Machine translation: mapping technological developments through scientometrics

Marileide Dias Esqueda
Doutora em Linguística Aplicada à Tradução pela Universidade Estadual de Campinas (Unicamp – SP – Brasil). Professora Associada da Universidade Federal de Uberlândia (UFU – MG – Brasil)
http://lattes.cnpq.br/3341029625579574
https://orcid.org/0000-0002-6941-7926
E-mail: marileide.esqueda@ufu.br

Flávio de Sousa Freitas
Mestre em Estudos Linguísticos pela Universidade Federal de Uberlândia (UFU – MG – Brasil)
http://lattes.cnpq.br/7428463010998872
https://orcid.org/0000-0002-8972-5870
E-mail: flaviofreitas@ufu.br

Data de submissão: 17/12/2022. Data de aceite: 28/11/2022. Data de publicação:30/12/2022.

ABSTRACT

Conceived as a scientometric study, this paper searches for comprehending the research status of machine translation on the IEEE Xplore database, of the American Institute of Electrical and Electronics Engineers, from 1956 to 2019. Documents were analyzed considering a series of measures such as most prominent academic institutions and countries that investigate machine translation, citation, co-authorship, keywords co-occurrence, reference coupling, and textual-based analysis retrieved from the documents' titles and abstracts. Through VOSviewer software and its tools for data collecting and visualization, machine translation research, in the circumscribed database and period of time, is focused on three main aspects: machine translation systems, statistical machine translation, and English language.

Keywords: Machine Translation. Translation Technologies. Scientometrics.

Tradução automática: mapeando desenvolvimentos tecnológicos por meio da cientometria

RESUMO

Delineado como um estudo cientométrico, este artigo busca compreender o estado da pesquisa em tradução automática na base de dados IEEE Xplore, do instituto americano Institute of Electrical and Electronics Engineers, entre os anos de 1956 e 2019. Os documentos foram analisados segundo uma série de indicadores, tais como as instituições acadêmicas e os países que mais investigam sobre a tradução automática, os índices de citações, a coautoria, a coocorrência de palavras-chave, o acoplamento bibliográfico e os elementos textuais extraídos dos títulos e resumos dos documentos. Com base no software VOSviewer e em suas ferramentas de compilação e análise de dados, as pesquisas em tradução automática, na base de dados e no recorte temporal estabelecidos, centram-se em três aspectos principais: os sistemas de tradução automática, a tradução automática estatística e a língua inglesa.

Palavras-chave: *Tradução Automática. Tecnologias da Tradução. Cientometria.*

Traducción automática: mapeo de desarrollos tecnológicos a través de la cienciometría

RESUMEN

Diseñado como un estudio cientométrico, este artículo busca compreender el estado de la investigación en traducción automática en la base de datos IEEE Xplore, del instituto americano Institute of Electrical and Electronics Engineers, entre los años 1956 y 2019. Los documentos han sido analizados según una serie de mediciones, como las instituciones académicas y los países más destacados que investigan acerca de la traducción automática, los índices de citas, la coautoría, la coocurrencia de palabras clave, el acoplamiento bibliográfico y los elementos textuales recogidos de los títulos y resúmenes de los documentos. Mediante el software VOSviewer y sus herramientas de recopilación y análisis de datos, las pesquisas en traducción automática, en la base de datos y en el recorte de tiempo circunscrito, se centran en tres aspectos principales: los sistemas de traducción automática, la traducción automática estadística y el idioma inglés.

Palabras clave: *Traducción Automática. Tecnologías de la Traducción. Cienciometria.*

CONTEXTUALIZING MACHINE TRANSLATION: GENERAL APPROACHES

Language is a powerful tool in global communication both for industry and academy. With more than 6,800 languages in the world, they will certainly reflect linguistic and cultural diversity (OLADOSU et al., 2016). Accordingly, human translators of the world, throughout history, have been struggling to provide culturally effective translation in order to tackle linguistic and cultural diversity. Nevertheless, as foresaw by Hutchins in the 1980s, there have not been enough translators to cope with the ever-increasing volume of material which has to be translated, and it is in this scenario of the necessity for global communication that machine translation systems have been studied and created throughout history (BOWKER, 2020; HUTCHINS, 1986).

Defined as a subfield of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another (OLADOSU et al., 2016), machine translation (MT) has been experiencing lots of improvements.

The long history of machine translation dates back to the 1950s in the United States, when it was believed that MT would be possible through a rule-based approach, by which researchers tried to program computers to process natural language using grammar rules and linguistic patterns. Since then, researchers have been trying to achieve the independence of the systems from human intervention, searching for developing a fully automatic approach that could be used in different situations with acceptable quality.

In the early 1960s, when the results of tests that have been implemented with the available machine translation systems were released, it was found that machine translation outputs were insufficient to deal with the complexity of natural language problems. An important report produced by the Automatic Language Processing Advisory Committee (ALPAC) put an end to the research on machine translation in the United States, demonstrating that the high cost of its implementation did not offset the tiny results achieved so far (HUTCHINS, 2015; WAY, 2018).

Arguably, this scenario has changed quite considerably, and research in MT has persisted in other countries, such as Canada, France and Germany. For the sake of illustration, in Montreal (Quebec, Canada), research on MT began in 1970 using syntactic transfer techniques for English–French translation. One of the achievements of the TAUM project (Traduction Automatique de l'Université de Montréal – Machine Translation Project created by University of Montreal) was METEO system for translating weather forecasts. Designed specifically for the restricted vocabulary and limited syntax of meteorological reports, METEO has been successfully operating since 1976 (BOWKER, 2020; MARTINS; NUNES, 2005).

As explained by Shiwen and Xiaojing (2015), the significance of linguistic knowledge in MT has been repeatedly reflected upon, which leads to an ever-growing understanding of the role that linguistic knowledge plays in MT. They affirm that:

> A more fundamental issue, however, is how to represent the linguistic knowledge so that it can be processed and utilized by MT systems. Basically, there are two types of formalized knowledge representations: dictionaries and grammar rules on the one hand, and corpora on the other. As explicit representations, the former adopt formal structures, such as relational databases and rewrite rules; as implicit representations, the latter use linear strings of words. (SHIWEN; XIAOJING, 2015, p. 195).

Thanks to computing power and machine learning applied to huge bitext (parallel corpora), apart from i) rule-based approach aforementioned (that followed, in time, by direct, transfer and interlingua models), new approaches and methods have been applied to MT systems. Broadly speaking, they can be named as:

i. corpus-based statistical approach, where computers are trained with parallel corpora and make use of probabilities. According to Kenny and Doherty (2014), in a statistical approach, rather than trying to encode a priori in the form of dictionaries, grammars and knowledge bases, all the linguistic and world knowledge required to translate a text from one language into another (the approach taken in rule-based and knowledge-based MT), systems simply learn how to translate from already existing human translations. In practice, "such learning involves the induction of statistical models of translation from parallel corpora, that is, source texts and their human translations" (KENNY; DOHERTY, 2014, p. 278); and more recently

ii. neural networks, or neural machine translation (NMT), where information processing system that is inspired by the way biological nervous systems, such as the brain, process information (BOWKER; BUITRAGO CIRO, 2019). The neural network machine translation approach finds patterns, such as contextual clues around the source phrase (TORAL; WAY, 2018).

