A CORPUS-BASED ANALYSIS OF TRENDS AND THEMES OF DIABETES A CASE STUDY OF AN ONLINE MALAYSIAN NEWSPAPER

Afendi Hamat[1], Azhar Jaludin[1], Haslina Rani[2], Ruhil Amal Azmuddin[3], Aznida Firzah Abdul Aziz[4], Tuti Ningseh Mohd-Dom[2*]

Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Bangi, Malaysia[1]
Faculty of Dentistry, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia[2]
Centre for Modern Languages, Universiti Malaysia Pahang, Pekan, Malaysia[3]
Faculty of Medicine, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia[4]

Corresponding author: 2*

Keywords:
diabetes education, corpora analysis, news frame, content analysis, internet

ABSTRACT

This paper presents the findings of an exploratory study that investigates the trends and themes related to the word 'diabetes' in the Malaysian media. The study utilizes the Malaysian Diabetes Corpus (MyDC), a specialized corpus consisting of online newspaper articles published between 2013 and 2018. The analysis combines corpus linguistics methodology with visualizations to examine the data. The results reveal a notable peak in the usage of the term 'diabetes' during the month of November across the studied years (2013-2018), coinciding with the increased attention given to World Diabetes Day. While there was a significant rise in the number of articles in 2018, the relative frequency of the word 'diabetes' remained stable throughout all the years. Among the various themes associated with 'diabetes', the most common ones include 'Awareness/Management', 'Type of (other related) disease', and 'Patient/Population'. This suggests that diabetes is increasingly portrayed in diverse contexts, with a greater emphasis on raising awareness and managing the condition rather than the traditional linguistic definition of the illness. The findings underscore the influential role of online media in promoting diabetes awareness and providing an effective platform for public diabetes prevention strategies. However, further investigation into the dynamics of online media and its impact on diabetes education is recommended to gain a deeper understanding of this relationship.

1. INTRODUCTION

The rising prevalence of diabetes mellitus (DM) has emerged as a significant public health concern, imposing burdens on individuals, national productivity, and healthcare systems. DM is a metabolic disease characterized by hyperglycemia resulting from either insufficient insulin secretion, inadequate insulin action, or a combination of both [1]. Moreover, diabetes is associated with a wide range of complications, including cardiovascular disease (CVD) [2], neuropathy [3], nephropathy [4], and retinopathy [5], thereby doubling the risk of mortality rates [6].

There are four main categories of diabetes: type 1 diabetes (T1DM), type 2 diabetes (T2DM), gestational diabetes (GDM), and specific etiology diabetes caused by monogenic factors, secondary drugs, and pancreatic factors [7]. T1DM and T2DM bear the greatest burden of the disease worldwide. In 2017, there were 425 million diagnosed cases of diabetes globally, a number projected to rise to 629 million by 2045 in the absence of adequate intervention measures [7]. Notably, the prevalence of diabetes is higher in low and middle-income countries (LMICs), with a prevalence rate of 13.5 percent compared to 10.4 percent in high-income countries [7].

Diabetes mellitus (DM) is a chronic condition that currently lacks a cure, but it can be effectively managed through the dissemination of knowledge regarding disease severity, manifestations, risk factors, complications, and management strategies. A key factor in this knowledge-sharing process is the role of mass media in health information dissemination. Mass media serves as a valuable tool for reaching a wide audience and providing organized information on mitigation plans, such as adopting a healthy diet, engaging in physical activity, and managing blood pressure [8]. By presenting information in various frames, mass media enables readers to gain different perspectives on health issues [9]. Analyzing diabetes-related topics in local newspapers serves as an ideal initial step in reaching target groups and informing the public about health policies associated with diabetes [10]. Language plays a crucial role in health promotion within the context of newspapers, making it a subject worthy of investigation. Corpus linguistics, a discipline that focuses on the large-scale analysis of language, is particularly relevant in this regard [11]. Utilizing corpora based on "language in use" rather than synthetic examples has gained popularity among scientists across various fields, including politics, social sciences, medicine, and education.

