

Prospects for Lexical Simplification: Bibliometric Analysis and New Directions of Research

Lisnawita^{1*}, Juhaida Abu Bakar²

{Lisnawita@unilak.ac.id¹, juhaida.ab@uum.edu.my²}

¹Universitas Lancang Kuning, Pekanbaru, Indonesia

^{1,2}School of Computing, Universiti Utara Malaysia, Malaysia

Abstract. This research investigates trends and advances in the field of Lexical Simplification (LS), namely efforts to simplify text. We used data from the Scopus database to see the types of publications that appeared most frequently. As a result, conference papers were the largest (65.6%), followed by journal articles and conference reviews. We also found that the number of publications fluctuated every year, with 2016 and 2020 being the peaks, indicating that there is a need to be more consistent and produce more publications in the future. This research also highlights the important contributions of several researchers and universities. From a language perspective, LS research mostly uses English, but some also uses Spanish and Mandarin. The main publication source is *Lecture Notes in Computer Science*. Keyword analysis shows that the main topics of this research are Lexical Simplification (18.34%), natural language processing systems (15.65%), computational linguistics (9.54%), semantics (7.85%), and Complex Word Identification (6.85%). These findings help understand the current state and future directions of LS research, emphasizing the importance of language variety and improving the quality of publications in this field.

Keywords: Lexical Simplification, Publication Trends, Bibliometric Analysis, Research Topics

1 Introduction

Language use and text comprehension are now essential components of many applications in the digital age, such as machine translation, automated text processing, machine learning, and many more areas. Planning future study directions therefore requires an understanding of the evolution of lexical simplification as well as the identification of prominent research trends.

Knowledge and communication technology advancements have broadened the scope of knowledge and increased accessibility to content for a worldwide audience. Individual variations in language comprehension skills, however, make it difficult to guarantee that knowledge can be shared inclusively and accessible to all. Lexical simplification (LS), which is the process of substituting simpler words or phrases for more complex ones without altering their original meaning, is crucial in this situation to promote understanding without omitting crucial details. This methodology has been used in studies on disorders including aphasia,[1][2] non-native speakers[3], dyslexia,[4] and autistic children[5][6], highlighting the urgent need for text simplification to support these groups. In addition, research on lexical simplification has been

conducted by many researchers in various languages, including English [7]–[15][16], French[17], Spanish[11], [18][19], Indonesian[20], Malaysian[3], Chinese[21], Portuguese[11], [22].

This study attempts to provide a clearer picture of the ways in which advancements in natural language processing have been influenced by lexical simplification and offers insightful information about areas that require more investigation and improvement. This research aims to offer useful insights for scholars, practitioners, and other parties interested in the development of Lexical Simplification and its future applications through bibliometric analysis and examination of current trends.

The necessity of lexical simplification in the context of developing information accessibility is first discussed in this paper. We then go over the bibliometric analysis methodology that was applied to gather and examine publication-related data. This study provides insights into methodological and technological advancements that have advanced text simplification by delving further into the literature. Using this method, the study aims to give a broad perspective of lexical simplification's future and how it may influence text simplification techniques to produce more inclusive and successful communication. This study aims to investigate recent publication trends in LS research, identify key researchers and institutions in the LS field, analyze the geographical distribution of LS research, uncover keywords and main topics in LS, ascertain major journals or publications covering LS, and determine the topics with the greatest influence on LS through clustering.

2 Literature Review

This section discussed on the overview of relevant studies is provided, offering insights into various approaches and methodologies employed in the field of lexical simplification.

2.1 Previous Studies

The 2020 study [15] presented an unsupervised machine learning technique called Bidirectional Encoder Representations from Transformers (BERT) for lexical simplification. With scores of 0.776 on LexMTurk5, 0.607 on BenchLS, and 0.423 on NNSeval, this work by Jipeng Qiang et al. [15] demonstrates that this strategy, called BERT-LS, enhances lexical simplification ability by 12% over baseline. Nevertheless, this approach is restricted to replacing a single word, whereas other approaches can replace many words.

Research by John Lee and Chak Yan Yeung [10] highlights the value of customization in lexical simplification by using Complex Word Identification (CWI) that is adapted to each user's level of vocabulary knowledge. According to this study, CWI personalization improved text simplification accuracy and precision, leading to a considerable gain in readability of up to 94.57%. This shows that lexical simplification systems can be significantly improved by being adjusted to user fluency. It also suggests the incorporation of more customized CWI models to satisfy individual reader needs.