Various approaches have been invented as relatively new methods and techniques for machine translation. Knowledge-based and example-based approaches are examples of such methods and techniques, and they can be considered as extensions of linguistic transfer rule-based approach and corpus-based statistical approach.

Nowadays, most machine translation technologies use a hybrid approach so that they can take advantage of the varied methods and techniques applied so far. It is worth noting that albeit emerged in the 1950s, rule-based machine translation methods perform well between very similar languages that could be considered dialects of each other (MELBY, 2020).

Considering the above, one can say that MT has evolved from different approaches, and neural machine translation, the last generation of MT systems, is rapidly becoming the dominant data-driven approach. Nevertheless, according to Melby (2020) sufficient training data for a viable NMT system are available only for a handful (perhaps twenty) of the over-four-thousand languages in the world. According to the author, for the rest (over 99%), either rule-based or statistical machine translation approaches, or, in most cases, human translation, are the only options.

Despite different possibilities of recounting the history of MT paradigms, which is beyond the scope of this paper, current trends show that MT research has been uniting linguists and computer engineers in long-lasting studies involving morphological, syntactic and semantic analyses, examples extracted from numerous pairs of source sentences and their respective translations. Besides that, some other research initiatives on MT use additional parallel data in which the source text is synthetically created, that is, the source text is machine translated from the target language to be used as a dataset, which has been reported to be a successful way of integrating target-language monolingual data into neural machine translation. Moreover, in-domain data training, for example, e-books are converted to plain texts through specific software, such as Calibre support tools (https://calibre-ebook.com/), and are then sentence-split for determined pair of languages, tokenized and finally sentence-aligned, as explained by Toral and Way (2018).

The facts that MT systems are getting better because they are making use of all traditional and modern MT approaches, and the more they are online and freely accessible to a wide range of external users, the better they get (PYM, 2013), are again attracting the attention of researchers within Translation Studies and other knowledge fields such as Computer Engineering or Computational Linguistics, and one way to follow MT developments and researchers' efforts on MT can be through scientometrics.

SCIENTOMETRICS AS A RESEARCH TOOL

At the present time, following the advances in science and technological innovations, there has been a noticeable increase in studies that allow us to evaluate the academic production of different areas of knowledge, referred to as bibliometrics.

By measuring the productivity of research centers and the intellectual production of their researchers, these studies seek to spotlight the most influential institutions, the areas and sub-areas with the greatest potential for innovation and, consequently, the priorities for allocation of financial resources by government or private institutions.

These studies, both at national and international scenarios, use the techniques of the subdisciplines of Information Science, such as bibliometrics, scientiometrics, informetrics, and webmetrics, to calculate the dissemination and impact of scientific knowledge. Other related subdisciplines have also been recently created, such as cybermetrics (which uses methodologies and results of bibliometric, scientometric or informetric studies associated to the Internet), and altmetrics (alternative metrics usually based on data from the social web). Even though the second word of the above compound nouns (-metrics) suggests the application of mathematical and statistical methods, they use different investigative procedures.

While bibliometrics, a term originally used in 1934 by Paul Otlet in *Traité de documentation* (OTLET, 1934), is dedicated to the measurement of productivity from a set of bibliographic materials, such as books, documents and scientific journals, scientiometrics, which gained notoriety with the beginning of the publication, in 1977, of the journal *Scientometrics* (an International Journal for all quantitative aspects of the Science of Science, Communication in Science and Science Policy), is dedicated to measure the production, on a larger scale, of scientific and technological areas, patents and the ways scientists communicate. On an even larger scale, informetrics investigates databases as a whole, seeking to improve the efficiency of information retrieval, making use of vector space models, Boolean retrieval frameworks, and probabilistic models, among others. For Vanti (2002) and Macías-Chapula (2001), informetrics encompasses bibliometrics, scientiometrics and, more recently, webmetrics, which, with technological advances and the emergence of new forms of online communication, seeks to measure the density of links and *sitations* (a neologism that refers to the ways of measuring how much a website is cited and its impact factor, referred to as Web Impact Factor).

For the Translation Studies researchers Luc van Doorslaer and Yves Gambier, the terms bibliometrics and scientiometrics are not used very consistently in the literature and are sometimes considered synonyms. According to the authors, bibliometrics is a hyperonym or sometimes a hyponym for scientiometrics. As described by Doorslaer and Gambier (2015), regardless of the sources and materials they investigate, these studies have a descriptive power, and seek for measuring the influence of academic centers and their intellectuals,

> When producing and transmitting scientific knowledge, authors weave a web of affinities: they cite some works to the detriment of others; they refer to certain publications; they set up more or less regular intellectual relationships. Nowadays, Translation (and Interpreting) Studies (TS) has the tools (journals, book series, bibliographies, encyclopedias, handbooks, readers, textbooks, etc.) which can trace and visualize outstanding developments in research and the most influential authors and centers so far. (DOORSLAER; GAMBIER, 2015, p. 305)

According to Gile (2015), research involving bibliometrics began in Translation Studies in the 2000s. According to the author, bibliometrics aims at measuring the production of texts and the parameters related to them, as opposed to the more general concept of scientiometrics, which applies to any measurement of scientific activities, and it has firstly started in China, with a series of studies implemented by Gao and Chai (2009), Wang and Mu (2009), Tang (2010), Wang (2015), among others. Gile (2015) also explains that while some of these studies analyze the statistics of the scientific production, others are dedicated to the analysis of citations (GARFIELD, 1997). For Gile (2015, p. 243), these studies evaluate the impact of Translation and Interpretation research on science:

> [...] the more often an author (or journal) is cited, the more influence (s)he has. In sociological terms, citations can be used to identify and track the evolution of research networks and their structure. In institutional terms, for academics, including TS scholars, the so-called impact factor has become important in one's professional development, which has raised some interest within TS. (GILE, 2015, p. 243)

According to the author, citations can also be used as indicators for purposes other than calculating the impact of a particular author or institution as an academic reference center. The analysis of citations can help to ascertain investigative cultures, research paradigms and their evolution, and may, on the one hand, be related to concepts, ideas and opinions associated with certain theoretical schools, and, on the other, the methodologies and results of empirical research, found mainly in publications that convey studies of this type.

Benefiting from advances in science and technological innovations, Translation Studies, as well as other knowledge fields, search for building databases in order to assemble scientific productions and measure their impacts on science.

The data registered in BITRA (Bibliography of Interpreting and Translation - Open Access Bibliography of Translation and Interpretation Studies), was created with this purpose in mind, and has been coordinated since 2001 by Professor Javier Franco Aixelá, from the Department of Translation and Interpretation of the University of Alicante, located in the province of Valencia, Spain. It comprises over 80,000 entries (approximately 10,000 books, 29,000 book chapters, 36,000 journal articles, 3,000 Ph.D. theses, 200 journals, etc.), with more than 42,000 abstracts (over 50% of the entries), 100,000 citations collected in the Impact Factor, and 4,000 tables of contents, involving over 15 languages (FRANCO AIXELÁ, 2020).