[12] conducted a corpus-based analysis focusing on the use of metaphors in discussions involving cancer patients and their caregivers. The study revealed that language plays a crucial role in improving information provision, diagnosis, support, self-management, and self-esteem, leading to the overall empowerment of both patients and healthcare professionals. Similarly, [13] employed a corpus-based approach to investigate the discourse surrounding eating disorders by examining content from two websites addressing anorexia. Their findings highlighted that this approach effectively captured the actual language used, aligning with the evidence-based practices emphasized by healthcare providers. In another study, [14] performed a corpus analysis of diabetes coverage in selected Australian newspapers from 2013 to 2017. Surprisingly, they discovered a mismatch in the coverage, with certain regional newspapers reporting less on diabetes despite higher cases of diabetes mellitus (DM) in those areas. This discrepancy raises concerns about the need for adequate representation and dissemination of diabetes-related information. Furthermore, it is worth noting that the continuous use of terms such as "diabetics" and "diabetes" in reference to individuals affected by DM should be carefully scrutinized to avoid potential misunderstandings among the public. Language in use should be examined and adjusted to ensure accurate and respectful communication.

Due to its multicultural makeup, Malaysia presents a diverse landscape for analyzing the framing of diabetes. In 2018, Malaysia recorded approximately 3.6 million adults with diabetes, representing a prevalence rate of 16.8 percent [15]. Alarmingly, the Malaysian National Health and Morbidity Survey in 2019 revealed that around 10 percent of the population remained unaware of their diabetes status [16]. This rise in diabetes cases necessitates collective efforts to enhance public awareness of diabetes prevention through the use of literacy-appropriate reading materials across various media platforms. Given the severity of the situation, it is crucial to initiate a corpus analysis of diabetes health promotion and information dissemination in Malaysia. This analysis would provide valuable insights into the current linguistic trends present in online media. To accomplish this, the current research combines the tools and techniques of corpus linguistics with data visualizations to investigate the representation of 'diabetes' within the Malaysian Diabetes Corpus (MyDC).

2. METHODOLOGY

2.1 Design

The present study adopted a corpus-driven mix approach and utilised the quantitative and qualitative methods. [17] described the corpus-driven technique as an inductive process where corpora are investigated from the bottom up and identified patterns explain linguistic regularities and exceptions of the language variety or genre exemplified by the corpora. Data from 2013 to 2018 were retrieved from MyDC and the outcomes were evaluated using past published theme-based bibliometric studies.

2.2 Data

The MyDC contained 212 articles from six sub-corpora (MyDC_2013 – MyDC_2018) with 10,904 types (the number of unique word forms) and 134,024 tokens (the number of individual words in the text). The corpus only included articles from one of the top Malaysian online English newspapers, *TheStar*, from 2013 to 2018. The term used for searching and building up the corpus was diabetes*.

2.3 Procedure

Stage 1: The raw frequency count for ‘diabetes’ in each sub-corpus was determined using the Word List function in AntConc to identify the diabetes trend over six years in the newspaper publication. Subsequently, the count was converted to relative frequency through a normalisation process for an accurate corpus or sub-corpus comparison in varied sizes. The normalisation process was performed using the following formula:

$$Relative\ frequency\ of\ word\ N\ in\ Corpus_i = \frac{Raw\ frequency\ of\ word\ N\ in\ Corpus_i}{Total\ token\ in\ Corpus_i} * Common\ base$$

The base used for the study was 1,000 and a line graph was plotted to indicate the occurrence trend of ‘diabetes’ from My_DC_2013 to My_DC2018.

Stage 2: The steps below were used to identify the themes associated with ‘diabetes’ within the corpus.

A. The MyDC was the first Part-Of-Speech (POS) tagged using TagAnt (https://www.laurenceanthony.net/software/tagant/) (Access date: 20th January 2022). The stop words were subsequently removed from the corpus.

