Web of Science and Scopus language coverage[1]

Miguel-Angel Vera-Baceta: Information Technologies Research Group, Facultad de Comunicación y Documentación. Universidad de Murcia, Campus de Espinardo, Murcia 30100, Spain orcid.org/0000-0003-3912-5882
Michael Thelwall, Kayvan Kousha: Statistical Cybermetrics Research Group, University of Wolverhampton, UK

The evaluation of research outputs in the form of journal articles is important to help with monitoring performance and to allocate funds. Elsevier's Scopus and Clarivate's Web of Science (WoS) are the two main sources for identifying outputs. For non-English-speaking countries, it is especially important that most of the scientific activity evaluated is represented in the bibliometric database used. All documents published in Scopus and WoS during 2018 (6,094,079 documents) were therefore analysed and compared for their languages and research areas. The most comprehensive source for each language and research area were identified and some coverage problems have been found.

Keywords: Research evaluation; Bibliometrics; Bibliographic database; Scopus; Web of Science

Introduction

Scientific and technical research, development and innovation contribute to social progress and welfare. They provide the keys to current global challenges, as well as helping the competitiveness and productivity of countries. Funding institutions often use bibliometric methods to help them decide on priorities and to evaluate funded research. These are often based on one of two main data sources: Clarivate Analytics' Web of Science (WoS) and Elsevier's Scopus.

Analyses based on WoS and Scopus are only valid if the databases offer representative coverage of the scientific activities evaluated (Mongeon & Paul-Hus, 2016). For non-English speaking countries, language is an important factor, as well as the extent of WoS and Scopus coverage of the different areas of research produced by each country (Van Leewen et al., 2001).

In some cases, there are no specific sources at the country or language levels, and WoS and Scopus can be used as references for each other to some extent. For example, in the case of Spanish, state institutions use these two databases as a source for the accounting of research outputs (Instituto Cervantes, 2009; Fundación Telefónica, 2013; FECYT, 2017), or global studies in which world scientific production is quantified, as in the case of the National Science Foundation of the USA, Scopus is taken as a source (NSF, 2018).

In Norway, an example of a country with an apparently comprehensive national index (Norwegian Science Index), 87% of publications are written in English, Scopus covers 72% of the total outputs and the Web of Science Core Collection covers 69% (Aksnes & Sivertsen, 2019). Other evaluations have been made based on citations (Van Leeuwen et al., 2001), and these estimate that 19% to 38% of non-English documents are omitted from Scopus or the Web of Science depending on the area (Martin-Martin et al., 2018).

Incomplete database coverage affects the citation counts of non-indexed documents (Moed, Markusov & Akoev, 2018) and it influences assessments of the impact of research in non-English languages. This has led some applications of bibliometrics, such as the

[1] Vera-Baceta, M., Thelwall, M. & Kousha, K. (in press). Web of Science and Scopus language coverage, *Scientometrics*.

Leiden Ranking[2], to only include publications written in English[3]. In contrast, the Leiden Manifesto[4] argues for the need to protect excellence in locally relevant research and the need to build metrics on high-quality non-English literature that would serve to identify and reward excellence in locally relevant research (Hicks et al., 2015).

Previous studies have investigated the journal coverage of WoS and Scopus based on field classifications, publisher countries and publication languages, reporting very general information (Gavel & Iselid, 2008; Archambault et al., 2009; Mongeon & Paul-Hus, 2016). Other studies have analysed coverage in specific areas, including: Social Sciences and Humanities (Archambault et al., 2006), Computer Sciences (Franceschet, 2009), Library and Information Science (Abrizah et al., 2012), Business Administration (Clermont & Dyckhoff, 2012) or Earth and Atmospheric Sciences (Barnett & Lascar, 2012). WoS and Scopus have also been compared for selected geographical areas: Latin American (Collazo-Reyes, 2014); Latin American and Caribbean (Santa & Herrero-Solana, 2010); Russia (Moed, Markusova & Akoev, 2018) and Norway (Aksnes & Sivertsen, 2019); and one study has combined a geographic area and subject: Spanish Psychology (Osca-Lluch et al., 2013).

