DBTechVoc: A POS-tagged Vocabulary of Tokens and Lemmata of the Database Technical Domain

Jatinderkumar R. Saini[1]*, Ketan Kotecha[2], Hema Gaikwad[3]
Symbiosis Institute of Computer Studies and Research, Symbiosis International Deemed University, Pune, India[1, 3]
Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International Deemed University, Pune, India[2]

Abstract*—Vocabulary of a language has a great role to play in the Natural Language Processing (NLP) applications. Such applications make use of lists like stop-word list, general service list, academic word list and technical domain word list. The technical domain word list differs with each domain and though it is available for fields like medicine, biology, computer science, physics and law, the domain of databases in specific has still not been explored. For the first time, we propose technical vocabulary comprising of POS-tagged unigram tokens and POS-tagged unigram lemmata for the technical domain of databases. This vocabulary has been called DBTechVoc with a coined term. Notably, the multi-word phrases have also been considered, without their further tokenization, to maintain their semantics. The empirical results, with more than 1000 high quality research papers collected over a period of 45 years from 1976 to 2021, prove that the technical general word list of the domain of computer science is different from the technical and specific word list of the domain of databases. The overlap was found to be less than 2%. The research titles use 6% Rainbow stop words while 13% of the words used for the research paper titles are inflectional forms of lemmata.

Keywords*—*Database; lemma; part-of-speech (POS); technical word list; token; unigram; vocabulary

I. Introduction

It has been empirically proved by Liu and Nation [1] that in order to comprehend a piece of text, at least 95% of the words should be recognized by the reader. In fact, this concept could be applied equally well to the listeners of a natural language too. Any person’s knowledge of a language is just limited by the knowledge of the vocabulary of that language. It is needless to mention that though grammar of a language has an important role to play too, it is the number of words known to a reader or listener that contributes to the comprehension of the semantics of a language. There are a number of specific terms like tokens, lemmata and stop words, just to name a few, which are used by the linguists, computational linguists as well as those working in the area of Natural Language Processing (NLP).

The importance of title of a research paper cannot be undermined. Several research works like those of Dewan and Gupta [2], Tullu [3], Mack [4] and Karagel and Karagel [5] have advocated and elaborated the importance of title of the research paper as a gist of the paper contents. Soler [6] conducted an exploratory study dedicated to the titles of scientific research papers. Hengl and Gould [7] emphatically highlighted that the title of the research papers should tend to clearly indicate the main contents of the research paper in addition to the actual discoveries discussed in the paper.

Any text used to convey the necessary semantics consists of words. The process used for separating the individual units of this text is called Tokenization and the units so received are called Tokens [8]. These tokens may in turn be formed of single word, two words, etc. which are technically referred to as unigrams, bigrams, etc. respectively. Unlike Kyle [9], we have not considered the relevance of single, double, etc. unigrams. Also, the consideration of the unigrams in the present research work is with respect to the number of words in a sentence rather than the number of letters in a word. More specifically, for a period of 45 years from 1976 to 2021, we have considered 1031 titles of the database domain related research papers as the sentences and extracted unigrams as well as lemmata from these titles. The process of lemmatization is deployed to find the base morphological form of a word [10]. This form is called ‘lemma’ if it is singular, and ‘lemmas’ or ‘lemmata’ if it is plural. Similarly, the Part-Of-Speech (POS) tagging is done in order to group the various tokens into different categories as well as to provide more information on the role of such tokens when used as words in a sentence [11]. The present research work makes use of Lemmatizer [12] provided by the Stanford University and the POS-tagger [13] provided by the University of Copenhagen. The POS-tagger [14] provided by the Princeton University has also been used.

Smith [15] has discussed three main types of lists, viz. Academic, General and Technical. He defines the Academic Vocabulary is the list containing words which could be used for discourse in the academic world including the usage during conferences. He defines the General Vocabulary as consisting of the most frequently used words for a language. Similarly, he advocates that the Technical Vocabulary consists of the discipline-specific words. In wake of this context, the present research work deals with proposing the lists which fit in the category of Academic Vocabulary and Technical Vocabulary. It does not fit in the definition of General Vocabulary as we have not considered the words based on their frequency. We have presented the token list as well as the lemmata list, both of which are POS-tagged, towards the academic and technical categories of words for the technical database domain.