B. Collocation extraction was performed for ‘diabetes’, which is a sequence of words or phrases that co-occur with another word (known as node word) more often than would be expected by chance [18]. Although collocation demonstrates linguistic significance (as in the simple example of phrasal verbs), it also reflects the wide range of human sociocultural constructs [19]. The word ‘candidates’ for diabetes was generated using the collocates function in AntConc. Three criteria were used to select the collocations:

i. The study employed Mutual Information (MI) score as the statistical measure between the node word with the collocate candidate, which was set at 3.0 and above (> = 3.0).

ii. The frequency of candidate occurrence (the normalised cut-off threshold) was set at 20 per million words [20].

iii. The distribution of the node word and collocations in the corpus was set at 10 percent of the texts.

C. The thematic analysis involved identifying the themes associated with ‘diabetes’ within the corpus. [21] emphasised two methods for determining the text themes: from the data (an inductive approach) and the investigator’s prior theoretical understanding of the phenomenon under study (prior or deductive approaches). The study adopted the second method based on the following procedures:

i. FrameNet is a linguistic knowledge graph (https://framenet.icsi.berkeley.edu/fndrupal/) to generate the preliminary topics or themes of 'diabetes'.

ii. The noun collocations for 'diabetes' were assigned manually to the relevant themes based on the dictionary definition of specific terms and the context in the corpus. In cases of uncertainty, the context of each noun was reviewed through concordance analysis and evaluated by the researchers.

iii. Any collocations outside of the themes produced by FrameNet were placed under the miscellaneous category.

iv. The last step in the procedure was to review, revise, and relabel the themes based on concordance review and common dictionary meanings.

3. RESULTS

3.1 The trend for the occurrence of 'diabetes'

A quantitative method was applied for the initial trend analysis of 'diabetes' in MyDC to clearly depict the patterns within the corpus. Figure 1 illustrates the number of article increments in November for the selected years in the corpus. Although the article pattern for 2013 was irregular, high article numbers were observed in November compared to other years. The article pattern numbers fluctuated for all years with no observed pattern identified from the corpus content analysis. Further analysis of cumulative frequencies determined the trend of the occurrence of 'diabetes' during the relevant periods. In Figure 2, November displayed the highest cumulative frequency of 86.3 percent followed by March (75.9 percent), while the least cumulative frequency was identified in Feb with 5.7 percent.