These studies have all identified that there is an overrepresentation of English language journals and English-speaking countries as well as an underestimation of documents from the Arts and Humanities and Social Sciences research areas. They also agree that there is a disparity between WoS and Scopus in the amount and type of coverage and therefore caution should be exercised when using these sources, especially for non-English analyses. In general Scopus tends to have greater coverage than WoS in all areas and languages investigated so far, but both have weak coverage of some languages and research areas.

Despite the above findings, the focus on journals and equating languages and countries (the same language may be spoken in several countries, while in the same country several languages may be spoken) mean that document-level findings are needed to give more complete information (Gavel & Iselid, 2008; Mongeon & Paul-Hus, 2016).

Research question

This study focuses on the language coverage of WoS and Scopus at the document level, separating out research areas in the results. Coverage here refers to the number of documents indexed in each database. The following question is addressed.

How many documents are indexed by WoS and Scopus by language and research area?

Methods

The research design was to investigate the subject area and language coverage of WoS and Scopus for 2018. This year was chosen as the most recent complete year, therefore giving the most relevant complete information. Initially, data from 2017 was also collected but this has been discarded because the results are very similar to those from 2018. Language and subject coverage information from each database was extracted with multiple queries submitted to each database, then collating and summarising the results.

WoS and Scopus during 2018 indexed a total of 6,071,821 documents. All the data was collected on May 30, 2019 to prevent time from affecting the comparisons.

[2] http://www.leidenranking.com/
[3] http://www.leidenranking.com/information/indicators
[4] http://www.leidenmanifesto.org/

The WoS information was obtained from the “Web of Science Core Collection”, making a basic query, first per year and then per year and language. No document types were excluded from the results. The WoS results analyser was then used to extract information about subjects. This data was extracted from WoS fields “Language” and " Research Area". The Web of Science Core Collection includes Science Citation Index Expanded (SCI-EXPANDED), Social Sciences Citation Index (SSCI), Arts & Humanities Citation Index (A&HCI), Conference Proceedings Citation Index- Science (CPCI-S), Conference Proceedings Citation Index- Social Science & Humanities (CPCI-SSH), Book Citation Index Science (BKCI-S), Book Citation Index Social Sciences & Humanities (BKCI-SSH), Emerging Sources Citation Index (ESCI), Current Chemical Reactions (CCR-EXPANDED) and Index Chemicus (IC).

The Scopus information was obtained with an advanced query to filter by publication year and the “Refine results” option was used to collect information about the language and subject. This data was obtained from the Scopus fields “Language”, "Subject Area" and "Document Type".

Languages

The documents analysed belong to 51 languages, although Scopus reported 382 documents as language undefined (0.006% of the total). In all cases, the number of documents per language sums to more than the total number of documents (never more than 0.43% more), indicating that some documents are indexed in several languages either by mistake or because they are published in multiple languages.

Languages that occurred in less than 0.01% of documents (312 Scopus documents or 294 WoS documents) in either database were excluded from the analysis: Afrikaans, Arabic, Azerbaijani, Basque, Bosnian, Bulgarian, Esperanto, Estonian, Finnish, Galician, Georgian, Greek, Hebrew, Icelandic, Indonesian, Irish Gaelic, Latin, Latvian, Lithuanian, Macedonian, Romanian, Scottish Gaelic, Serbian, Serbo-Croatian, Tagalog and Welsh. Discarded language documents account for 4,932 publications (0.04% of all documents).

The remaining 25 languages occurred in at least 0.01% WoS and Scopus documents: Catalan, Chinese, Croatian, Czech, Danish, Dutch, English, French, German, Hungarian, Italian, Japanese, Korean, Malay, Norwegian, Persian, Polish, Portuguese, Russian, Slovak, Slovenian, Spanish, Swedish, Turkish and Ukrainian.

To evaluate the evolution of language coverage, information on the total number of documents indexed in each language was also collected for each year since 2004 (the year Scopus was launched).

Since English dominates the data collected, the results for non-English languages are analysed separately in some cases.

Subjects

A total of 152 WoS Research Areas were included in the results, although 16,550 documents had no subject information (0.56% of WoS). A total of 28 Scopus Subject Areas were found, with 1,135 documents lacking subject information. Adding 382 documents without a defined language, this gives 0.05% of Scopus in 2018.