Rest of the paper is structured as follows: Section 2 presents the pertinent literature review. Section 3 elaborates the methodology, followed by section 4 presenting the results and discussion. The paper ends with the last Section 5 on conclusion and limitations of the present research work. Many

*Corresponding Author.

application areas and directions of future work are also presented in the last section.

II. Literature Review

Ever since the researchers recognized the importance of the vocabulary of a language, they have been working for the generation and in the field of word-lists. The research has gained more interest in the wake of several developments including the growth of various interdisciplinary fields like Computational Linguistics (CL), Natural Language Processing (NLP) and Foreign Language Understanding (FLU), among others. First of its kind, a general service list, comprising of most commonly used English words was proposed way back in 1953 by West [16]. This list has seen two updates, viz. new-GSL by Brezina and Gablasova [17] and NGSL by Browne et al. [18], in recent times with inclusion of additional words owing to consideration of bigger corpora respectively for more than 12 billion words and 2 billion words respectively. It is important to note that where GSL by West [16] contained some 2000 words, NGSL by Browne et al. [18] contained 2818 lemma words. Gilner and Morales [19] used the existing GSL and presented a speech-based analysis of the words in the list. Gilner [20] also presented an introductory note on the description of GSL with an aim to aid and ease its comprehension. Nation and Waring [21] argued that the most part of the text is actually composed of only a few words which occur frequently in the text.

The stop-word lists of the various natural languages contribute to the creation of a language itself. Though such stop words or noise words are believed to statistically irrelevant and mostly useless from point of view of NLP applications too, they enable the spoken use and representation of the vocabulary of a language through speech, dialects and scripts. They help in putting the vocabulary words together to make sense to the listener and reader. This way they contribute to a special vocabulary domain in its own right. Researchers have presented various types of stop-word lists as well as those for several languages. Researchers have also worked a lot on the analysis and classification of stop-word lists for various languages. Fayaza and Farhath [36] presented a stop word list for Tamil language. Similarly, Kaur and Saini [22, 23] worked for the stop-word list of Punjabi language, Rakholia and Saini [24, 25] worked for the stop-word list of Gujarati language while Raulji and Saini [31, 32] worked for the stop-word list of Sanskrit language. A stop word list based on the Rainbow statistical text has also been presented by Shuson [38].

Similar to the concept of GSL and NGSL, Coxhead [27] proposed the concept of an Academic Word List (AWL). However, Hancioglu et al. [26] argued that it is inappropriate to treat AWL and GSL as separate lists. Billurogˇlu and Neufeld [28] used a rather simplistic approach of list generation by filtering out the unique common words from the corpus created by the merging of all existing and commonly used lists. The concept of various lists has gained importance and interest as the various words contained in such lists provide a glimpse into the vocabulary of a language or a specific domain thereof. It is the knowledge of this vocabulary and understanding of words which helps one to understand and learn a language.

In addition to the core research works on various types of lists, researchers have also explored other similar and pertinent domains. For instance, a number of methods exist for the extraction of the terms from the scripted version of the language. The extracted tokens are in turn used by various downstream operations in the field of CL and NLP. Bakaric et al. [29] evaluated many such methods for the German language. Choy [30] proposed an innovative method for generation of stop word list by making use of combinatorial values. Venugopal et al. [33] presented lemmata for the Hindi corpus stop words while Saini and Rakholia [34] presented a detailed statistical analysis for such lists for various international languages.

Saed et al. [39] presented a lemmata list of the various categories related to biological and medical sciences including for the classes of diseases and the recent COVID-19 outbreak. Das et al. [40] used various sources and presented the technique of generating a list of words for the specific domain of Finance. They presented a typical comparison and contrast of their lexicographic approach with the conventional machine learning based approaches. Ahsanuddin et al. [41] attempted to create a list of words for the vocabulary learning by the students aiming to learn languages like Indonesian, English, German and Arabic. They used nearly 380 Thousand tokens for the corpus creation. Joensuu [42] presented an innovative description of the lists of menus and recipes for the culinary domain. The language researched by the author was Finnish.