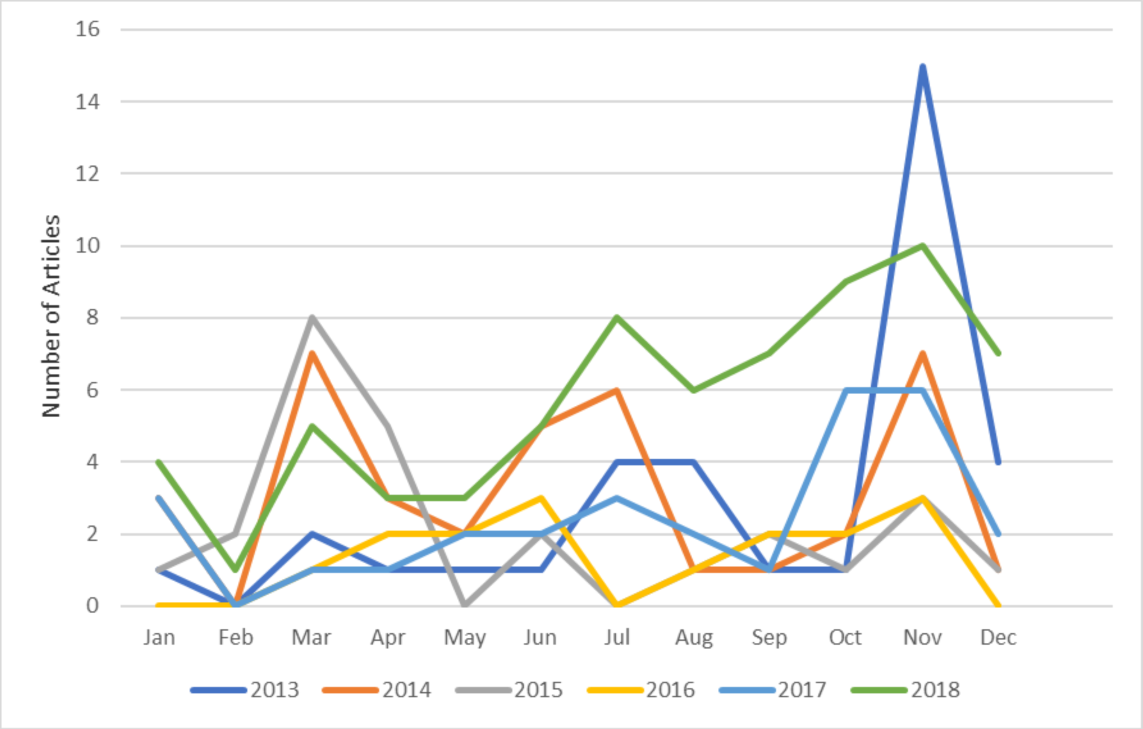

Figure 1. The occurrence trend of 'diabetes' in newspaper articles from January to December of 2013 to 2018

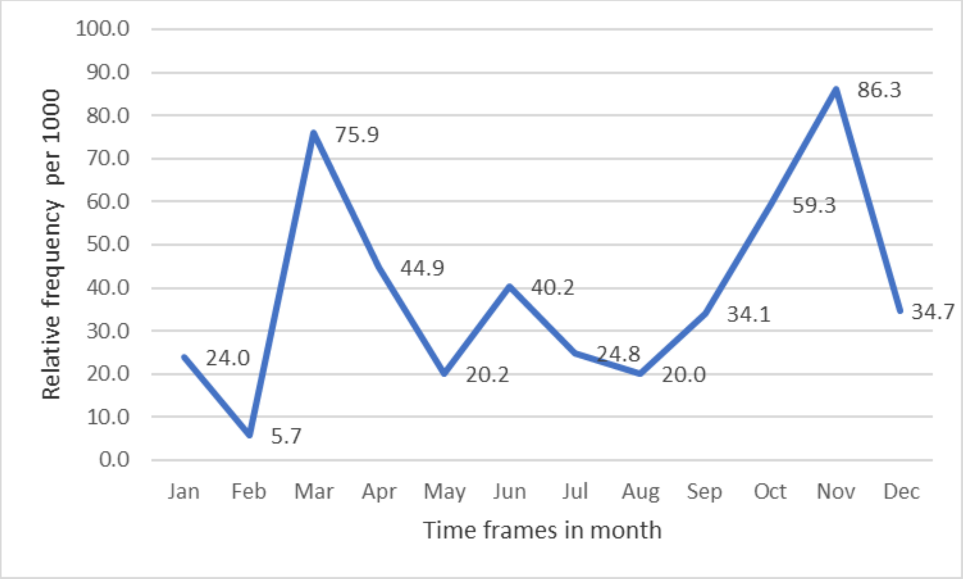

Figure 2. The cumulative frequency trend of ‘diabetes’ from January to December 2013 to 2018

Although the 2018 corpus contained the highest number of articles with ‘diabetes’, a different pattern displayed cumulative relative frequency (see Figure 3). A low cumulative relative frequency was observed in 2018 similar to 2017. The highest cumulative relative frequency was in 2015 with 13.0 followed by 10.9 in 2013. Figure 4 illustrates the density versus articles for each investigated year where the highest number was in 2015 followed by 2013. The years 2014, 2017, and 2018 recorded a close resemblance to the pattern. In Figure 4, the density for 2018 was the lowest and less than half of the highest year, 2015.