A broad discipline classification of subjects was done by matching WoS and Scopus subjects. As WoS proposes (Clarivate Analytics, 2019), the following classification was used: Arts & Humanities; Life Sciences & Biomedicine; Physical Sciences; Social Sciences and Technology. All subjects were reclassified under one of these categories except the Scopus subject "Multidisciplinary" (0.67% of Scopus) and subject “Undefined” which was excluded. The reclassification of Scopus materials, therefore, has been carried out based

on the WoS criteria, being distributed as follows: Arts and Humanities as Arts & Humanities; Agricultural and Biological Sciences, Biochemistry Genetics and Molecular Biology, Dentistry, Earth and Planetary Sciences, Environmental Science, Health Professions, Immunology and Microbiology, Medicine, Neuroscience, Nursing, Pharmacology Toxicology and Pharmaceutics and Veterinary as Life Sciences & Biomedicine; Chemistry, Mathematics and Physics and Astronomy as Physical Sciences; Business Management and Accounting, Decision Sciences, Economics Econometrics and Finance, Psychology and Social Sciences as Social Sciences; and Chemical Engineering, Computer Science, Energy, Engineering and Materials Science as Technology. Each document can be classified in more than one category, so the sum of the subjects in all cases is higher than the total number of documents.

Results

Language coverage

As is already known, English dominates both WoS and Scopus (92.64% of the documents indexed in Scopus are in English and this percentage is even higher in the WoS with 95.37% compared to the second language with the highest number of documents in Scopus, Chinese, with 2.76% and the second language in WoS, Spanish, with 1.26%). In general, Scopus offers greater coverage of English and non-English documents (of all types) than WoS. In both cases, Scopus indexes about 100,000 more documents than WoS. In the non-English case this means 25% more documents in Scopus and in the case of English this means 2% more documents in Scopus. Thus, the main advantage of Scopus over WoS is in its greater coverage of non-English documents.

The greater Scopus coverage of non-English documents does not apply to all languages (Figure 1). Among the most represented languages, WoS indexes more documents than Scopus in Spanish (37,113 versus 30,632 documents) and Portuguese (12,560 versus 11,734 documents). WoS also indexes more documents for Catalan, Croatian, Italian, Malay, Norwegian and Turkish. In contrast, Scopus indexes over ten times more documents in Chinese (86,821) than does WoS (7,231). It also indexes more documents in most languages, including Danish, Japanese, Persian and Swedish, where WoS has almost no representation, and Russian, where Scopus has indexed 18,955 documents versus 13,500 for WoS.