Using the lists like AWL by Coxhead [27], Wingrove [43] attempted to analyze the introduction of TED talks for English learners. The list of words extracted from the talks and other lectures was analyzed for the possibility of vocabulary enrichment of the language learners. On the sidelines, he also analyzed the richness of such talks from the perspective of the usage of different lexicons. Alasmary [44] presented a technical list of words for the domain of mathematics. He sourced the corpus from the textbooks of the mathematics course at the graduate-level of students.

Though a list of the database terms is provided by raima.com [35], it consists of only a limited 150 terms, without POS and more in the form of a dictionary. Also it has not made use of lemmatization to present the lemmata list. The present research work considers all such points by providing an improved set of lists. Smith [15] has presented a few subject-specific lists like for Medicine, Law, Computer Science, Physics, Chemistry and Accounting but he has not presented a specific technical word list for the subject of Database which happens to be a sub-field under the umbrella of Computer Science. Also, the Computer Science subject related word list provided by him is very different from the vocabulary used in the Database domain.

After a thorough literature review, it was concluded that though several types of lists like stop-word lists of different types and for different languages, general service lists of various types and many academic word lists exist, the area of technical domain word lists is rather unexplored. This is particularly true for the highly technical domains like that of databases. Additionally, as no such list exists for the specific field of databases, there is no research work which has elaborately annotated such a list with Parts-of-Speech (POS). In order to bridge this gap, this research work presents a technical domain word list for the domain of databases. It is remarkable that as the field of databases itself is a sub-set of the field of computers, the proposed lists could also be used with backward inclusion in the technical domain list of the parent field of computers in general. Hence, the contributions of the present research work are manifold in terms of presentation of vocabulary and word lists. In the increasing order of generality, firstly, it presents a list of vocabulary words for databases, secondly it presents a technical word list and finally it also presents the vocabulary word list for the field of computer science and engineering as well as information technology.

III. Methodology

All the executions of the multiple codes needed at different junctures of the present research work were done using the open source Java programming language with version 17.0.1 2021-10-19 LTS for Java Development Kit (JDK), build 17.0.1+12-LTS-39 for the Standard Edition (SE) Runtime Environment and build 17.0.1+12-LTS-39 with mixed mode and sharing features for Java HotSpot(TM) 64-bit server Virtual Machine (VM). The execution was done on a machine with Intel(R) Core(TM) i3-8145U CPU with 2.10 GHz, 8 GB RAM and a licensed Windows 10 Pro 64-bit operating system.

In order to create a subject-specific vocabulary of the technical domain of databases, two Part-Of-Speech (POS) tagged lists were created. The diagrammatic representation of the process is depicted in Fig. 1. As a first step, the list of titles of research papers published in the field of databases from 1976 to 2021 were collected. In order to assure the duration of publications, quality of research publications and the scope of the present research work, only the papers published in the ACM Transactions on Database Systems (ACM TODS) [37] were considered. The collected list of 1031 titles from all the research papers of this duration was subjected to tokenization in order to extract the words from the titles. The tokenization was performed without considering the case of the words but maintaining the Multi-word phrases (MWP). Only unigrams were considered for the present research work. The resultant list consisted of 8139 words. This list could be considered a technical word list.

Cleaning was performed in this list to remove various noise words in context of the present research work. This constituted removal of unigrams like years (e.g. 1977, 2005, etc.), numbers (e.g. 3, 6, etc.) and special characters (e.g. *, #, etc.). The resultant list with 7994 words was used to find unique tokens. It is noteworthy that this point onwards, in order to emphasize the unique words in the list, we term the words as tokens. The count of such tokens was 1900. Stop words were removed from this list. We considered the 526 stop words provided by the standard Rainbow Stop Word List [38] for the present research work. The Rainbow Stop Word List had no MWP and its snapshot is provided in Table I. The resultant list was the refined technical list containing unique, lower-cased and non-stop-word 1791 tokens.