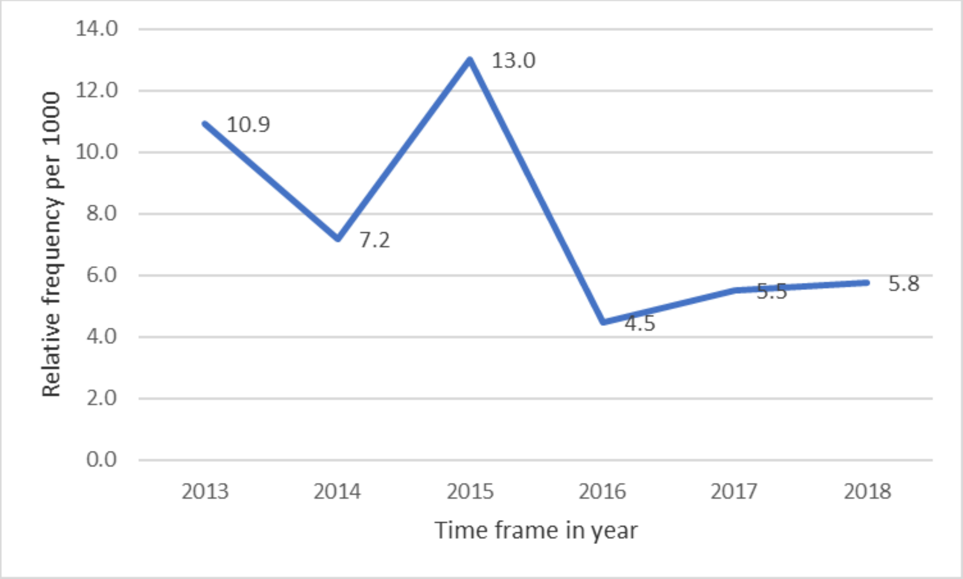

Figure 3. The relative cumulative frequency trend of ‘diabetes’ in newspapers published from 2013 to 2018

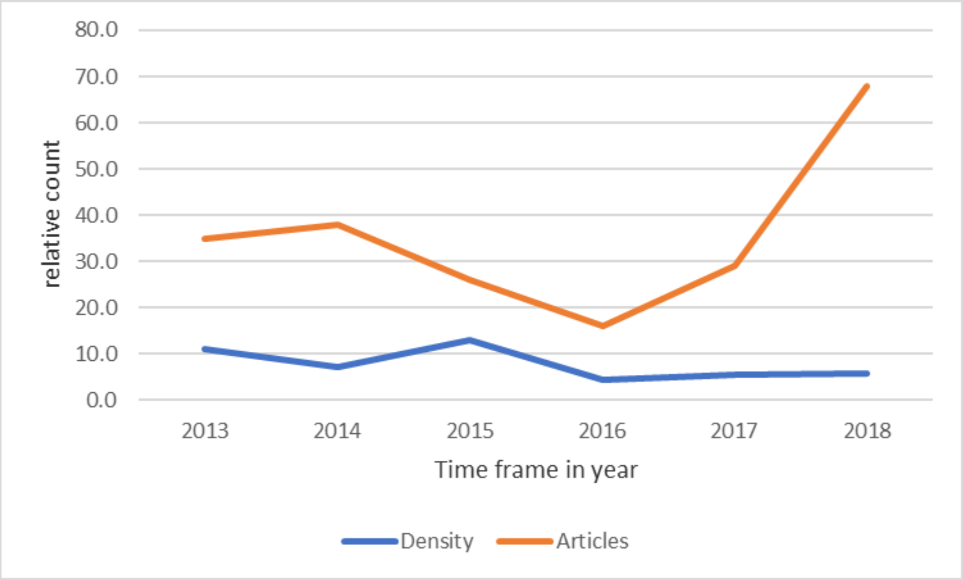

Figure 4. Density versus the number of articles from 2013 to 2018

3.2 Collocation analysis and theme identification

The corpus-based analysis of collocations identified 108 noun collocations of ‘diabetes’, which fit Stage 2 criteria in the procedure above. Figure 5 presents nine themes from diabetes news within the corpus. Overall, the ‘awareness/management’ theme contained the largest number of noun collocations, while ‘body part (affected)’ registered the least number of collocations. The largest number of collocations was ‘awareness/management’ with 60 (55 percent) collocations, while the others include ‘body part (affected)’ with 1 (1 percent) collocation, ‘symptom of the condition’ with 2 (2 percent) collocations, ‘patient or population’ with 9 (8 percent) collocations, ‘name to identify the condition’ with 2 (2 percent) collocations, ‘type of (other related) disease’ with 19 (18 percent) collocations, ‘people’ with 4 (4 percent) collocations, ‘location’ with 4 (4 percent) collocations, and ‘association (special interest group)’ with 7 (6 percent) collocations. Different theme categories from the corpus are presented in the Supplementary file section.