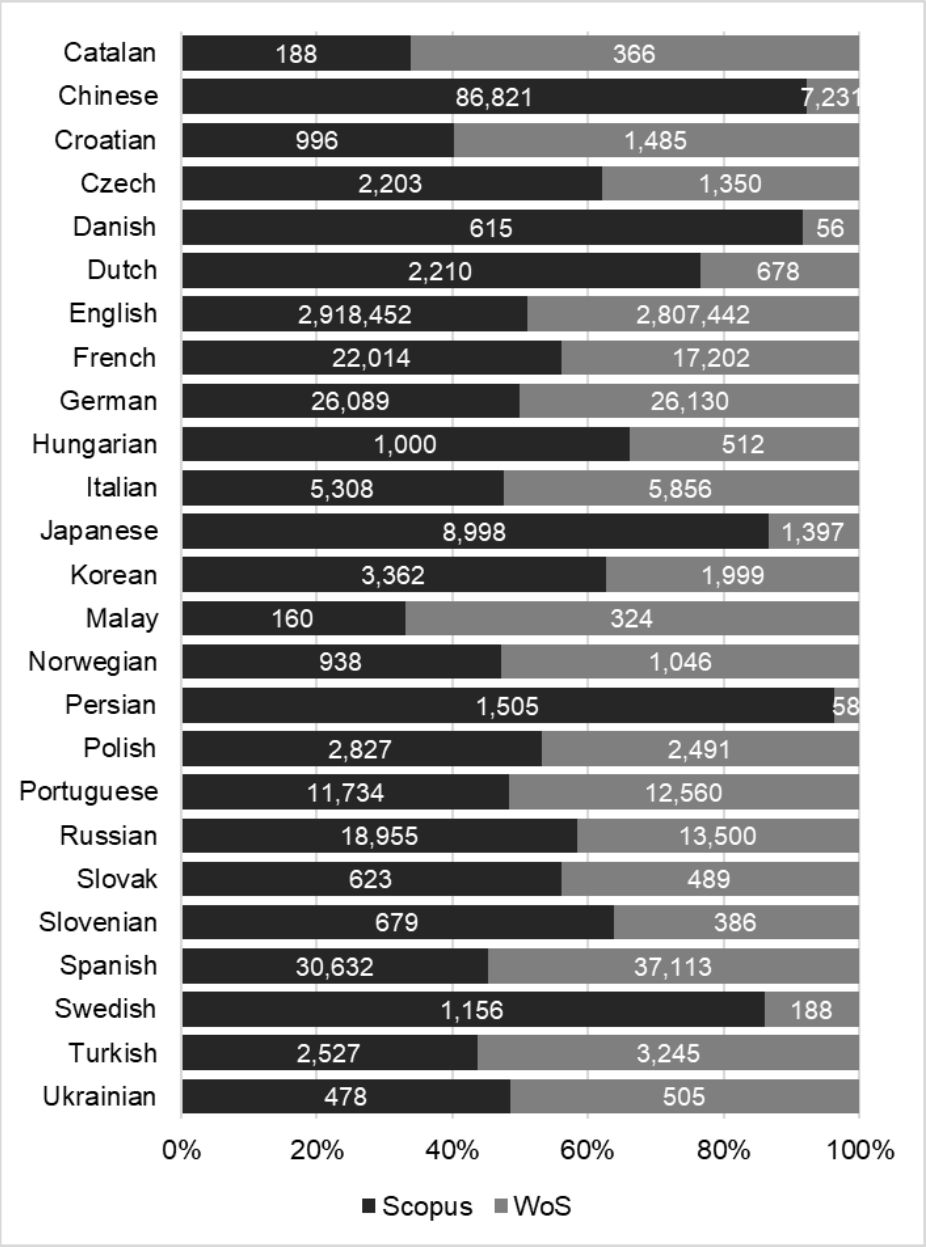

Fig.1 Proportion and number of documents indexed in WoS and Scopus by language.

Research area coverage

Grouping the results by research area, as derived from the WoS and Scopus category names, Scopus has a greater number of indexed documents in all areas except Arts & Humanities, which has the fewest documents (around 1.5%) but its biggest coverage advantage is for Technology (Figure 2).

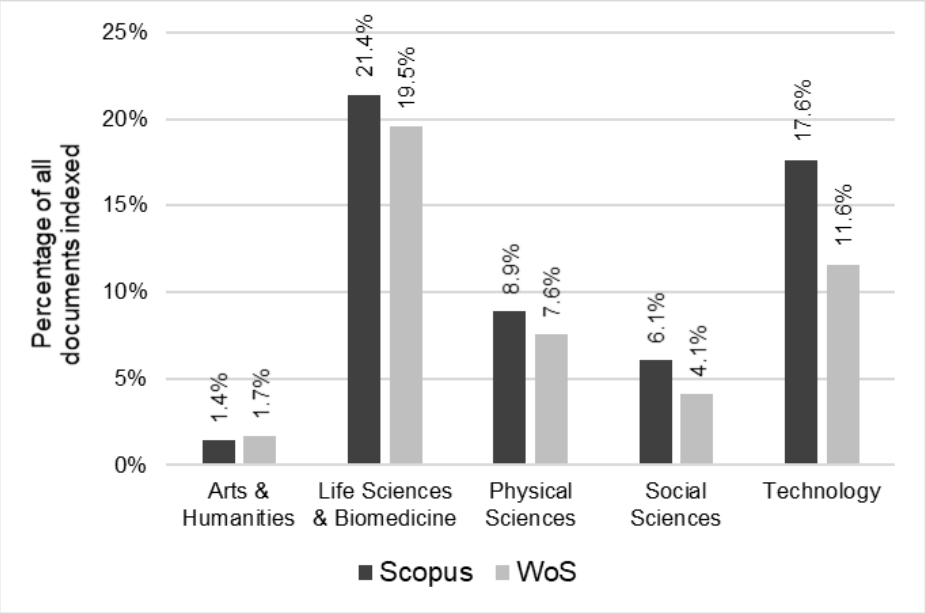

Fig. 2 Proportion of documents indexed in WoS and Scopus by area.