Research by Rodrigo Alarcon [19] and his team developed a lexical simplification approach that uses ensemble methods, including word embedding techniques, word length, and frequency, along with a Support Vector Machine (SVM) in Spanish. Using Multilingual CWI from Spanish Wikipedia, this study managed to achieve fourth place at the 2018 BEA Workshop with an F1 score of 74.97%. This research also suggests improving features by integrating the

Sense2Vec and Char2Vec models and considering the application of Deep Learning techniques to improve the simplification process.

Research by Harchit Mahajan, Prateek Koul, and Sukhjeet Singh [12] introduces a web extension that uses supervised learning and Bi-LSTM and BERT algorithms to make it easy to simplify text on web pages with one click. Their model, effective for single words, showed a very low loss of 0.03 after five epochs and achieved an F1 score of 0.68, but was unable to simplify phrases.

The 2021 research published in the IEEE journal by [21], Jipeng Qiang and his colleagues focuses on lexical simplification techniques for the Mandarin language, utilizing varied approaches such as synonym-based, word embedding, BERT, sememe, and hybrid methods. The results indicate that BERT-based and hybrid methods outperform others, with the hybrid method achieving a precision of 80% and an accuracy of 70%. This study also suggests enhancing the Mandarin lexical simplification system by integrating existing knowledge into the trained language model.

The research published by IEEE in 2022 by Salehah Omar and their team [3] developed a lexical simplification model for non-native Malay language speakers using Machine Learning methods, including Gradient Boosted Tree and Random Forest. This model, focusing on morphology and word stemming, achieved impressive results in terms of accuracy, precision, and F1-score, notably the Gradient Boosted Tree with an F1-score of 92.09%. This study underscores the importance of adapting the CWI model to the unique morphology of each language to enhance lexical simplification.

As indicated in the literature above, several methods and techniques have been utilized in lexical simplification, including unsupervised machine learning techniques such as BERT, customization in lexical simplification, ensemble methods, and unique morphology. Several languages have been discovered in the literature, such as English, Mandarin, and Malay. Standard metrics used for performance evaluation include accuracy, precision, recall, and F1 score.

3 Research Methodology

This study adopts a bibliometric analysis methodology, many have already carried out bibliometric analysis such as [23]–[25] to map and evaluate the future of Artificial Intelligence. This methodology includes the use of annual publication tabulation, the number of works and author citations, supported by VOS viewer software for visual analysis and the creation of collaboration maps and research trends. Bibliometric analysis transforms qualitative data into quantitative data, facilitating the identification of dominant research topics and quantities, as well as evaluating developments, structures, and new trends in the field.

The research process begins by defining the scope and sustainability of the study, leveraging data from the Scopus database. Methodological steps include the selection and filtering of Lexical Simplification research data, followed by the analysis of publication metadata. The collected data is extracted and exported in CSV and RIS formats, then processed using Microsoft Excel and analytical tools such as VOS Viewer and Harzing's Publish or Perish software for citation and collaboration analysis. The research structure encompasses introductory, literature review, methodology, results and discussion, as well as conclusions and limitations sections. Descriptive analysis will involve the assessment of document types, sources, languages, publication years, publication and citation frequencies, and the number of authors per document. Visualization of citation networks and collaborations will be used to interpret patterns and

relationships in Lexical Simplification research. Figure 1 depicts the flowchart of the search strategy.

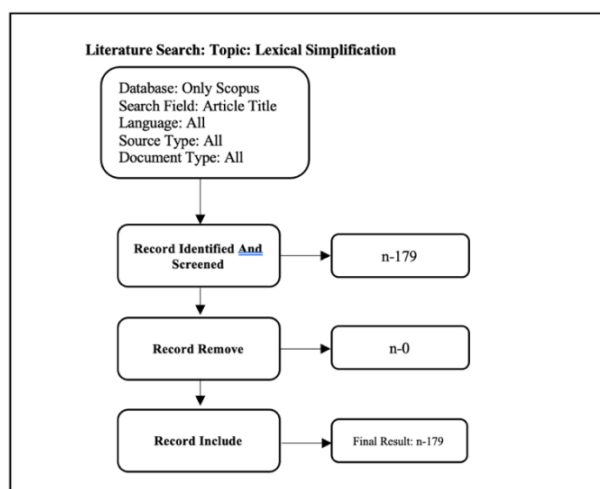


Figure 1. Flow Diagram of the Search Strategy

3.1 Bibliometric Analysis

In the context of lexical simplification, this chapter explores bibliometric analysis used to examine trends and progress based on statistical and mathematical methods in research. This tool allows researchers to measure and evaluate the quality and quantity of scientific publications, which is crucial for identifying current trends and finding gaps in the existing literature. Bibliometric analysis facilitates mapping of the current literature, allows assessment of research growth and describes social relationships and collaboration patterns between researchers, which can direct future research paths.