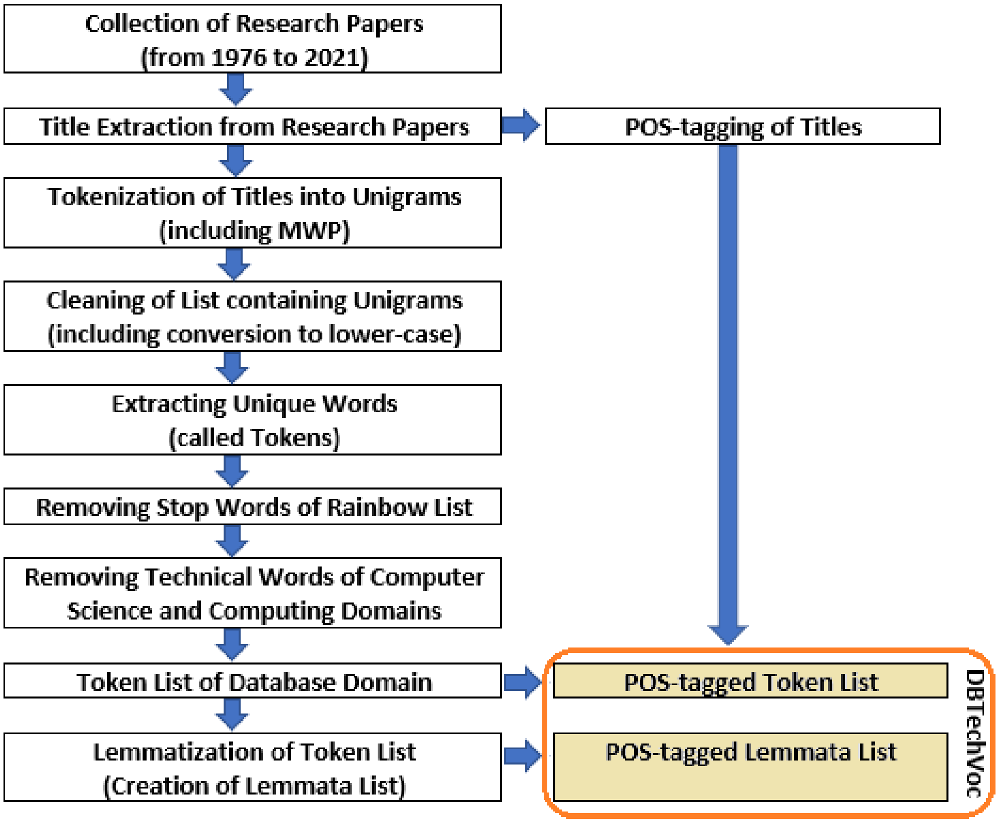
Collection of Research Papers
(from 1976 to 2021)
Title Extraction from Research Papers
POS-tagging of Titles
Tokenization of Titles into Unigrams
(including MWP)
Cleaning of List containing Unigrams
(including conversion to lower-case)
Extracting Unique Words
(called Tokens)
Removing Stop Words of Rainbow List
Removing Technical Words of Computer
Science and Computing Domains
Token List of Database Domain
POS-tagged Token List
Lemmatization of Token List
(Creation of Lemmata List)
POS-tagged Lemmata List
DBTechVoc

Fig. 1. Diagrammatic Representation of the Methodology to Create DBTechVoc.

The aim of present research work was the development of a database-specific vocabulary. Hence, at this stage, we referred another standard technical word list [15] containing the words from general domains of computer science as well as computing. The list in its raw form had 150 entries which were expanded to 252 words as there were many entries in the list with non-atomic values. For instance, the first entry with ‘access’ and ‘memory access’ was treated as two words viz. ‘access’ and ‘memory access’. It is notable that ‘memory access’ is a MWP and was treated as a unigram with the form ‘memory-access’. The snapshot of the expanded technical word list is presented in Table II. After the subtraction of 252 words from 1791 tokens, the resultant list had 1758 tokens. It is notable that only matching words were subtracted.

Finally, the POS-tagger [12] provided by the Stanford University was used for tagging the words in the paper titles. The resultant list had 2067 entries corresponding to 1758 tokens. The number of entries in the resultant list is more as the same token could be tagged with multiple parts of speech. As we were interested in the exhaustive coverage of the vocabulary, we considered all possible POS tags for a token.

Having created a POS-tagged token list of the database domain, the first part of the aim of creating an exhaustive vocabulary of the database domain, was achieved. For the second and last part, we targeted the creation of a POS-tagged lemmata list. This was achieved by lemmatizing the tokens in the list having 1758 tokens. It is noteworthy that we did not use the stemmer and used the Lemmatizer directly to obtain the lemma for each token rather than the non-lemma root for each token. The Lemmatizer [13] provided by the University of Copenhagen was used for this. The resultant list had 1530 lemmata.

