the visual examination of dendrograms (Figure 1). Five or seven clusters were defined, depending on basic settings.

It is characteristic that two thirds or three quarters of the terms are grouped in one or two clusters, the remaining terms set in the other three of five clusters. The internal structure of the clusters were differentiated by increasing the spars value but, of course, the place of terms did not change.

Table 5. Correlation among some of the most mentioned terms according to frequency of terms

Terms	Agriculture/ agricultural	Develop	Production	Poland/ Polish	High	Product	Cost	Change	Increase	Hungary/ Hungarian
Agriculture/agricultural	1	0.069	0.776*	0.642	0.839**	0.637	0.477	-0.465	0.685*	-0.028
Develop	0.069	1	0.171	0.404	0.151	0.295	0.506	-0.455	0.173	-0.680*
Production	0.776*	0.171	1	0.416	0.669*	0.548	0.533	-0.195	0.727*	0.267
Poland/Polish	0.642	0.404	0.416	1	0.275	0.928**	0.515	-0.264	0.601	-0.195
High	0.839**	0.151	0.669*	0.275	1	0.242	0.293	-0.738*	0.477	-0.139
Product	0.637	0.295	0.548	0.928**	0.242	1	0.669*	-0.049	0.816**	-0.008
Cost	0.477	0.506	0.533	0.515	0.293	0.669*	1	-0.075	0.798**	-0.346
Change	-0.465	-0.455	-0.195	-0.264	-0.738*	-0.049	-0.075	1	-0.049	0.452
Increase	0.685*	0.173	0.727*	0.601	0.477	0.816**	0.798**	-0.049	1	0.082
Hungary/ Hungarian	-0.028	-0.680*	0.267	-0.195	-0.139	-0.008	-0.346	0.452	0.082	1

Remark: ** Sig 0.99, * Sig 0.95

Source: own study

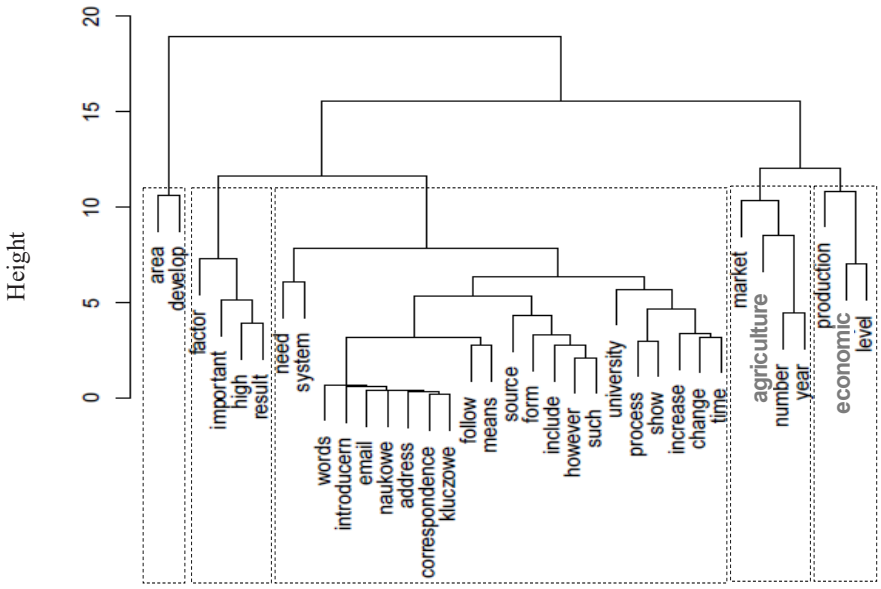

Figure 1. Groups of the most mentioned words in 2009 (hierarchic cluster dendrogram of terms by the Ward.D2 method, sparse = 0.2)