Using indicators such as publication type, authorship, author affiliation, and H index, this study analyzes datasets from selected databases to determine the impact and influence of publications, seen from the number of citations received.

The results of this analysis will offer broad insight into future prospects in the field of lexical simplification, as well as form the basis for further research aimed at filling the gaps identified in previous research.

4 Results and Discussion

This section discussed on the Trends in Document Profiles , Key researchers and institutions in the field of LS, Lead Institution in LS Field, Geographic Distribution of Research Journal or Major Publication at LS, Keywords and Key Topics in LS.

4.1 Trends in Document Profiles

From this research, data was obtained from Scopus from 2013 to October 2023. The results show trends based on analysis of documents and annual publications. As shown in Table 1, six documents were analyzed. It was found that the most publications were conference papers, with a total of 112 publications, accounting for around 65.6% of the total. This was followed by journal articles with 47 publications or 26.3%, and conference reviews with 13 publications or 7.3%. There are 3 review documents, 2 book chapters, while books and letters each amount to 1 publication or 0.6% of the total publications.

Table 1. Trends in Document Profiles

Document Type	Total Publication	Percentage (%)
Conference Paper	112	65.6 %
Article	47	26.3 %
Conference Review	13	7.3 %
Review	3	1.1 %
Book Chapter	2	1.7 %
Book	1	0.6 %
Letter	1	0.6 %
Total	179	100 %

4.1.1 Publication Trends by Year

The publication trend from year to year shows fluctuations, which indicates uncertainty. It was recorded that the highest number of publications occurred in 2016 and 2020, with 20 publications each. This was followed by 18 publications in 2018, and 16 publications in 2017.

Table 2. Publication Trends by Year

Year	Documents	Percentage (%)
2023	18	10.06 %
2022	20	11.17 %
2021	18	10.06 %
2020	20	11.17 %
2019	15	08.38 %
2018	18	10.06 %
2017	11	06.15 %
2016	20	11.17 %
2015	16	09.34 %
2014	9	05.03 %
2013	14	08.22 %
Total	179	100 %

Based on the analysis of this graph, it can be concluded that there is a need to increase the consistency and quantity of publications in the future so that this trend becomes more stable and increases.

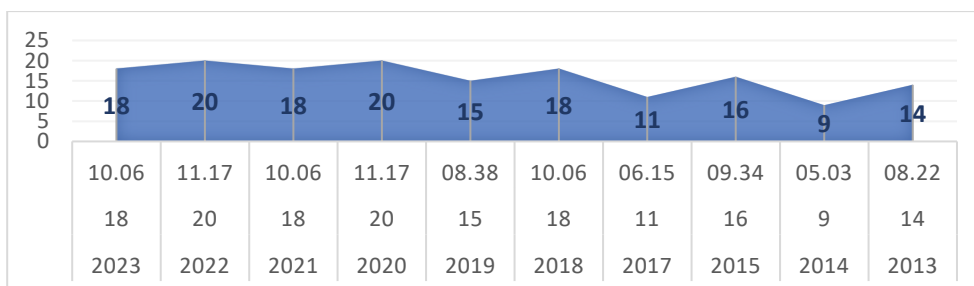


Figure 2. Total Publications and Citations by Year

4.2 Key Researchers and Institutions in the Field of LS

In this study, it was found that the most productive researchers in the field of lexical simplification were Saggion with a total of 13 publications. In second position, Paetzold, 12 publications Martínez 11 publications Moreno, Rello, Specia, produced 9 publications, followed by Alarcon, with 8 publications. This graph shows the distribution of research productivity in the field of lexical simplification among these researchers.