Translation Studies Bibliography (TSB), the continuously updated database coordinated by John Benjamins Publishing, has the same purpose of assembling studies of the field and it now contains over 30,000 annotated records. The database has been created in 2010 and it has as editors the Professors Yves Gambier, from University of Turku and Kaunas University of Technology (KTU), located in Lithuania, and Luc van Doorslaer, from University of Tartu and KU Leuven. In December 2019, the bibliography had a new partnership with Guangxi University for structural and substantial supply of Chinese bibliographic records.

In terms of general databases, Google Scholar, Microsoft Academic, CrossRef, Web of Science, Scopus, among others, play also an important role when researchers decide to carry out bibliometric, scientometric or other related analyses in Translation Studies, and in other knowledge fields.

In the case of machine translation, the focus of this paper, Gupta and Dhawan (2019) provide a quantitative and qualitative description of machine translation research published from 2007 to 2016 in Scopus, a database of abstracts and quotations from Elsevier, the Dutch publication and analysis company of scientific, technical and medical content. Gupta and Dhawan (2019) concluded that machine translation research registered a high 12.35 per cent growth, and cumulated 5,181 publications during these nine years. A total of 2,174 authors from 683 organizations and 93 countries contributed to the research during the period. According to Gupta and Dhawan (2019), considering a qualitative perspective, machine translation research averaged a medium level citation impact of 6.03 citations per paper. Only five countries could achieve relative citation index above the world average of 1.17: Canada (2.26), United States of America (2.12), United Kingdom (1.44), Germany (1.38), and France (1.27) during 2007 to 2016. Once the researchers compare the statistics of these countries to India, they have also pointed out another important qualitative conclusion:

> […] Developing world countries have yet to make their impact in this field. […] The main problems that India faces in the area of MT software are syntactic and semantic in nature since each Indian language has own distinct structure. It is not easy to capture such grammatical nuances across languages when it comes to software development for machine translation of Indian languages. Nevertheless, MT in India has over the years made a notable progress in the field. In order to catalyse machine translation research, India needs a long-term policy with a view to prioritise R&D areas in MT, identify role of private sector in system development and identify organizations that have major potential to undertake machine translation research. (GUPTA; DHAWAN, 2019, p. 37)

Accordingly, conceived as a scientometric study, which deals with quantitative and qualitative data, this paper, inspired by the aforementioned studies, searches for comprehending the research status of machine translation. Despite the existing important Translation Studies databases such as BITRA or TSB, or even the available general databases, the IEEE database was chosen as a source for this study.

The IEEE Xplore digital library offers access to scientific content published by the United States Institute of Electrical and Electronics Engineers (IEEE) and its partners. According to information retrieved from its website, this digital library holds more than five million documents from the most cited publications of Electrical Engineering, Computer Science and Electronics. Apart from its robustness of resources for scientometric analyses, this database can be freely accessed through the Internet network of Universidade Federal de Uberlândia, located in Minas Gerais, Brazil, the academic institution to which we are affiliated to.

It is worth noting that, as in Gupta and Dhawan's (2019) study on machine translation research developments, as well as in Voss and Zhao's (2005), or Dong and Chen's (2015) studies on the same topic, even with the intense work in a determined database, it does not mean that authors have total control over everything that may has been published on the topic. In the best of the cases, scientometric studies provide some relevant quantitative and qualitative data of a limited set of data, stimulating other future research topics. As Bornmann and Leydesdorff (2014, p. 1230) aptly put it: "citations need time to accumulate". A reception of a paper can be rather timid in a determined period of time, but it may become one of the highest cited papers in subsequent years or may become associated with some trendy research. In our scientometric study, the results are inevitably tied to the circumscribed time cut and database.

ADDRESSING METHODOLOGICAL PROCEDURES

All research papers, including conference papers and articles, were collected from the IEEE Xplore online database and will be referred to as *documents*. Documents were analyzed considering a series of measures such as most prominent academic institutions and countries that investigate machine translation, citation, co-authorship, keywords co-occurrence, and reference coupling. Moreover, in line with our academic interest as Translation Studies researchers and Translation teachers, documents were also investigated on the interface between Machine Translation and Translation Teaching through textual-based analysis (detailed below). In Doherty *et al*.'s (2018) words:

> It is also true that for decades there was hardly any exchange between MT researchers and developers on the one hand, and professional translators and translation theorists on the other; this was mostly because translators have historically tended to see MT as a threat […], and (like translation theorists) the difficulties that MT faced in the days of rule-based systems were too banal from their point of view to take MT seriously […]. (DOHERTY et al., 2018, p. 99).

Still seeking to meet our interests as Brazilian researchers, we also aim to identify the participation of Brazilian researchers in MT research (see Section 4.1, 4.5 and 4.6).

In view of that, the keyword "machine translation" was applied to IEEE Xplore database in order to collect data through its advanced search feature. After four preliminary searches, through the advanced search applied to all metadata, we have found that some of the documents were not related to machine translation research itself (e.g.: *SMPTE Periodical - Mechanical and Optical Equipment for the Stereophonic Sound-Film System*). The total of recovered documents in these preliminary phases was 4,210.

In order to obtain more reliable data, we have delimited the searches to the titles (document title) and keywords (index terms) of the documents. In this phase, the database has recovered 692 documents. Nevertheless, when exporting the bibliographical data to .ris file extension (RIS - Reference Information System), 78 documents could not be included in the study once they have not presented their DOIs (A Digital Object Identifier is a persistent identifier or handle used to identify objects uniquely, standardized by the International Organization for Standardization).

The IEEE database export feature does not offer the option for automatic retrieval of DOIs, so the .ris file (henceforth RIS file) was accessed to collect and list the DOIs from the 614 documents in another file (henceforth DOIs file). It is worth noting that the creation of the RIS file is particularly useful for scientometric search, once it is only through this format that we can map the terms contained in titles and abstracts using VOSviewer software.

Therefore, data search and collection can be summarized as:

1. Database: IEEE Xplore;
2. Keyword applied to the database: machine translation;
3. Total of recovered documents: 614;
4. Period of time recovered by the database: 1956-2019 (63 years).

Using VOSviewer software (VAN ECK; WALTMAN, 2018), maps for the following scientometric statistics were made:

1. Co-authorship: the repeated occurrence of two or more authors or organizations in a given number of documents;
2. Co-occurrence of keywords: the relationship between keywords and the number of documents in which they concurrently occur;
3. Bibliographic coupling: the relationship between two documents based on the number of common references cited by them;
4. Terms of greater relevance in the titles and abstracts of the documents.

Using the two files (RIS file and DOIs file), maps and visualizations of two types were obtained: maps based on bibliographic data (authors, keywords, year of publication), and maps based on textual data (terms in the titles and abstracts). With the DOIs we were able to recover, using VOSviewer, the authors, the years of publication and the keywords of the documents. In addition, it is important to highlight that, with the file containing the DOIs, VOSviewer works on APIs. The API used was the one that recovered all the documents, in this case COCI (OPENCITATION, 2022).