For English documents (Fig. 3), coverage is like that for the overall results (Fig.3). For non-English documents (Fig.4) the difference between Scopus and WoS is much greater, except for Arts & Humanities and Social Sciences.

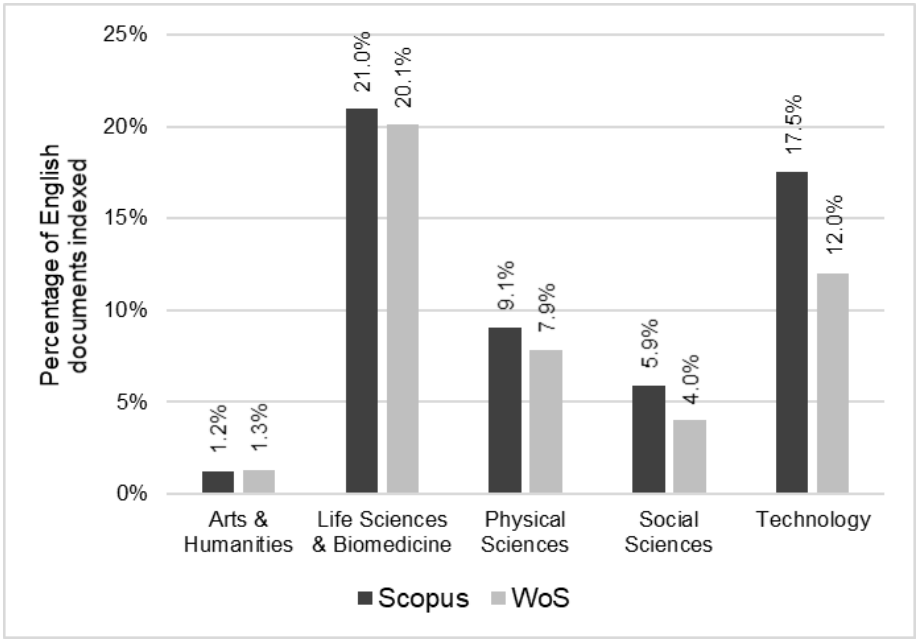

Fig. 3 Proportion of English documents indexed in WoS and Scopus by area.

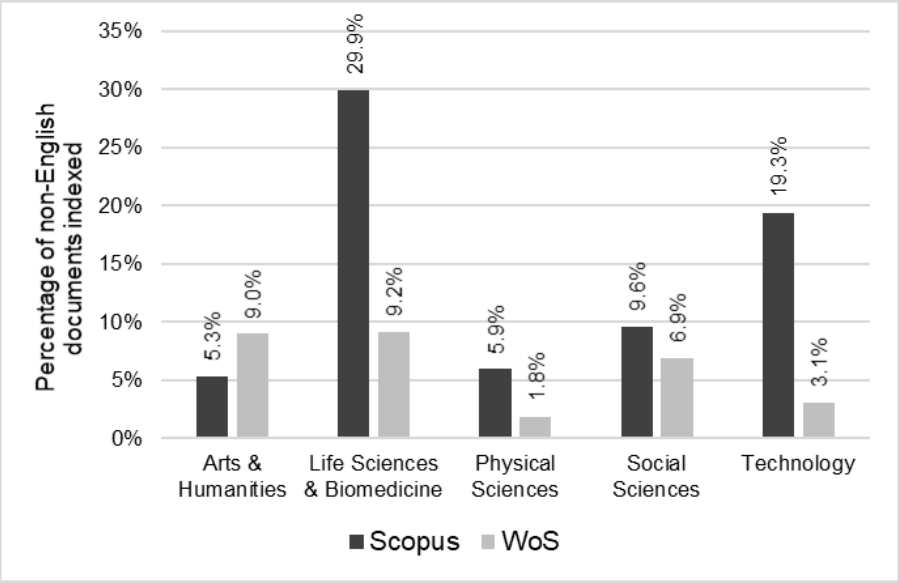

Fig. 4 Proportion of non-English documents indexed in WoS and Scopus by area.