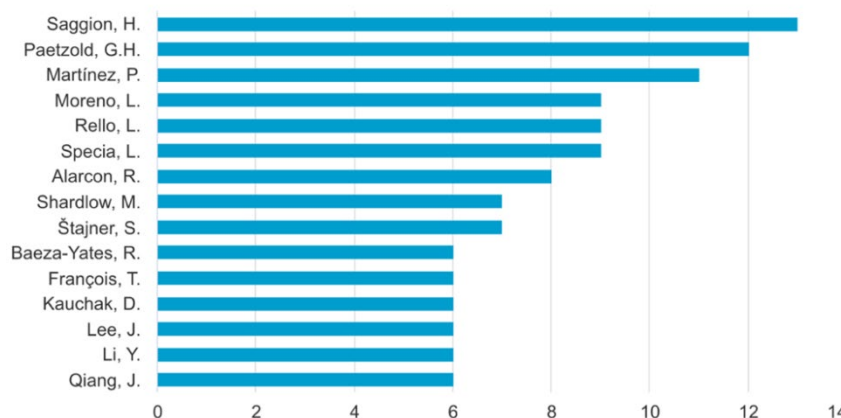


Figure 3. Researcher in the Field of LS

4.2.1 Citation Analysis by Documents

The most cited documents are the authors glavas, Štajner (2015), Paetzold, Specia, (2016) (2017) Shardlow(2013) Madella (2018)

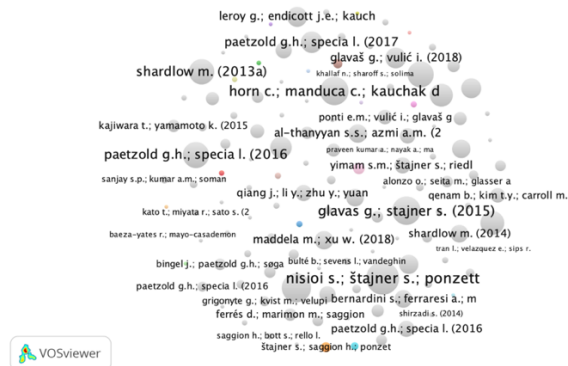


Figure 4. Network Visualisation Map of the Citation by Documents

4.2.2 Citation Analysis by Author

The authors most cited by Specia, Saggion, Kauchak, Stajner, and Glavas are the authors who have made significant contributions to this research

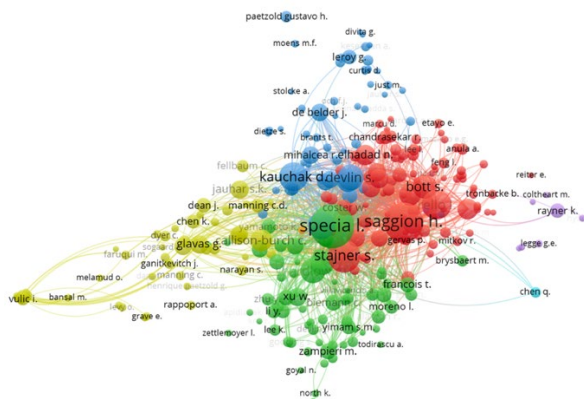


Figure 5. Network Visualisation Map of the Citation by Documents

4.3 Lead Institution in LS Field

The main institution making significant contributions in the field of lexical simplification is Universitat Pompeu Fabra in Barcelona, with 17 publications. This is followed by Universidad Carlos III de Madrid with 13 publications, University of Sheffield, with 12 publications. Universität Mannheim has produced 8 publications, while the University of Cambridge and Université Catholique de Louvain have recorded 7 publications. This data shows the important role of these institutions in the development of research in the field of lexical simplification.