The list of institutions where the authors are affiliated to was done manually. The task has proved to be particularly challenging because each of the 614 documents were accessed at IEEE Xplore database and the institutional affiliations of each author were collected. Many authors write the name of the same institution differently and several of them use only acronyms to refer to institutions. Unfortunately, VOSviewer software does not build maps with data on institutions.

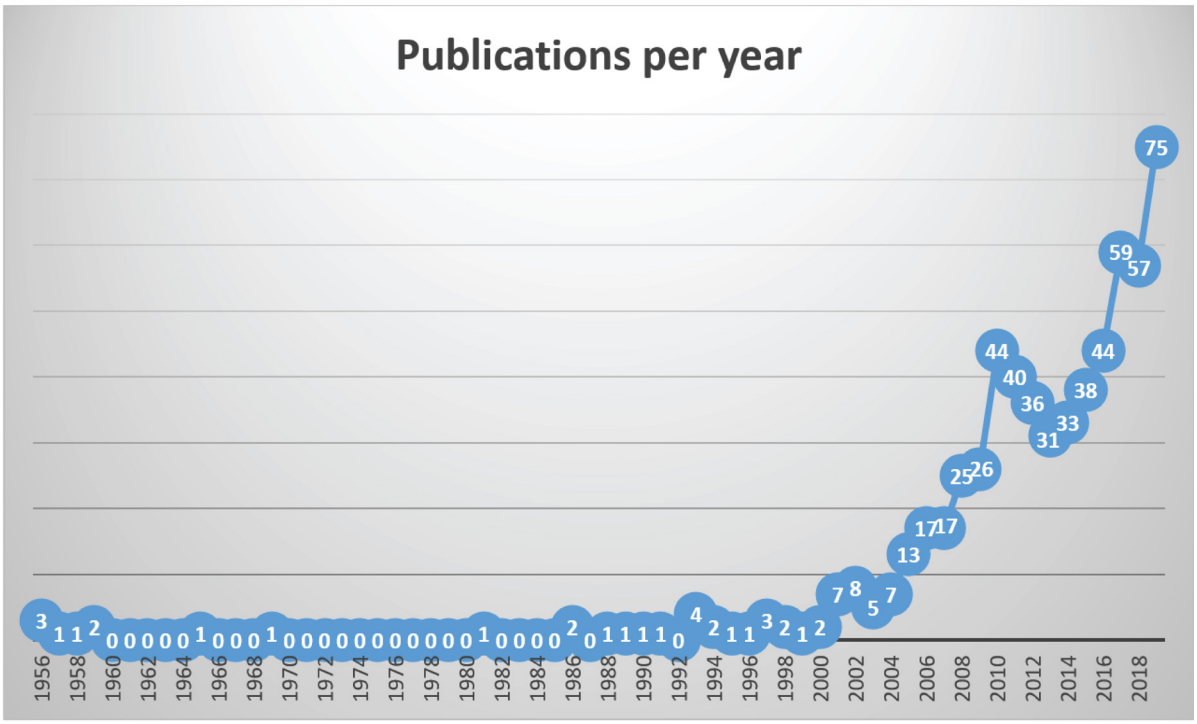
SCIENTOMETRIC ANALYSES

GENERAL ANALYSES

Out of 614, 597 (96%) of the documents were published as conference papers, and 15 (2.44%) as articles, showing the importance of scientific congresses or forums for scientific dissemination of MT research. Although we have applied a filter to retrieve only conference papers and articles, IEEE database has recovered two books (1.56 %), containing synthesis of monographs on topics related to natural language processing, computational linguistics, information retrieval, and spoken language understanding.

Even MT has been researched since 1956, 2010, 2017 and 2019 are the years with the highest number of research papers on MT, with 44, 59, and 75 publications, respectively, as shown in figure 1.

Figure 1 – Grow in MT research over the years: 1956 to 2019



Source: The authors.

As previously mentioned, the 614 documents were accessed at IEEE Xplore database and the institutional affiliations of each author were collected. A total of 451 academic institutions were found in the corpus, and table 1 shows the top 10 institutions that have published the highest numbers of documents on MT (see also GUPTA; DHAWAN, 2019).

Table 1 − Number of published documents per institution

Institutions	N. of publications
Chinese Academy of Sciences, China	22
Beijing Normal University, China	15
Harbin Institute of Technology, China	15
University of Moratuwa, Sri Lanka	14
Soochow University, Suzhou, China	10
Xiamen University, China	10
Amirkabir University of Technology, Tehran, Iran	9
IBM T. J. Watson Res. Center, USA	9
Tsinghua University, China	9
National Institute of Information & Communications Technology, Japan	8

Source: The authors.

China is the country which has the highest number of documents in the corpus, a total of 194 publications. This country also appears in Gupta and Dhawan's (2019) scientometric study, as one of the top 10 most productive countries in MT during 2007-2016 based on Scopus database.

Back to IEEE database, Brazil has registered six documents on MT. Authors from the following Brazilian institutions appear in the database: São Paulo University (São Paulo and São Carlos campuses); CEFET/RJ (Federal Center for Technological Education of Rio de Janeiro, Rio de Janeiro campus); Federal University of São Carlos, São Carlos; Federal University of Amazonas, Manaus; and Pontifical Catholic University of Rio Grande do Sul, Porto Alegre; Federal University of Paraná, Curitiba. Fours documents are registered as conference papers, one as an article, and one as a book. These Brazilian documents have been written in coauthorship with researchers from University of Bari, Italy, and University of Sheffield, England.

A total of 1,554 authors participated in machine translation research during 1956-2019, according to the data retrieved from IEEE database. Table 2 shows the list of the top 10 authors who have the highest ranks of productivity.

Table 2 − Ten authors ranked top in terms of their research productivity

Authors and their institutions	Number of publications
1. Yaohong Jin (Beijing Normal University, China)	15
2. Tiejun Zhao (Harbin Institute of Technology)	11
3. Shahram Khadivi (Amirkabir University of Technology, Iran)	10
4. Eiichiro Sumita (National Institute of Information & Communications Technology, Japan)	9
5. Deyi Xiong (Tianjin University, China)	7
6. Fuji Ren (The University of Tokushima, Japan)	7
7. Hemant Darbari (Centre for Development of Advanced Computing, India)	7
8. Ayu Purwarianti (Bandung Institute of Technology, Indonesia)	7
9. Qun Liu (Chinese Academy of Sciences, China)	7
10. Bo Xu (Chinese Academy of Sciences, China)	7

Source: The authors.

Results show that there is no direct relationship between the number of published documents per institution (table 1) and the ten authors ranked top in terms of their research productivity (table 2), because some institutions have many researchers who publish low numbers of documents.

Even though Chinese Academy of Sciences is the institution with the highest number of publications, only two of its authors are amongst the ten most productive authors in table 2. In addition, some authors are affiliated to more than one institution. Deyi Xiong, for example, published seven works that were attributed to four different institutions. The institutions in table 2 are the authors' current affiliations. In order to reveal more reliable data, journals should better guide authors when it comes to credit their institutions.