Language	Documents	Source	Documents by source	Graph	Documents by researh area				
					Arts & Humanities	Life Sciences & Biomedicine	Physical Sciences	Social Sciences	Technology
Catalan	554	Scopus	188		108	43	0	158	0
		WoS	366		151	1	0	230	15
Chinese	94,052	Scopus	86,821		766	47,437	21,903	1,487	68,362
		WoS	7,231		383	279	3,790	374	3,047
Croatian	2,481	Scopus	996		377	402	3	374	124
		WoS	1,485		558	142	184	454	192
Czech	3,553	Scopus	2,203		463	1,593	134	475	72
		WoS	1,350		421	584	141	391	16
Danish	671	Scopus	615		35	579	0	2	0
		WoS	56		15	26	12	13	13
Dutch	2,888	Scopus	2,210		169	2,662	0	277	1
		WoS	678		379	212	3	84	12
English	5,725,894	Scopus	2,918,452		116,271	1,968,260	850,269	555,264	1,645,358
		WoS	2,807,442		123,003	1,885,107	737,698	375,557	1,128,087
French	39,216	Scopus	22,014		5,251	15,475	557	8,176	1,331
		WoS	17,202		9,676	5,952	243	3,067	155
German	52,219	Scopus	26,089		3,218	21,477	717	5,259	3,723
		WoS	26,130		6,401	17,083	566	2,768	2,957
Hungarian	1,512	Scopus	1,000		248	663	29	385	114
		WoS	512		2	410	0	64	36

Language	Documents	Source	Documents by source	Graph	Documents by researh area				
					Arts & Humanities	Life Sciences & Biomedicine	Physical Sciences	Social Sciences	Technology
Italian	11,164	Scopus	5,308		2,590	1,907	28	2,606	468
		WoS	5,856		4,440	454	50	1,002	184
Japanese	10,395	Scopus	8,998		255	6,909	969	475	4,123
		WoS	1,397		17	570	527	14	429
Korean	5,361	Scopus	3,362		43	2,032	620	524	2,440
		WoS	1,999		27	363	592	32	1,081
Malay	484	Scopus	160		29	12	3	77	8
		WoS	324		58	28	0	110	88
Norwegian	1,984	Scopus	938		45	842	0	74	0
		WoS	1,046		44	952	0	78	0
Persian	1,563	Scopus	1,505		19	1,403	60	72	99
		WoS	58		34	24	0	0	0
Polish	5,318	Scopus	2,827		558	1,479	373	648	1,027
		WoS	2,491		486	785	547	242	968
Portuguese	24,294	Scopus	11,734		1,302	6,264	692	5,652	2,373
		WoS	12,560		3,290	2,827	492	5,281	1,191
Russian	32,455	Scopus	18,955		2,754	14,205	2,143	5,963	5,177
		WoS	13,500		4,485	2,313	1,194	4,115	1,578
Slovak	1,112	Scopus	623		354	270	2	258	10
		WoS	489		320	44	8	109	23
Slovenian	1,065	Scopus	679		447	133	0	416	20
		WoS	386		217	10	0	139	20
Spanish	67,745	Scopus	30,632		6,546	18,621	662	12,927	3,887
		WoS	37,113		11,678	10,676	482	14,371	2,434
Swedish	1,344	Scopus	1,156		68	1,055	0	84	2
		WoS	188		148	4	0	53	0
Turkish	5,772	Scopus	2,527		276	1,317	112	502	1,267
		WoS	3,245		735	1,425	36	708	743
Ukrainian	983	Scopus	478		56	227	224	88	330
		WoS	505		146	41	153	115	53

Table 1. Number of documents indexed in WoS and Scopus by language and research area.

Introducing language to the research areas analysis (Table 1), for most languages there is a greater WoS coverage in Arts and Humanities and a greater Scopus coverage for the remaining four areas. Some exceptions are described below.

- In the Arts & Humanities, Scopus has more documents indexed in Chinese and Slovenian, with twice as many documents as WoS; in Hungarian and Japanese, where the WoS coverage is negligible; and similar coverage in Czech, Polish and Slovak.
- In Danish, Dutch, Persian and Swedish, the coverage is almost entirely limited to the Scopus Life Sciences & Biomedicine category (no WoS coverage). This also happens

with Norwegian, although in this case WoS has some coverage (53.1% WoS versus 46.9% Scopus).