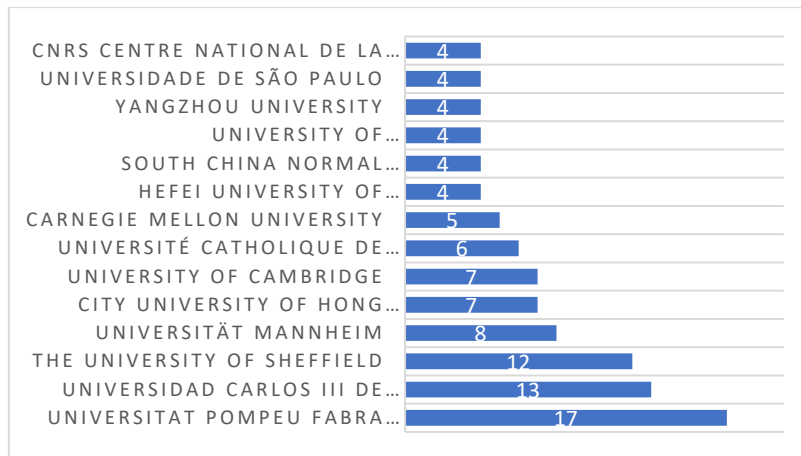


Figure 6. The Main Institution on LS

4.4 Geographic Distribution of Research

The majority of research on Lexical Simplification (LS) uses English, with 62 publications or 34.64%. In second place, Spanish contributed with 38 publications or 21.23%. Chinese takes third place with 17 publications, which accounts for 9.50%. German follows in fourth place with 15 publications or 8.38%. Meanwhile, Hindi from India comes in fifth place with 11 publications, or 6.15% of the total. This data shows the linguistic distribution in Lexical Simplification research globally.

Table 3. Language

Country/Territory	Documents	Percentage (%)
United Kingdom	62	34,64
Spain	38	21,23
China	17	9,50
Germany	15	8,38
India	11	6,15
France	8	4,47
Belgium	7	3,91
Brazil	7	3,91
Japan	7	3,91
Italy	4	2,23
Indonesia	1	0,56
Malaysia	1	0,56
Sweden	1	0,56

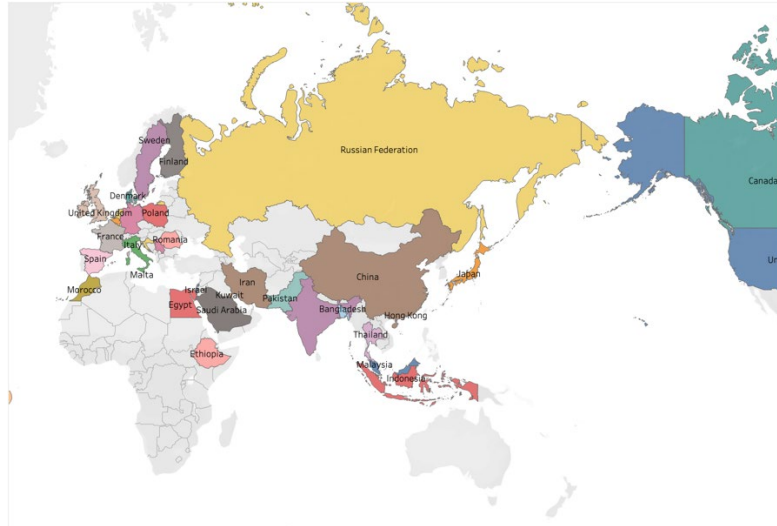


Figure 7. Country Research Mapping

4.5 Journal or Major Publication at LS

Figure 8 provides information on the main sources of publications in this field. "Lecture Notes in Computer Science" stands out as the top source with 15 publications. It is followed by "CEUR Workshop Proceedings", which accounts for 10 publications. "Procesamiento del Lenguaje" comes in third with 7 publications. Finally, "IEEE Access" which focuses on Natural Language Processing accounts for 4 publications. This data shows the important role of these publications in providing the latest knowledge and research in the field.

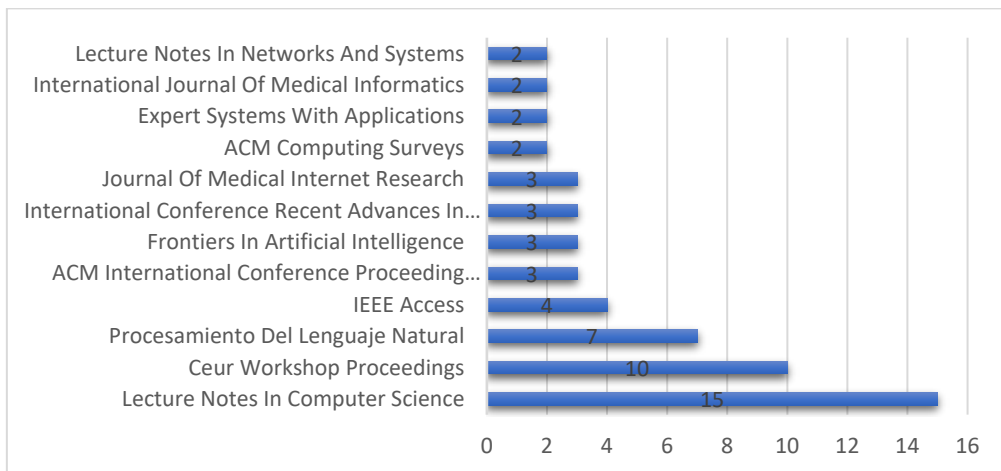


Figure 8. Main Publication on LS

4.6 Keywords and Key Topics in LS

We conducted a co-occurrence analysis to identify the main keywords in the Lexical Simplification (LS) topic. In this analysis, 'Lexical Simplification' appeared 75 times (18.34%), 'natural language processing systems' 64 times (15.65%), 'computational linguistics' 39 times (9.54%), 'semantics' 32 times (7.85%).