Source: own study

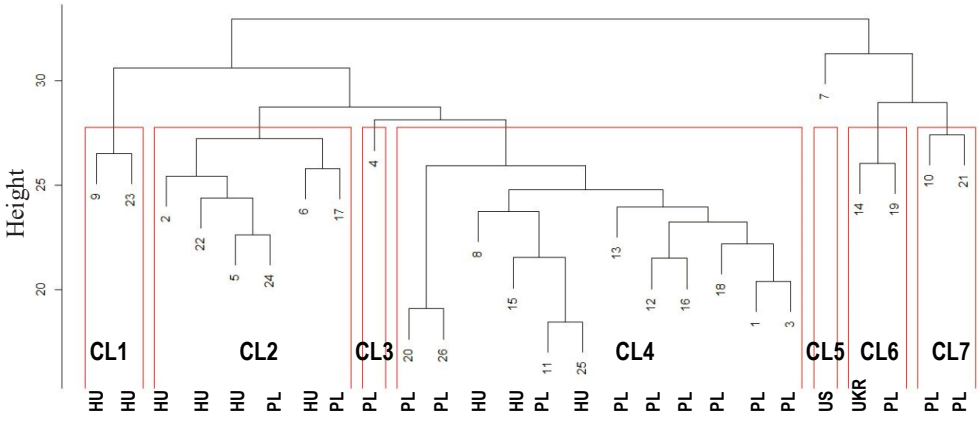

Figure 2. Groups of articles and nationality of authors in 2009 (hierarchic cluster dendrogram by the Ward.D2 method, sparse 0.65)

Source: own study

Table 6. The most mentioned words of the clusters in the articles in 2009

Cluster	Term				
	1.	2.	3.	4.	5.
CL1.	market	agriculture	production	time	increase
CL2.	agriculture	number	area	develop	year
CL3.	factor	high	time	source	important
CL4.	develop	economic	area	factor	level
CL5.	need	system	area	production	level
CL6.	production	system	develop	economic	agriculture
CL7.	area	develop	result	change	form

Source: own study

Cluster analysis of the terms was carried out by applying the Term-Document Matrix (Tdm). The transposed text matrix, the Document-Term Matrix (Dtm) is suitable for hierarchical clustering of documents (in this case the papers of the journal) (Figure 2).

It can be stated that even papers of Hungarian and Polish authors belong to separate clusters, separated from each other in 2009. Later, mixed clusters became more typical. As clustering is based on the frequency of terms, one of the factors of group formation itself is the topic of the paper. That is why the keywords typical of clusters could be identified (Table 6). It also shows the differences of focuses between research teams, but may also refer to the differences in the use of terms.

As an example: by frequency of terms coming from 2009 (due to the dendrogram), papers number 11 [Kołodziejczak 2009] and 25 [Vajna Istvánné Tangl 2009] are close to each other. Examining the papers, the following can be concluded: both papers are theoretical and although they do not seem relevant to the titles (Integrated Agriculture in Poland [Kołodziejczak 2009], and Correlation of Quality Development Methods to Efficiency and Competitiveness in Agricultural Companies and their Appearance in the Accounting System [Vajna Istvánné Tangl 2009]), there is a particularly high number of "quality", "develop~", "control~" and "system~".

The word cloud and the Correlation plots formerly mentioned in the research process were also completed but, due to the limitation of this paper, the charts are not presented, but the results obtained from them confirm the above described tendencies.

CONCLUSIONS

The Polish Association of Agricultural and Agribusiness Economists (i.e. SERiA) celebrated the 25th anniversary of its foundation in 2018 and published the 20th year of the journal (i.e. Roczniki Naukowe or the English issue entitled Annals PAAAE). During the last two decades the journal published an increasing number of English papers from foreign and domestic authors, giving a wider platform to become international. The journal only accepts papers in English as the next step in the process from 2019 onwards.

These changes led to the identification of the characteristics of changes in publication using developing tools of text mining and text analysis.

More than 40% of papers came from foreign authors of 393 English papers published in the last ten years. The share of Hungarian authors exceeded one third of the papers. The main outcomes of the analysis are as follows:

1. The journal is a leading publication platform for the society of Hungarian agricultural economists; there are a large number of researchers with high professional prestige among Hungarian authors; the fact that the Class of Agricultural Sciences of The Academy of Hungarian Sciences accepts this international foreign scientific journal, highlights the journal's reputation in Hungary.
2. Besides the expressions in the ranking of terms in scientific publications – clearly following the journal's mission – the words "market; products" referring to the role of agriculture in the market are determinant. It is surprising that the term "competitive" has never been included in the first 30 and its average position is 59.6. At the same time, the appearance of keywords shows the direction of change in research focus. By the end of the examined ten years less were mentioning "rural", "ecotourism", "revitalise", "tradition", and more mentioned "rural" "production" and "cost".
3. The papers of the journal are becoming less and less distinct by the author nationality.
4. The number of papers written by teams of authors from different countries is gradually increasing.

The Polish articles were not subject to the present examination, and based on the examination of previously published articles in English that contribute to the internationalization of the journal, it was concluded that the articles published by Polish and non-Polish authors were distinctly different by text context. Over the past ten years, convergence has been identified, showing more homogeneous text relationships and word-for-word habits for both sides.

BIBLIOGRAPHY

Balińska Agata. 2009. Tourism as a form of non-agricultural activity in the opinion of the selected rural communities inhabitants. *Annals PAAAE* XI (6): 5-10.