The list with 1530 lemmata was further subjected to POS-tagging. The complete process was achieved through the use of multiple POS-taggers. Also, the different POS-taggers provided us with different sub-forms of POS tags but for simplicity the results were captured under the more common supersets. For instance, all entries from set {JJ (Adjective), JJR (Adjective, comparative), JJS (Adjective, superlative)} were considered to be just ‘adjective’ while entries from sets { NN (Noun, singular or mass), NNS (Noun, plural), NNP (Proper noun, singular), NNPS (Proper noun, plural)} and { VB (Verb, base form), VBD (Verb, past tense), VBG (Verb, gerund or present participle), VBN (Verb, past participle), VBP (Verb, non-3rd person singular present), VBZ (Verb, 3rd person singular present)} were considered to be ‘noun’ and ‘verb’ respectively.

Firstly, the Stanford University’s POS-tagger [12] was used which resulted in 1504 entries corresponding to 1078 lemmata. The remaining unprocessed 452 lemmata were attempted to be POS-tagged using the Princeton University’s POS-tagger [14]. This still resulted in 383 entries corresponding to only 261 additional lemmata. The remaining 191 remnant lemmata were manually POS-tagged. This resulted in 205 entries corresponding to 191 lemmata. Hence, the total number of entries corresponding to 1530 lemmata was 2092 in total. The summary of this data is presented in Table III.

TABLE I. A SNAPSHOT OF THE RAINBOW STOP WORD LIST [38]

Sr. No.	Stop Word
1	a
2	able
3	about
…	…
526	zero

TABLE II. A SNAPSHOT OF THE EXPANDED TECHNICAL WORD LIST FOR COMPUTER SCIENCE

Sr. No.	Technical Word
1	Access
2	access-time
3	Accumulator
…	…
252	Window

TABLE III. STATISTICS ON PROCESSING OF LEMMATA FOR POS USING DIFFERENT POS-TAGGERS

Sr. No.	POS-tagger	Lemma Count	POS Count
1	Stanford University POS-tagger [12]	1078	1504
2	Princeton University POS-tagger [14]	261	383
3	Manual POS-tagging (for remnant lemmata)	191	205
Total	3	1530	2092

Similar to the token POS-tagging case, for lemmata too there were multiple occurrences of a lemma having more than one POS-tag. Like the token POS-tagging case, in order to have an exhaustive coverage of the vocabulary of the technical domain of databases, we considered all possible POS tags for each lemma. The technical vocabulary of the database domain called DBTechVoc, which is a coined term, is formed by the POS-tagged token list and POS-tagged lemmata list.

IV. RESULTS AND DISCUSSION

The present research work was initiated with the motive of generating the technical vocabulary for the database domain. The titles of high-quality research papers in the field of database were believed to be the best source for populating the corpus. In order to make sure that no bias creeps in and also to assure that the basic terminology from the early days of development of databases as well as the latest terminology of the field of databases is covered, the duration of 45 years was considered for the present research work. Notably, this time period is not just long enough but also coinciding with the time period of evolution as well as proliferation of the field of databases. Also, it is both significant as well as relevant to consider the titles of research papers as something new in the field is first promulgated through a research paper and authors always include the important terms in the title of the research paper. It is with passage of time that those terms then become the part of the technical conversation of the field and thereby generating the technical field specific vocabulary.

TABLE IV. COMMON WORDS OF COMPUTER SCIENCE DOMAIN AND DATABASE DOMAIN

Sr. No.	Common Word	Sr. No.	Common Word
1	access	18	interface
2	allocation	19	interoperability
3	architecture	20	interpreter
4	backup	21	overhead
5	block	22	partition
6	buffer	23	pointer
7	cache	24	processor
8	capacity	25	protocol
9	disc	26	resolution
10	disk	27	retrieval
11	document	28	simulation
12	editor	29	software
13	error	30	statement
14	execution	31	storage
15	fragmentation	32	utility
16	hardware	33	window
17	instruction		