Table 6. Frequently Used Keywords

No	Keyword	Occurrences	Percentage (%)
1	Lexical simplification	75	18,34 %
2	natural language processing systems	64	15,65 %
3	computational linguistics	39	9,54 %
4	semantics	32	7,82 %
5	Complex Word Identification	28	6,85 %
6	text simplification	26	6,36 %
7	simple++	26	6,36 %
8	human	25	6,11 %
9	language model	13	3,18 %
10	linguistics	12	2,93
11	non-native speakers	12	2,93
12	syntactics	12	2,93
13	artificial intelligence	10	2,44
14	syntactic simplification	10	2,44
15	embeddings	9	2,20
16	word sense disambiguation	9	2,20
17	Wikipedia	7	1,71

4.6.1 Topics that have influence on LS

Based on co-occurrence analysis, topics related to Lexical Simplification (LS) can be grouped into 10 main categories. These categories include Natural Language Processing, approaches for non-native language learners, Machine Translation, Word Sense Disambiguation, Lexical Substitution, Dyslexia-related research, studies on Synonyms, Rule-Based methods, and applications in the English context. This analysis indicates that each of these categories has potential for further development in the future, suggesting that there are various aspects that can still be explored in LS research.

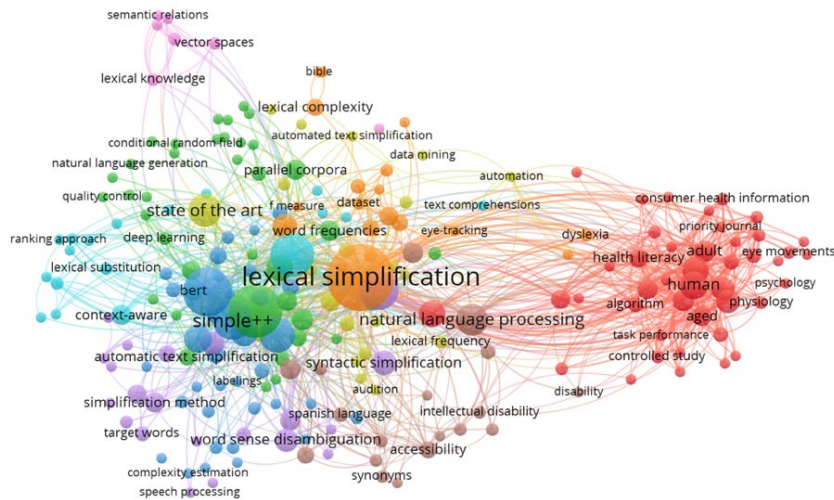


Figure 9. Network Visualisation Map of the Author's Keywords

Figure 10 illustrates that "lexical simplification" was central to the discussion, with close connections to concepts such as "natural language processing" (NLP), "simple++", "BERT", and "context-aware". Additional terms such as "consumer health information" and "dyslexia" indicated applications of NLP in easing access to health information and providing support for individuals facing dyslexia.

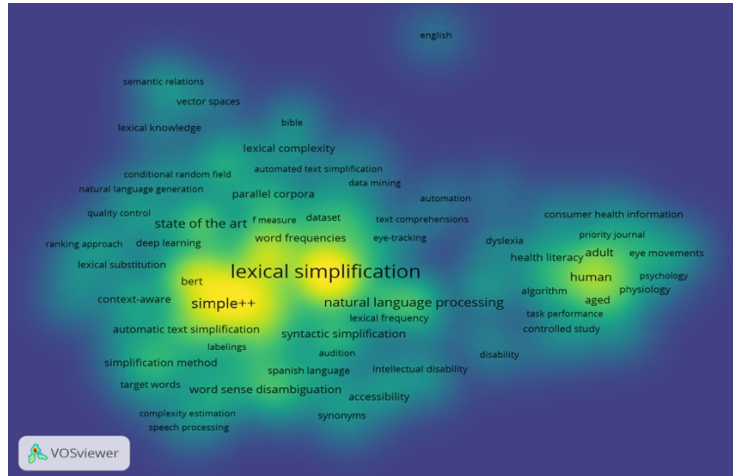


Figure 10. Density Visualisation of the Author's Keywords

4.7 Discussion

This section discussed on the Trends and Contributions in Lexical Simplification Research, Future Focus and Development Potential, Limitations of this study