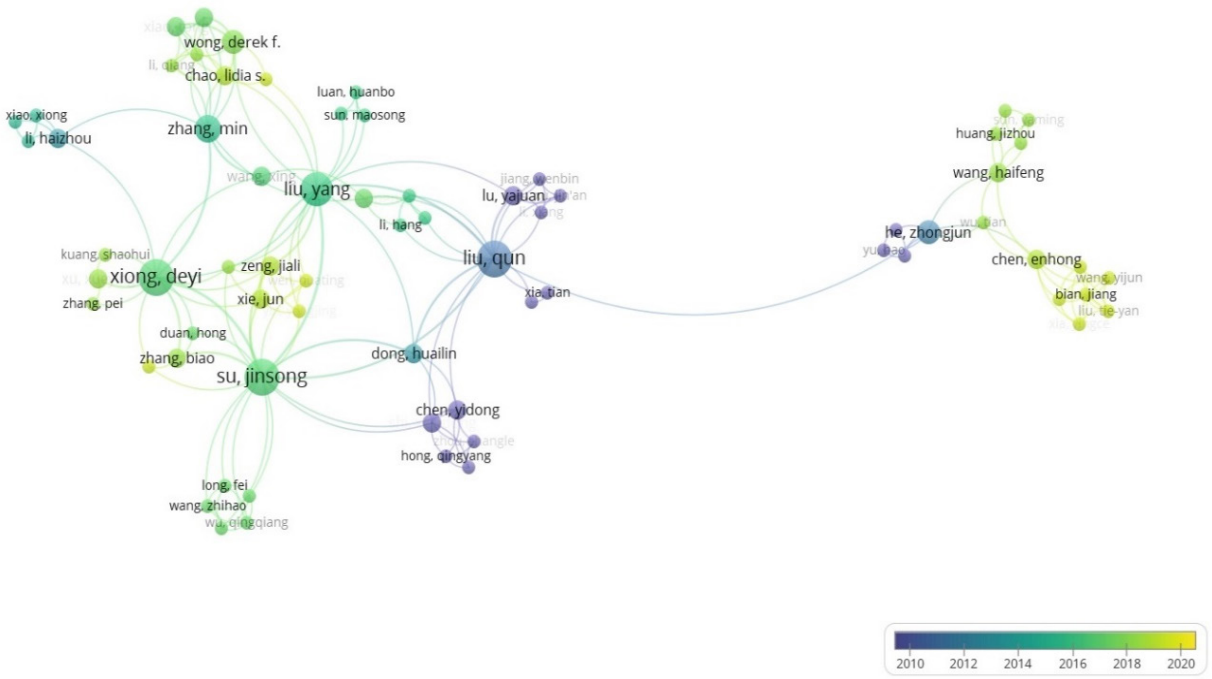
CO-AUTORSHIP

Co-authorship was measured using COCI API and the DOIs file. Using authors as a unit of analysis in the metadata, the counting method used was full counting, and the minimum number of documents per author was one. Out of 1,554 authors, 69 perform a co-authorship network and are distributed in 11 clusters. The largest cluster is composed of 10 authors and, amongst them, two authors who collaborate the most with other authors are Zhongjun He and Haifeng Wang.

One of the advantages of scientiometric studies is particularly the fact that they can update our view of scientific production. Research publications (books, conference papers and articles), considered the most prestigious types of academic publication, show the evolution of science in terms of content and form. In the past, science seems to be performed in an unsystematic way by rich and curious men, while today it has reached a global and organized system, shown by co-authorship networks (OLOHAN, 2016).

figure 2 shows the VOSviewer network visualization of the co-authorship map, where yellow and light yellow circles represent the authors which have co-authorship networks, ranging from years 2010 to 2019.

Figure 2 – Overlay visualization of co-authorship map



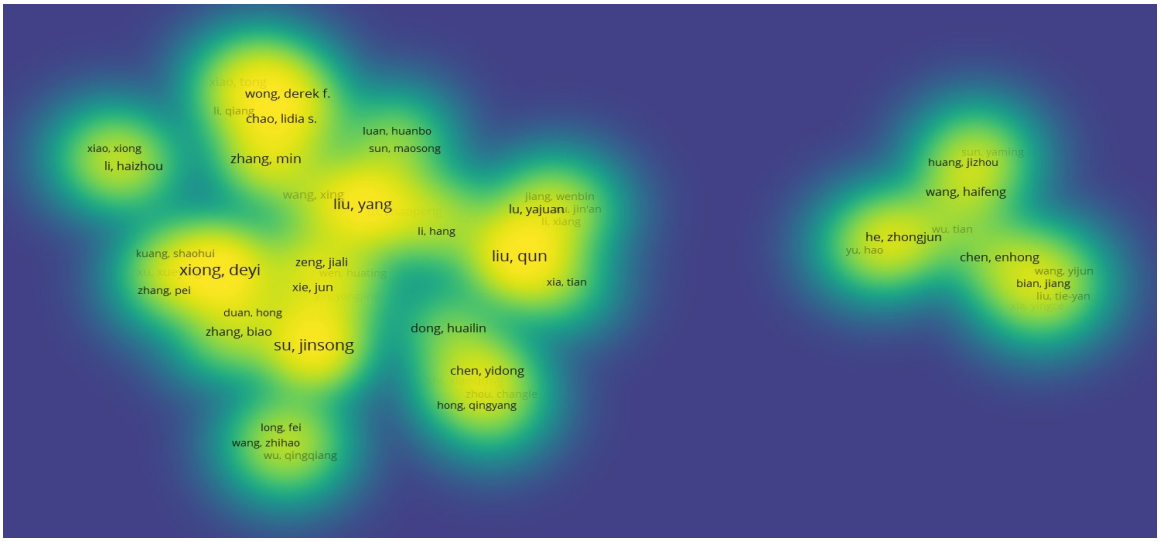
Source: The authors using VOSviewer Software (2020).

In figure 3, each item on the map, the authors' names, are displayed in shades of yellow. The more authors collaborate one another, the closer the authors' names will be to golden yellow. Likewise, the less authors collaborate, the closer the authors' names will be to light yellow or green.