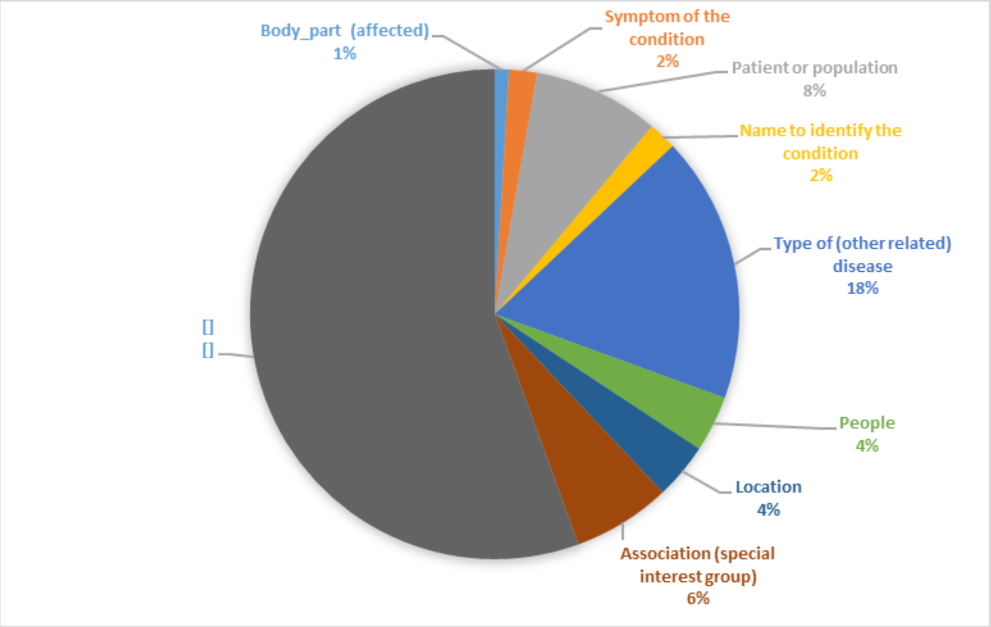

Figure 5. The nine themes of noun collocations retrieved from diabetes news in MyDC Corpus

4. DISCUSSION

Public education on diabetes prevention is considered the optimal strategy for addressing the concerns surrounding individuals, communities, and healthcare systems. In this study, a corpus analysis was conducted on the Malaysian Diabetes Corpus (MyDC) to identify trends in the usage of the term 'diabetes' and related themes within English newspaper articles in Malaysia. This pioneering research in Malaysia explores diabetes-related issues in local online newspapers using corpus analysis.

The analysis of the corpus revealed an upward trend in the usage of the term 'diabetes' during the month of November across all six years (see Figures 1 and 2). This observation aligns with the fact that World Diabetes Day falls on the 14th of November, prompting newspapers to place greater emphasis on publishing articles related to the disease. The findings from the corpus content analysis are consistent with the objective of World Diabetes Day, which was established by the United Nations in 2006 to promote international awareness of diabetes-related topics. Notably, there was a significant increase in the number of articles in 2018 compared to the preceding three years, indicating a growing attention to diabetes-related matters within the mass communication media. Although the number of articles increased over the years, the overall density remained relatively stable (see Figures 3 and 4).

The involvement of newspapers in regional health campaigns cannot be overlooked, as the power of repetition through advertising plays a significant role. Previous research by [23] has shown that repeated exposure to terms and term combinations can lead to significant changes in behavior. Consequently, the steady increase in the number of diabetes-related articles over the years reflects the influence of this repetition. The usage of the term 'diabetes' in different contexts has evolved from a narrow focus on illness-related contexts to a more comprehensive representation, as demonstrated in Figure 5, where the 'awareness/management' category exhibited a higher number of themes compared to others.