4.7.1 Trends and Contributions in Lexical Simplification Research

The study, which took data from Scopus, showed an interesting trend in Lexical Simplification publications. This trend is dominant in conferences with 65.6% of publications, followed by journal articles and conference reviews. Significant fluctuations in annual publications, with peaks in 2016 and 2020, indicate imperfections that require increased consistency and quantity in the future. Meanwhile, Saggion and Paetzold stand out as prolific researchers, and institutions such as Universitat Pompeu Fabra and Universidad Carlos III de Madrid play an important role in this research. The global linguistic distribution shows the dominance of research in English, followed by Spanish, Mandarin, and others, signaling the diversity in Lexical Simplification studies.

4.7.2 Future Focus and Development Potential

The co-occurrence analysis identified 'Lexical Simplification' as the main focus, which includes areas such as Natural Language Processing, Complex Word Identification, Machine Translation, and Lexical Substitution. The journals "Lecture Notes in Computer Science" and "CEUR Workshop Proceedings" are recognized as contributing significantly to the knowledge in this area. The analysis results also indicate the potential for growth and innovation in LS, with ten major categories identified, including strategies for non-native language learners, Word Sense Disambiguation, Dyslexia research, synonym studies, Rule-Based methods, and applications in English.

4.7.3 Limitations of this Study

This study reviews Lexical Simplification (LS) trends based on Scopus data from 2013 to October 2023, with limitations such as reliance on a single data source and a focus on publications in a few specific languages. The review also highlights the importance of developing effective text simplification frameworks to assist individuals with disabilities and supporting LS research, to improve accessibility and understanding of information at large

5 Conclusion

Research on Lexical Simplification (LS), which relies on data from the Scopus database, has attracted significant attention. Over the past decade, there have been 179 publications in this area, but research remains limited in scope and depth. Most publications are conference papers, reflecting its popularity within the research community. Fluctuations in annual publications signal the need to improve consistency and increase the number of publications in the future. This study also recognizes the important contributions of several leading researchers and institutions, such as Saggion and Universitat Pompeu Fabra in Barcelona, who were key contributors.

One of the main weaknesses of this research is the lack of in-depth exploration of the current issues of LS in different countries, leading to an unseen phenomenon of gaps in individual LS issues. Despite its limitations, this research also highlights the importance of LS in the context

of text simplification, particularly to support individuals with disabilities. This marks an urgent need for further research in this area.