BIBLIOGRAPHIC COUPLING

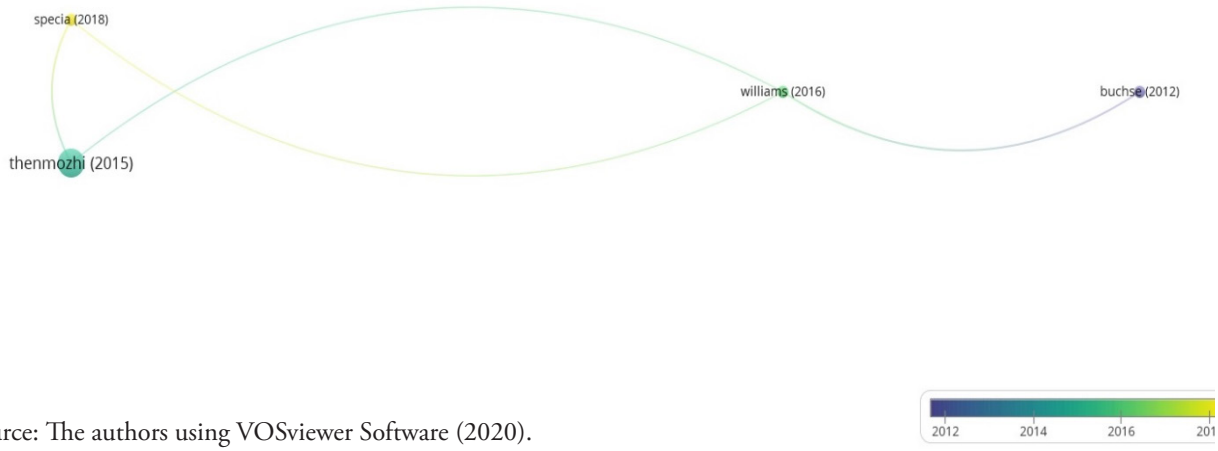
Bibliographic coupling was measured using COCI API, and the DOIs files. The unit of analysis in the metadata was documents with a full counting method. A minimum number considered for bibliographic coupling per document was one. Out of 614 documents, 186 have at least one bibliographic coupling. Out of 186, four documents have the highest rates of bibliographic coupling and are distributed in two groups, each group containing two documents. The document that has the highest connection rate with other documents is *Syntax-based statistical machine translation*, written by Williams *et al.* (2016). The most cited work of the two clusters, however, is the article *Paraphrase identification by using clause-based similarity features and machine translation metrics*, by Thenmozhi and Aravindan (2015) as shown by figure 4.

Figure 3 – Visualization of density of the co-authorship map



Source: The authors using VOSviewer Software (2020).

Figure 4 – Overlay visualization of the bibliographic coupling map



Source: The authors using VOSviewer Software (2020).

KEYWORDS

Based on the RIS file, and using keywords as the unit of analysis in the metadata, with a full counting method, the minimum number of occurrence of a keyword was one. Results show that 3,407 keywords have been retrieved, all of them occurring at least once in the metadata. When making the map we have selected all keywords. Out of 3,407 keywords, 3,345 form a network of co-occurrence, and are distributed in 86 clusters. The largest cluster is composed of 102 keywords. Amongst the 102 keywords, the three keywords that have the highest co-occurrence rate are "machine translation" (even though this was the keyword applied to the database), "statistical machine translation" and "neural machine translation", as shown in figure 5.

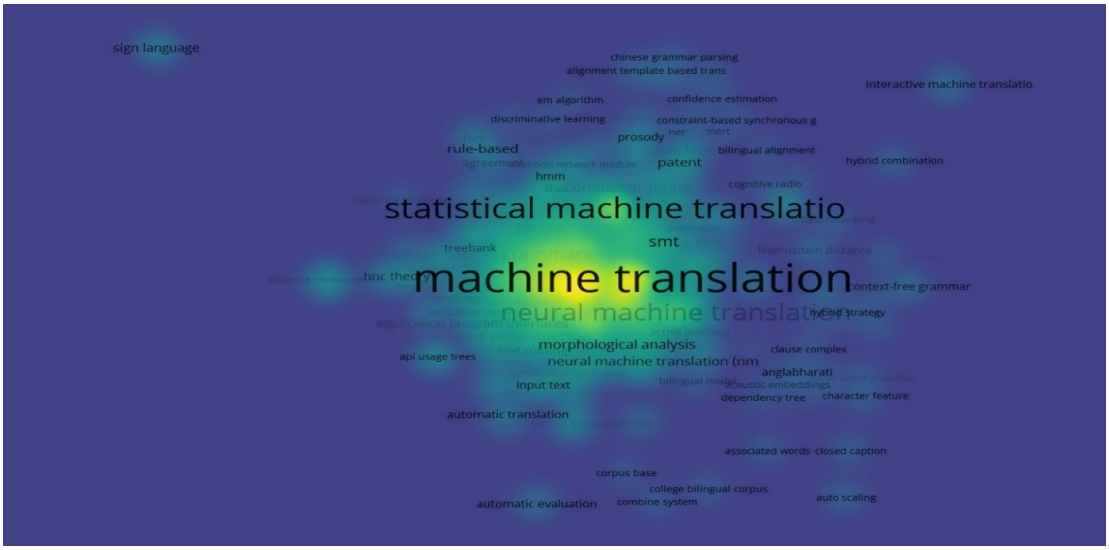
It is worth noting that VOSviewer software has a limit of 30 characters to show texts in the map. This is why instead of "statistical machine translation" in figure 6, the final "n" in "translation" is missing. The same occur in the other maps.

Figure 5 – Co-occurrence of keywords

Selected	Keyword	Occurrences	Total link strength
☑	machine translation	219	1668
☑	language translation	217	4505
☑	natural language processing	166	2970
☑	statistical machine translation	105	980
☑	natural languages	73	1301
☑	training	61	1417
☑	decoding	57	1203
☑	neural machine translation	57	1086
☑	statistical analysis	46	1044
☑	dictionaries	42	770
☑	computational linguistics	37	706
☑	humans	34	591
☑	semantics	33	756
☑	text analysis	32	769
☑	learning (artificial intelligence)	31	746
☑	vocabulary	31	656
☑	neural nets	29	715

Source: The authors using VOSviewer Software (2020).

Figure 6 – Density display of the co-occurrence map of keywords



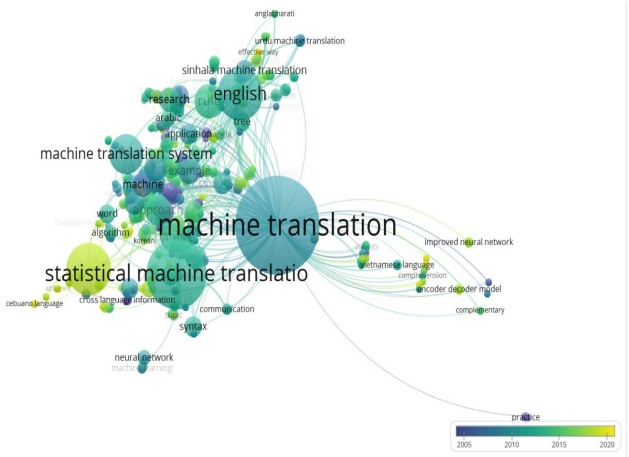
Source: The authors using VOSviewer Software (2020).

TERMS CO-OCCURRENCE IN TITLES

In order to identify possible content related to machine translation research, textual-based maps were made from terms that occur in the titles of the documents. Using the RIS file, a minimum number of occurrences of a term was one. We have selected "all" as counting method.

Out of 1,165 terms, 863 perform a network of co-occurrence of terms in the titles and are distributed in 130 clusters. The largest cluster is composed of 35 terms. The terms that most co-occur are "machine translation", "statistical machine translation", "machine translation system" and "English", as displayed in figure 7.