An important discovery from the data is the appearance of the term 'prediabetes' only in the 2013 sub-corpus. This finding highlights the significance of promoting awareness of prediabetes, as it can potentially prevent the progression of the disease [24]. Therefore, the term 'prediabetes' should consistently be included in public discourse. Further investigation using different languages in local news sources is warranted to determine if the observed patterns hold true across different linguistic contexts. Additionally, to complement this study, which provides a sample of educational materials available in the public domain, it would be valuable to conduct studies that examine the uptake and comprehensibility of publicly available diabetes materials through online reading or media engagement and public health digital publications. Currently, there is one local study that analyzed the readability of such materials [25], but it would be beneficial to determine if people are engaging with and understanding these materials.

This study has several limitations. First, relying solely on quantitative data may not provide sufficient depth to explore the effectiveness of local corpus materials in raising health awareness among the public. Incorporating additional metrics to gauge reader engagement with online materials could contribute to a more comprehensive qualitative analysis of the corpus. Furthermore, since the study was initiated in 2019, the analyzed articles only reflect reporting trends on diabetes up until 2018 and may not capture current trends.

From a language perspective, health promotion in Malaysia faces challenges due to its multilingual and ethnically diverse population. While Malay is the official language and English is considered a second language, various vernacular languages are used by different ethnic groups, including numerous Chinese dialects, Tamil, and diverse Borneo languages and dialects. Therefore, future research should explore the development of sub-corpora to encompass languages beyond English.

5. CONCLUSION

The frequency of the term 'diabetes' in the sampled articles remained consistent across all years, indicating a stable relative frequency. However, a noticeable upward trend in its usage was observed specifically during World Diabetes Day, which falls on the 14th of November each year. This finding highlights the increased emphasis placed on diabetes-related discussions during this significant day. Among the various themes associated with 'diabetes', the most common ones were 'awareness/management', 'type of (other related) disease', and 'patient/population'. These themes hold great relevance and appropriateness in the context of national strategies aimed at mitigating the rise of Malaysians living with or at risk of diabetes. By incorporating these themes into national initiatives, efforts can be made to address the challenges posed by diabetes effectively.

Availability of data and materials

The data that support the findings of this study are openly available in GitHub at https://doi.org/10.5281/zenodo.7033889,

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Higher Education, Malaysia under the Malaysian Research University Network Grant (MRUN-Rakan-RU-2019–02/1). The funding body does not provide any input in the analysis and interpretation of data, and in writing the manuscript.

COMPETING INTERESTS

The authors report no conflict of interest in the project.

6. REFERENCES

[1] Verhulst MJ, Loos BG, Gerdes VE, et al. Evaluating all potential oral complications of diabetes mellitus. Front. Endocrino 2019; 10: 56. https://doi.org/10.3389/fendo.2019.00056

[2] De Rosa S, Arcidiacono B, Chiefari E, et al. Type 2 diabetes mellitus and cardiovascular disease: genetic and epigenetic links. Front. Endocrino. 2018;9:2. https://doi.org/10.3389/fendo.2018.00002

[3] Kiyani M, Yang Z, Charalambous LT, et al. Painful diabetic peripheral neuropathy: health care costs and complications from 2010 to 2015. Neurology 2020;10(1):47-57. https://doi.org/10.1212/CPJ.0000000000000671

[4] Koch EA, Nakhoul R, Nakhoul F, et al. Autophagy in diabetic nephropathy: a review. Int Urol Nephrol. 2020;52(9):1705-12. https://doi.org/10.1007/s11255-020-02545-4

[5] Khazai B, Adabifirouzjaei F, Guo M, et al. Relation between Retinopathy and Progression of Coronary Artery Calcium in Individuals with Versus Without Diabetes Mellitus (From the Multi–Ethnic Study of Atherosclerosis). Am J Cardiol. 2021;149:1-8. https://doi.org/10.1016/j.amjcard.2021.03.026

[6] Lin X, Xu Y, Pan X, et al. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. Sci Rep. 2020;10(1):1-1. https://doi.org/10.1038/s41598-020-71908-9