References

- [1] S. Devlin and G. Unthank, "Helping aphasic people process online information," 2006, <https://doi.org/10.1145/1168987.1169027>.
- [2] J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait, "Practical simplification of English Newspaper Text to Assist Aphasic Readers," *AAAI-98 Work. Integr. Artif. Intell. Assist. Technol.*, no. July 1998, pp. 7–10, 1998, [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.1145&rep=rep1&type=pdf>
- [3] N. Y. Salehah Omar, Juhaida Abu bakar, Maslida Mohd Nazir, Nor Hazlyna Harun, "Malay lexical simplification model for non-native speaker," *2022 International Conference on Intelligent Systems and Computer Vision, ISCV 2022*. 2022. <https://doi.org/10.1109/ISCV54655.2022.9806133>.
- [4] L. Rello, R. Baeza-Yates, S. Bott, and H. Saggion, "Simplify or help?: text simplification strategies for people with dyslexia," 2013, <https://doi.org/10.1145/2461121.2461126>.
- [5] R. Evans, C. Orasan, and I. Dornescu, "An evaluation of syntactic simplification rules for people with autism," 2014, <https://doi.org/10.3115/v1/W14-1215>.
- [6] E. Barbu, M. T. Mart, and L. A. Ure, "Open Book : a tool for helping ASD users ' semantic comprehension," *Proc. 2th Work. Nat. Lang. Process. Improv. Textual Access.*, no. June, pp. 11–19, 2013.
- [7] E. Sulem, "Simple and effective text simplification using semantic and neural methods," *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1. pp. 162–173, 2018. <https://doi.org/10.18653/v1/p18-1016>.
- [8] A. Garain, A. Basu, R. Dawn, and S. K. Naskar, "Sentence Simplification using Syntactic Parse trees," *2019 4th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2019*, pp. 672–676, 2019, <https://doi.org/10.1109/ISCON47742.2019.9036207>.
- [9] S. Seneviratne, E. Daskalaki, and H. Suominen, "{CILS} at {TSAR}-2022 Shared Task: Investigating the Applicability of Lexical Substitution Methods for Lexical Simplification," *Proc. Work. Text Simpl. Access. Readability*, pp. 207–212, 2022, [Online]. Available: <https://aclanthology.org/2022.tsar-1.21>
- [10] J. Lee and C. Y. Yeung, "Personalizing lexical simplification," *COLING 2018 - 27th Int. Conf. Comput. Linguist. Proc.*, no. 2010, pp. 224–232, 2018.
- [11] S. Štajner, D. Ferrés, M. Shardlow, K. North, M. Zampieri, and H. Saggion, "Lexical simplification benchmarks for English, Portuguese, and Spanish," *Front. Artif. Intell.*, vol. 5, pp. 1–32, 2022, <https://doi.org/10.3389/frai.2022.991242>.
- [12] H. Mahajan, "Lexical analyser web extension for text simplification," vol. 8, no. 4, pp. 24–30, 2022.
- [13] A. Dmitrieva, "A Multi-task Learning Approach to Text Simplification," *Communications in Computer and Information Science*, vol. 1357. pp. 78–89, 2021. https://doi.org/10.1007/978-3-030-71214-3_7.
- [14] G. Paetzold, "Reliable Lexical Simplification for Non-Native Speakers," 2015, <https://doi.org/10.3115/v1/N15-2002>.

- [15] J. Qiang, "Lexical simplification with pretrained encoders," *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*. pp. 8649–8656, 2020. <https://doi.org/10.1609/aaai.v34i05.6389>
- [16] P. Przybyła, "Multi-Word Lexical Simplification," *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*. pp. 1435–1446, 2020. [Online]. Available: https://api.elsevier.com/content/abstract/scopus_id/85137271176
- [17] N. Gala and J. Ziegler, "Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia," *Proc. Work. Comput. Linguist. Linguist. Complex.*, pp. 59–66, 2016, [Online]. Available: <https://www.aclweb.org/anthology/W16-4107>
- [18] R. Alarcon, "Lexical Simplification System to Improve Web Accessibility," *IEEE Access*, vol. 9, pp. 58755–58767, 2021, <https://doi.org/10.1109/ACCESS.2021.3072697>.
- [19] P. M. Rodrigo Alarcon, Lourdes Moreno, Isabel Segura-Bedmar, "Lexical simplification approach using easy-to-read resources," *Proces. del Leng. Nat.*, vol. 63, pp. 95–102, 2019, <https://doi.org/10.26342/2019-63-10>.
- [20] S. S. Al-Thanyyan and A. M. Azmi, "Simplification of Arabic text: A hybrid approach integrating machine translation and transformer-based lexical model," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 8, p. 101662, 2023, <https://doi.org/10.1016/j.jksuci.2023.101662>.
- [21] J. Qiang, X. Lu, Y. Li, Y. Yuan, and X. Wu, "Chinese Lexical Simplification," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1819–1828, 2021, <https://doi.org/10.1109/TASLP.2021.3078361>.
- [22] A. Nathan S Hartmann, Gustavo H Paetzold, Sandra.M, "A dataset for the Evaluation of Lexical Simplification in Portuguese for Children," no. 1, pp. 1–14, 2020, https://doi.org/10.1007/978-3-030-41505-1_6.
- [23] L. Zhang, J. Ling, and M. Lin, "Artificial intelligence in renewable energy: A comprehensive bibliometric analysis," *Energy Reports*, vol. 8, pp. 14072–14088, 2022, <https://doi.org/10.1016/j.egy.2022.10.347>.
- [24] F. Gao *et al.*, "Bibliometric analysis on tendency and topics of artificial intelligence over last decade," *Microsyst. Technol.*, vol. 27, no. 4, pp. 1545–1557, 2021, <https://doi.org/10.1007/s00542-019-04426-y>.
- [25] P. Song and X. Wang, "A bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years," *Asia Pacific Educ. Rev.*, vol. 21, no. 3, pp. 473–486, 2020, <https://doi.org/10.1007/s12564-020-09640-2>.