Figure 7 – Overlay map of co-occurrence of terms in the documents' titles

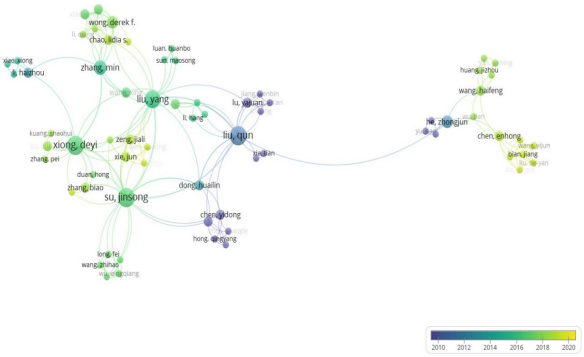


Source: The authors using VOSviewer Software (2020).

TERMS CO-OCCURRENCE IN ABSTRACTS

Similarly, textual-based maps were made from terms that occur in the abstracts of the documents. Using the RIS file, a minimum number of occurrences of a term was one, and so we have selected the full counting method. Out of 8,608 retrieved terms, all terms occur at least once in the abstracts. 8,608 terms form a network of co-occurrence of terms in the abstracts and are distributed in 84 clusters. The largest cluster is composed of 172 terms. The three most co-occurring terms are "paper", "English sentence" and "target word", as displayed in figure 8.

Figure 8 – Network visualization of the map of co-occurrence of terms in the documents' abstracts



Source: The authors using VOSviewer Software (2020).

Based on the textual-based maps, "machine translation", "statistical machine translation", "machine translation system", "English language", "target word" and "paper" are highly considered in the titles and abstracts for describing and identifying MT research. Languages other than English are not included, and neural machine translation, the last tendency in machine translation technology, has not yet been highly investigated compared to statistical machine translation, at least in IEEE database and in the circumscribed period of time. It is worth noting that "paper" is one of the top words retrieved from metadata, probably due to the fact that researchers commonly use this word when expressing the purposes of theirs papers, such as in the very common sentence: "This paper aims at …"

SUMMARIZING AND CONCLUDING

This scientometric study provides a quantitative and qualitative description of machine translation research published during the period of 1956 to 2019. The data for the study was retrieved from IEEE database covering 63 years.

Machine translation research registered an increase in the number of publications ranging from 0 to 3 a year in the 1950s to 75 in 2019.

As highlighted in the introduction of this paper, machine translation systems have been increasingly studied to cope with the ever-growing volume of material which has to be translated for global communication purposes (HUTCHINS, 1986, 2015; BOWKER, 2020). Once MT can be applied to e-learning, e-health, commerce, government organizations, production of scientific and technical documentation, localization of software, speech translation, information retrieval and information extraction, inter alia, the study, creation and fine-tuning of this technology show no signs of reversal (JIMÉNEZ-CRESPO, 2018; MOORKENS et al., 2018).

A total of 1,554 authors from 451 institutions contributed to the research on machine translation. At a qualitative level, five countries have achieved the highest ranks: China, Sri Lanka, Iran, USA, and Japan. As already affirmed by Gupta and Dhawan (2019), developing world countries have yet to make their impact in this field. Both India and Brazil, based on scientometric studies, need a long-term policy to identify academic institutions and organizations that have major potential to undertake machine translation research.

The results of this scientometric study also show that, as affirmed by Doherty *et al.*'s (2018), there is still missing an exchange between MT researchers and developers and Translation Studies. Considering the textual-based analysis retrieved from titles and abstracts here investigated, Translation Studies researchers, human translators, teachers and students, and machine translation teaching and learning have not yet been included in MT research. We advocate that if Translation Studies researchers, human translators, teachers and students are all excluded from MT research, they will continue to use MT outputs "as is" (MELBY, 2020), without considering possible alternatives offered by different degrees of MT technologies, such as pre-editing and post-editing (BOWKER, 2020; BOWKER; BUITRAGO CIRO, 2019).

In a nutshell, we believe that to improve both human and machine translation, MT researchers and developers, and Translation Studies researchers and translation professionals should search for better practices to collaborate one another, particularly because machine translation systems are not agentless, or timeless, but built thanks to collaborative human labor.

REFERENCES

BORNMANN, L.; LEYDESDORFF, L. Scientometrics in a changing research landscape. *EMBO Reports*, v. 15, pp. 1228-1232, 2014. DOI: https://doi.org/10.15252/embr.201439608. Available at: https://www.embopress.org/doi/full/10.15252/embr.201439608. Latest access: Apr. 28 2020.

BOWKER, L. Fit-for-purpoe translation. *In:* O'HAGAN, M. *The Routledge Handbook of Translation and Technology*. New York: Routledge, 2020. pp. 319-335.

BOWKER, L.; BUITRAGO CIRO, J. *Machine Translation and Global Research*: Towards Improved Machine Translation Literacy in the Scholarly Community. United Kingdom: Emerald Publishing Limited, 2019.

DOHERTY, S. *et al*. On Education and Training in Translation Quality Assessment. *In*: MOORKENS, J.; CASTILHO, S.; GASPARI, F.; DOHERTY, S. (ed.). Translation Quality Assessment: From Principles to Practice. Machine Translation Series. Switzerland: Springer International Publishing, 2018. pp. 95-106.

DONG, D.; CHEN, M.-L. Publication trends and co-citation mapping of translation studies between 2000 and 2015. *Scientometrics*, v. 105, n. 2, pp. 1111-1128, 2015. DOI: https://doi.org/10.1007/s11192-015-1769-1. Available at: https://link.springer.com/content/pdf/10.1007/s11192-015-1769-1.pdf. Latest access: Apr. 29 2020.

DOORSLAER, L. V.; GAMBIER, Y. Measuring relationships in Translation Studies. On affiliations and keyword frequencies in the Translation Studies Bibliography. *Perspectives:* Studies in Translatology, v. 23, n. 2, pp. 305-319, 2015. DOI: 10.1080/0907676X.2015.1026360. Available at: https://www.tandfonline.com/doi/pdf/10.1080/0907676X.2015.1026360?needAccess=true. Latest access: Apr. 29 2020.

FRANCO AIXELÁ, J. 2001-2020. *BITRA (Bibliography of Interpreting and Translation).* Open-access database. DOI: 10.14198/bitra. Available at: http://dti.ua.es/en/bitra/introduction.html.

GAO, B.; CHAI, M. A bibliometric analysis of new developments in simultaneous interpreting studies in the West. *Chinese Translators Journal*, v. 2, pp. 17-21, 2009.

GARFIELD, E. Validation of citation analysis. *Journal of the American Society for Information Science*, v. 48, pp. 962-964, 1997.

GILE, D. Analyzing Translation studies with scientometric data: from CIRIN to citation analysis. *Perspectives:* Studies in Translatology, v. 23, n. 2, pp. 240–248, 2015. DOI: https://doi.org/10.1080/0907676X.2014.972418. Available at: https://www.tandfonline.com/doi/pdf/10.1080/0907676X.2014.972418?needAccess=true. Latest access: Apr. 29 2020.