- Greater WoS coverage occurs for Turkish Life Sciences & Biomedicine, Polish Physical Sciences, Croatian Technology; and similar coverage occurs in Czech Physical Sciences, Korean Physical Sciences and Polish Technology.
- For the Social Sciences, WoS has greater coverage of Catalan, Croatian, Malay, Spanish and Turkish, and similar coverage of Portuguese.

Discussion

The results obtained differ from previous studies in two main aspects. On the one hand, some of these studies are based on data prior to 2015. That year WoS extended its Core Collection including the Emerging Sources Citation Index, which meant an increase in the coverage of non-English languages such as Spanish (15,101 documents in 2014 to 39,504 in 2015), Portuguese (5,628 documents in 2014 to 14,685 in 2015), Russian (3,584 documents in 2014 to 13,665 in 2015), Italian (4,259 in 2014 to 7,587 in 2015) and Turkish (1,742 documents in 2014 to 4,308 in 2015) (Figures 5 and 6).

Fig. 5 WoS Spanish documents evolution.

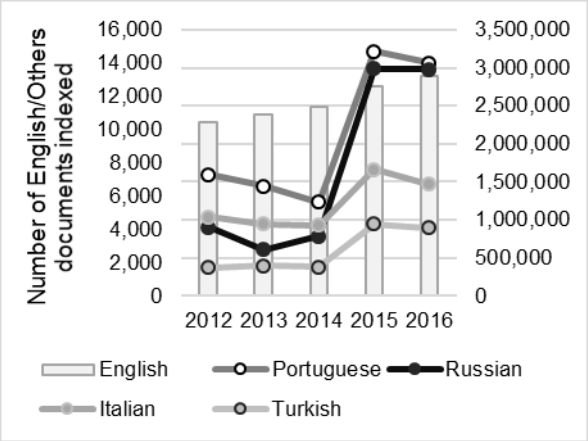

Fig. 6 WoS indexed documents evolution.

On the other hand, it has been possible to verify that aggregate values, both at the language level and research area, differ from the values obtained when the two variables are considered. So, for example, in the case of Spanish, the WoS has a greater number of documents indexed in total computation, but if we analyse the results by research areas, Scopus has a greater number of documents in three of the five areas: Life Sciences & Biomedicine, Physical Sciences and Technology. Similarly, a review Life Sciences & Biomedicine category, where Scopus has a much larger total of documents, the WoS has greater coverage for the Norwegian and Turkish languages. Even if we consider a single source, the overall proportion by research area is different when analysed at language level.

The results obtained suggest that the WoS has made an effort to have better coverage for the area of Arts & Humanities and the main Latin American languages, while Scopus has extensive coverage of Technology, Life Sciences & Biomedicine areas and Asian languages, such as Chinese, Japanese and Korean. Scopus' attention to the Life Sciences & Biomedicine area is especially evident in languages such as Czech, Danish, Dutch, German, Hungarian, Norwegian, Persian or Swedish, where a large majority of indexed documents for these languages belong to this category, with a small proportion of documents for the rest of the areas. Similarly, in languages such as French, German, Italian

and Russian, where Scopus has a greater total number of indexed documents, this majority is maintained in all areas except for Arts & Humanities where the WoS has more documents.

On the other hand, we must consider whether the volume of documents, by itself, is a sufficient indicator that the source is adequate or if other quality factors should be taken into account. We have not checked whether any of the documents in either database are incorrect, for example.

Coverage problems

In addition to the languages that do not appear in any of the two databases and the languages discarded in the methodology for having a representation of less 0.01%, among the languages studied, the following research areas have been identified because the sum of documents from both sources does not reach 0.0015% of the total documents (at least 91 documents per area). Total number of documents in each area and language is indicated:

- Catalan – Physical Sciences (44) and Technology (15).
- Czech – Technology (88).
- Danish – Arts & Humanities (50), Physical Sciences (12), Social Sciences (15) and Technology (13).
- Dutch – Physical Sciences (3) and Technology (13).
- Hungarian – Physical Sciences (29).
- Italian – Physical Sciences (78).
- Korean – Arts & Humanities (70).
- Malay – Arts & Humanities (87), Life Sciences & Biomedicine (40) and Physical Sciences (3).
- Norwegian – Arts & Humanities (89), Physical Sciences (0) and Technology (0).
- Persian – Arts & Humanities (53), Physical Sciences (60) and Social Sciences (72).
- Slovak – Physical Sciences (10) and Technology (33).
- Slovenian – Physical Sciences (0) and Technology (40).
- Swedish – Physical Sciences (0) and Technology (2).