Brelik Agnieszka. 2009. Rural tourism development in Poland. *Annals PAAAE* XI (6): 17-20.

House Lisa. 2017. *Choose your words wisely. Presidential Address*. [In] Southern Agricultural Economics Association 49th Annual Meeting. February 4-7, 2017, Mobile, Alabama (handout, provided by W. J. Florkowski, 2019).

Kołodziejczak Anna. 2009. Integrated agriculture in Poland. *Annals PAAAE* XI (6): 62-65.

Koreleski Dariusz. 2009. Interregional differentiation of the development level in Poland with regard to the chosen parameters. *Annals PAAAE* XI (6): 66-71.

Murtagh Fionn, Pierre Legendre. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification* 31: 274-295.

Staszewska Sylwia. 2009. Rural Areas – a New Space for Urban Development. *Annals PAAAE* XI (6): 115-120.

Takács István, Zsolt Baranyai. 2009. Agricultural products noted on commodities exchange and Global Financial Crisis. *Annals PAAAE* XI (6): 121-127.

Vajna Istvánné Tangl Anita. 2009. Correlation of Quality Development Methods to efficiency and competitiveness in agricultural companies and their appearance in the accounting system. *Annals PAAAE* XI (6): 139-142.

Wieliczko Barbara. 2018. Financial instruments in Cap 2020+. *Annals PAAAE* XX (4): 205-209.
Williams Graham. 2016. *Hands-on data science with R Text Mining*, https://onepager.togaware.com/TextMiningO.pdf, access: 30.05.2019.
Witten Ian H. 2004. *Text mining*, https://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf, access: 30.05.2019.
Wojcieszak Monika, Jan Zawadka. 2018. Cultural values as a determinant of the development of tourism in rural areas and their popularity among Poles based on the example of Folk Culture Museums. *Annals PAAAE* XX (1): 149-155.
Zhao Yanchang. 2013. *R and Data Mining: Examples and Case Studies*. http://www2.rdatamining.com/uploads/5/7/1/3/57136767/rdatamining-book.pdf, access: 30.05.2019.

***

GŁÓWNE POLA ZAINTERESOWANIA ANGIELSKICH ARTYKUŁÓW BADAWCZYCH PUBLIKOWANYCH W ROCZNIKACH NAUKOWYCH SERIA W OSTATNICH DZIESIĘCIU LATACH

Słowa kluczowe: eksploracja tekstu, język programowania R, rolnictwo, publikacje, terminologia

ABSTRAKT

Celem badań było dokonanie analizy artykułów naukowych publikowanych w języku angielskim w Rocznikach Naukowym SERiA w latach 2009-2018. Wykorzystano analizę i eksplorację tekstu, dokonano pomiaru zmian w częstotliwości pojawiania się kluczowych terminów oraz wyciągnięto na tej podstawie wnioski dotyczące możliwej zmiany zainteresowań badawczych u autorów. Przedmiotem badania były 393 artykuły w języku angielskim, które zostały opublikowane w Rocznikach Naukowych SERiA w latach 2009-2018 (tomy XI-XX), które są dostępne w formacie pdf na stronie wydawnictwa. Eksplorację i analizę tekstu przeprowadzono za pomocą dostosowanych do potrzeb własnych procedur stworzonych w języku programowania R. W ramach grupowania hierarchicznego zastosowano metodę Warda D, a powiązanie częstotliwości występowania terminów obliczono za pomocą korelacji Pearsona. Stwierdzono, że wzrosła liczba kwerend w kontekście produktu-produkcji-rynku, jednak rzadko korzystano z terminu „konkurencyjność". Ważnym doświadczeniem było stwierdzenie, że w przypadku grupowania artykułów według narodowości autorów, spadło znaczenie pochodzenia autora jako czynnika determinującego, co mogłoby sugerować jednolitość terminologii, a także podobieństwo tematów badań i zainteresowań badaczy. To z kolei może przełożyć się na zainicjowanie współpracy i prowadzenie badań wspólnych. Wzrost liczby artykułów przygotowanych przez zespoły międzynarodowe jest również prognozą tej tendencji.

AUTHORS

ISTVÁN TAKÁCS PROF. DR. HAB.
ORCID: 0000-0002-0847-3844
Óbuda University
Institute of Economics and Social Sciences
Keleti Faculty of Business and Management
15-17 Tavaszmező St., 1084 Budapest, Hungary

KATALIN TAKÁCS-GYÖRGY PROF. DR. HAB.
ORCID: 0000-0002-9129-7481
Óbuda University
Institute of Management and Organisation
Keleti Faculty of Business and Management
8 Népszínház St., 1081 Budapest, Hungary