GUPTA, B. M.; DHAWAN, S. M. Machine Translation Research: A Scientometric Assessment of Global Publications Output during 2007-16. *DESIDOC Journal of Library & Information Technology*, v. 39, n. 1, pp. 31-38, 2019. DOI: 10.14429/djlit.39.1.13558. Available at: https://www.researchgate.net/profile/Brij_Mohan_Gupta/publication/331326073_Machine_Translation_Research_A_Scientometric_Assessment_of_Global_Publications_Output_during_2007_16/links/5dd79fce92851c1feda58cc8/Machine-Translation-Research-A-Scientometric-Assessment-of-Global-Publications-Output-during-2007-16.pdf. Latest Access: Apr. 20 2020.

HUTCHINS, W. J. Machine translation: history of research and applications. *In*: SIN-WAI, C. (ed.). The Routledge encyclopedia of translation technology. London/New York: Routledge, 2015. pp. 120-136.

HUTCHINS, W. J. Machine translation: past, present, future. Chichester: Ellis Horwood, 1986.

JIMÉNEZ-CRESPO, M. A. Crowdsourcing and Translation Quality: Novel Approaches in the Language Industry and Translation Studies. *In:* MOORKENS, J.; CASTILHO, S.; GASPARI, F.; DOHERTY, S. (ed.). Translation Quality Assessment: From Principles to Practice. Machine Translation Series. Switzerland: Springer International Publishing, 2018. pp. 69-93.

KENNY, D.; DOHERTY, S. Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer*, v. 8, n. 2, pp. 276-294, 2014. DOI: https://doi.org/10.1080/1750399X.2014.936112. Available at: https://www.tandfonline.com/doi/pdf/10.1080/1750399X.2014.936112?casa_token=N5ph6XsUy-4AAAAA:QFS5MTPuKHEs3XJprB2bhJr7qq8bFCnUfJc2fZeSqrz4a9mOf4X4Q72Zbm4ow1vNxPQ3uvfqlyDtZGiU. Latest access: Apr. 21 2020.

MACÍAS-CHAPULA, C. A. Papel de la informetría y de la cienciometría y su perspectiva nacional e internacional. *Acimed*, v. 9, pp. 35-41, 2001. Available at: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352001000400006. Latest access: Apr. 29 2020.

MARTINS, R. T.; NUNES, M. G. Noções gerais de tradução automática. *Notas didáticas do ICMC – USP*, São Carlos, n. 68, pp. 1-26, 2005. Available at: http://www.nilc.icmc.usp.br/nilc/download/NotasDidaticasICMC_68.pdf. Latest access: Apr. 2 2020.

MELBY, A. K. Future of machine translation: Musings on Weaver's memo. *In:* O'HAGAN, M. The Routledge Handbook of Translation and Technology. New York: Routledge, 2020. pp. 419-436.

MOORKENS, J.; CASTILHO, S.; GASPARI, F.; DOHERTY, S. (ed.). *Translation Quality Assessment*: From Principles to Practice. Machine Translation Series. Switzerland: Springer International Publishing, 2018.

OLADOSU, J. et al. Approaches to Machine Translation: A Review. *FUOYE Journal of Engineering and Technology*, v.1, n.1, pp. 120-162, 2016. Available at: http://engineering.fuoye.edu.ng/journal/index.php/engineer/article/view/26/pdf. Latest access: Apr. 20 2020.

OLOHAN, M. *Scientific and Technical Translation*. London and New York: Routledge, 2016.

OPENCITATION. *COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations*. 2022. Available at: https://opencitations.net/index/coci. Latest access: Dec. 6 2022.

OTLET, P. *Traité de documentation*: le livre sur le livre – théorie et pratique. Bruxelas: Editiones Mundaneum, 1934.

PYM, A. Translation Skill-Sets in a Machine-Translation Age. *Meta*, v. 58, n. 3, pp. 487–503, 2013. DOI: https://doi.org/10.7202/1025047ar. Available at: https://www.erudit.org/en/journals/meta/2013-v58-n3-meta01406/1025047ar.pdf. Latest access: Apr. 1 2020.

SHIWEN, Y.; XIAOJING, B. Rule-based machine translation. *In*: SIN-WAI, C. (ed.). The Routledge encyclopedia of translation technology. London/New York: Routledge, 2015, pp. 186-200.

TANG, F. A bibliometric analysis of empirical interpreting studies in China: Based on data of experimental research papers. *Foreign Language World*, n. 2, pp. 39-46, 2010.

THENMOZHI, D.; ARAVINDAN, C. Paraphrase Identification by Using Clause-Based Similarity Features and Machine Translation Metrics. *The Computer Journal*, v. 59, n. 9, pp. 1289-1302, 2015. DOI: https://doi.org/10.1093/comjnl/bxv083.

TORAL, A.; WAY, A. What Level of Quality Can Neural Machine Translation Attain on Literary Text? *In:* MOORKENS, J.; CASTILHO, S.; GASPARI, F.; DOHERTY, S. (ed.). *Translation Quality Assessment*: From Principles to Practice. Machine Translation Series. Switzerland: Springer International Publishing, 2018, p. 263-287.

TRANSLATION STUDIES BIBLIOGRAPHY. Amsterdam: John Benjamins Publishing Company. Available at: https://benjamins.com/online/tsb/.

VAN ECK, N. J.; WALTMAN, L. *VOSviewer manual*. Leiden: Univeristeit Leiden, v. 1, n. 1, 2018. Available at: https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.10.pdf. Latest access: Apr. 29 2020.

VANTI, N. A. P. Da bibliometria à webmetria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. *Ciência da Informação*, Brasília, DF, v. 31, n. 2, pp. 152–162, 2002. DOI: http://dx.doi.org/10.1590/S0100-19652002000200016 . Available at: https://www.scielo.br/scielo.php?pid=s0100-19652002000200016&script=sci_arttext&tlng=pt. Latest access: Apr. 20 2020.

VOSS, S.; ZHAO, X. Some steps towards a scientometric analysis of publications in machine translation. IASTED In Conference on Artificial Intelligence and Applications. Innsbruck, Austria: Multi-Conference on Applied Informatics, 2005. pp. 651-655.

WANG, B. Describing the progress of interpreting studies in China: A bibliometrical analysis of CSSCI/CORE journal articles during the past five years. *Babel*, v. 61, n. 1, pp. 62-77, 2015.

WANG, B.; MU, L. Interpreter training and research in mainland China: Recent developments. *Interpreting*: International Journal of Interpreting Theory and Practice, n. 11, pp. 267–283, 2009.

WAY, A. Quality Expectations of Machine Translation. *In:* MOORKENS, J.; CASTILHO, S.; GASPARI, F.; DOHERTY, S. (ed.). *Translation Quality Assessment:* From Principles to Practice. Machine Translation Series. Switzerland: Springer International Publishing, 2018, pp. 159-178.

WILLIAMS, P. *et al.* Syntax-based statistical machine translation. *Synthesis Lectures on Human Language Technologies*, v. 9, n. 4, pp. 1-208, 2016.