Conclusions

The results obtained at document level, combining language and research area, differ from journal-level analysis and from the average values of aggregate data, even at the same document level. The document level analyses reported here should be more relevant for evaluations because, for example, the omission of a tiny journal would have little effect on any evaluation. In contrast, large coverage of a language through a single huge journal would also not be desirable. Therefore, it is necessary to consider both document and journal coverage to decide which the best source in each case is. A disaggregated analysis at the document level considering the language and the research area is the logical starting point for each language and research area, however. In contrast to previous studies, the current analysis has found some areas in which WoS is better than Scopus and so Scopus is not always the most comprehensive source.

References

Abrizah, A., Zainab, A. N., Kiran, K., & Raj, R. G. (2012). LIS journals scientific impact and subject categorization: A comparison between Web of Science and Scopus. Scientometrics, 94(2), 721–740.

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. Scientometrics, 68(3), 329-342.
Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. Journal of the American society for information science and technology, 60(7), 1320-1326.
Aksnes, D. W., & Sivertsen, G. (2019). A Criteria-based Assessment of the Coverage of Scopus and Web of Science. Journal of Data and Information Science, 4(1), 1-21.
Barnett, P., & Lascar, C. (2012). Comparing unique title coverage of Web of Science and Scopus in Earth and atmospheric sciences. Issues in Science and Technology Librarianship.
Clarivate Analytics. (2019). Web of Science platform: Web of Science: Summary of Coverage. Retrieved from https://clarivate.libguides.com/webofscienceplatform/coverage.
Clermont, M., & Dyckhoff, H. (2012). Coverage of business administration literature in Google Scholar: Analysis and comparison with Econbiz, Scopus and Web of Science. Rochester, NY: Social Science Research Network.
Collazo-Reyes, F. (2014). Growth of the number of indexed journals of Latin America and the Caribbean: the effect on the impact of each country. Scientometrics, 98(1), 197–209.
Delgado, J. L. G., Alonso, J. A., & Jiménez, J. C. (Coords.). (2013). El español, lengua de comunicación científica (Vol. 12). Fundación Telefónica.
Franceschet, M. (2009). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. Scientometrics, 83(1), 243–258.
Fundación Española para la Ciencia y la Tecnología. (2017). Indicadores del sistema español de ciencia, tecnología e innovación. Retrieved from https://icono.fecyt.es/sites/default/files/filepublicaciones/libro_indicadores_2017.pdf
Gavel, Y., & Iselid, L. (2008). Web of Science and Scopus: A journal title overlap study. Online Information Review, 32(1), 8–21.
Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. Nature News, 520(7548), 429.
Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. Journal of Informetrics, 12(4), 1160-1177.
Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. Scientometrics, 116(2), 1153-1180.
Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106(1), 213-228.
National Science Foundation. (2018). Science and engineering indicators, chapter 5: Academic research and development. Retrieved from https://nsf.gov/statistics/2018/nsb20181/assets/nsb20181.pdf
Osca-Lluch, J., Miguel, S., Gonzalez, C., Penaranda-Ortega, M., & Quinones-Vidal, E. (2013). Coverage and overlap of the Web of Science and Scopus in the analysis of the Spanish scientific activity in Psychology. Anales De Psicología, 29(3), 1025–1031.
Santa, S., & Herrero-Solana, V. (2010). Cobertura de la ciencia de América Latina y el Caribe en Scopus vs Web of Science. Investigación bibliotecológica, 24(52), 13-27.
Van Leeuwen, T. N., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & van Raan, A. F. J. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. Scientometrics, 51(1), 335–346.
Vivancos Cervero, V. (Coord.). (2009). El español, lengua para la ciencia y la tecnología. Madrid: Instituto Cervantes.