[7] Forouhi NG, Wareham NJ. Epidemiology of diabetes. Medicine. 2019;47(1):22-7. https://doi.org/10.1016/j.mpmed.2014.09.007

[8] Timpel P, Harst L, Reifegerste D, et al. What should governments be doing to prevent diabetes throughout the life course?. Diabetologia. 2019;62(10):1842-53. https://doi.org/10.1007/s00125-019-4941-y

[9] Ohlsson R. Public discourse on mental health and psychiatry: Representations in Swedish newspapers. Health. 2018;22(3):298-314. https://doi.org/10.1177/1363459317693405

[10] Gounder F, Ameer R. Defining diabetes and assigning responsibility: how print media frame diabetes in New Zealand. J Appl Commun Res. 2018;46(1):93-112. https://doi.org/10.1080/00909882.2017.1409907

[11] Bowker L. Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research. Library Hi Tech. 2018; 36(2): 358-371. https://doi.org/10.1108/LHT-12-2017-0271

[12] Semino E, Demjén Z, Hardie A, et al. Metaphor, cancer and the end of life: A corpus-based study. In Metaphor, Cancer and the End of Life: A Corpus-Based Study (pp. 1–306). Routledge; 2017.

[13] Hydén LC. Storytelling in dementia: collaboration and commun ground. Living with demencia. London: Palgrave. 2017; 23:116-34.

[14] Bednarek M, Carr G. Diabetes coverage in Australian newspapers (2013-2017): A computer-based linguistic analysis. Health Promot J Austr. 2020;31(3):497-503. https://doi.org/10.1002/hpja.295

[15] Ganasegeran K, Hor CP, Jamil MF, et al. A systematic review of the economic burden of type 2 diabetes in Malaysia. Int J Environ Res Public Health. 2020;17(16):5723. https://doi.org/10.3390/ijerph17165723

[16] Institute for Public Health 2020. National Health and Morbidity Survey (NHMS) 2019: Non-communicable diseases, healthcare demand, and health literacy—Key Findings. Ministry of Health Malaysia. 2020. https://iptk.moh.gov.my/images/technical_report/2020/4_Infographic_Booklet_NHMS_2019_-_English.pdf [accessed 21st September 2022).

[17] McEnery T, Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press; 2011.

[18] Shammas NA. Collocation in English: Comprehension and use by MA students at Arab universities. Int J Humanit Soc Sci. 2013;3(9):107-22.

[19] Bosque I. On the conceptual bases of collocations: restricted adverbs and lexical selection. In Collocations and other lexical combinations in Spanish: theoretical, lexicographical and Applied perspectives 2016 (pp. 9-20). Routledge.

[20] Biber D, Conrad S, Cortes V. If you look at…: Lexical bundles in university teaching and textbooks. Applied Linguistics. 2004;25(3):371-405.

[21] Ryan GW, Bernard HR. Techniques to identify themes. Field Methods. 2003 ;15(1):85-109.

[22] Pandey SK, Sharma V. World diabetes day 2018: battling the emerging epidemic of diabetic

retinopathy. Indian J Ophthalmol. 2018;66(11):1652. https://doi.org/10.4103/ijo.IJO_1681_18

[23] Zhan L, Guo D, Chen G, Yang J. Effects of repetition learning on associative recognition over time: Role of the hippocampus and prefrontal cortex. Front Hum Neurosci. 2018;12:277. https://doi.org/10.3389/fnhum.2018.00277

[24] Okosun IS, Lyn R. Prediabetes awareness, healthcare provider's advice, and lifestyle changes in American adults. Int J Diabetes Mellit. 2015 ;3(1):11-8. https://doi.org/10.1016/j.ijdm.2010.12.001

[25] Hamat A, Jaludin A, Mohd-Dom TN, Rani H, Jamil NA, Abdul Aziz AF. Diabetes in the News: Readability Analysis of Malaysian Diabetes Corpus. Int J Environ Res Public Health. 2022;19(11):6802. https://doi.org/10.3390/ijerph19116802