An important finding was obtained during the text processing with removal of stop words. It was observed that only 5.74% (or approx. 6%) of the tokens constituted the stop words. The same has been calculated using the formula: { [n(Tokens_with_SW) – n(Tokens_without_SW)] / n(Tokens_with_SW) } x 100 = Percentage of SW in Text; i.e. { [1900 – 1791] / 1900 } x 100 = 5.74 %. Here SW stands for Stop Words and n(entity) indicates the count of specified entity. This finding is in line with our assumption that the technical vocabulary of the database domain could be created from the titles of research papers as they contain more of important words rather than irrelevant words (like stop words). This holds true from multiple viewpoints of research, linguistics as well as statistics. Similarly, it was expected that a large number of words will be removed from the token list when computer science and computing domain technical word list will be considered. Actually, this step resulted in removal of just (1791-1758=) 33 common words. This means that there is only a { (33 / 1791) x 100 = 1.84% }, i.e. nearly 2% overlap of the technical lists of the domains of computer science and databases. The list of these removed common technical words is presented in Table IV. This finding is also very important and in line with our assumption that the technical word list of computer science domain will not be the same as the technical word list for the specific domain of databases.

Stemming was not used and directly lemmatization was used for the present research work to obtain the lemma for each token. Notably, this stage resulted in reduction of 13% entries from 1758 tokens to 1530 lemmata. This is important for the current context as it indicates the highly inflectional use of a few tokens by the researchers in the database domain. This is also important as it yielded a more refined vocabulary of the domain and hence let us meet the research objective.

TABLE V. DBTECHVOC (PART A): LIST OF TOKENS[A] AND CORRESPONDING POS

Sr. No.	Token	POS
1	abstract	noun
2	abstraction	noun
3	abstractions	noun
4	abstractions	verb
5	accelerating	noun
6	accelerating	verb
7	acceleration	noun
8	accesses	noun
9	accessibility	noun
10	account	noun
11	accuracy	noun
12	accurate	adjective
13	accurate	noun
14	achieving	noun
…	…	…
2065	xsketch	noun
2066	xsq	noun
2067	years	noun

[A]:Total unique tokens: 1758
Total unique POS: 5

TABLE VI. ANALYSIS OF FREQUENCIES OF POS TYPES OF TOKENS

Sr. No.	Token POS Type	Frequency	Share of POS Type (in %)
1	Noun	1341	64.88
2	Adjective	364	17.61
3	Verb	332	16.06
4	Adverb	25	1.21
5	Foreign Word (FW)	5	0.24
Total	5	2067	100

After following the various stages of methodology mentioned in section III, a final refined list of tokens was generated which was further POS-tagged. A snapshot of this list is presented in Table V. This table presents the glimpse of first 14 POS-tagged tokens and last 3 POS-tagged tokens from a total of 2067 POS-tagged tokens corresponding to 1758 unique tokens fortified with 5 unique POS. A summary on frequency of these unique 5 POS tags for this list is presented in Table VI. It can be observed from Table VI that nouns followed by adjectives constitute more than 82% of the total POS types.

Similar to the POS-tagged token list, another list for POS-tagged lemmata was also generated. A snapshot of this list is presented in Table VII. This table presents the glimpse of first 11 and last 8 POS-tagged lemmas out of a total of 1859 such POS-tagged lemmas corresponding to 1530 unique lemmas. A summary on frequency of the 5 unique POS tags found for this list (already presented in Table VII) is presented in Table VIII. It can be observed from Table VIII that the nouns and adjectives together constitute more than 80% of all the POS tags.

TABLE VII. DBTECHVOC (PART B): LIST OF LEMMATAA AND CORRESPONDING POS

Sr. No.	Lemma	POS
1	abstract	noun
2	abstraction	noun
3	acceleration	noun
4	access	noun
5	access	verb
6	accessibility	noun
7	account	noun
8	accuracy	noun
9	accurate	adjective
10	accurate	noun
11	achieve	verb
…	…	…
1852	xml	noun
1853	xpath	noun
1854	xqbe	noun
1855	xquery	noun
1856	xsd	noun
1857	xsketch	noun
1858	xsq	noun
1859	year	Noun

A:Total unique lemma: 1530
Total unique POS: 5

To summarize, Table V and Table VII present the POS-tagged token list and lemmata list respectively. The frequency break-up of the unique POS tags for Table V and Table VII is presented respectively in Table VI and Table VIII. The two lists viz. POS-tagged token list and POS-tagged lemmata list, together constitute the technical vocabulary of the database domain and have been addressed with a coined term DBTechVoc. Table V and Table VII represent the two parts, viz. A and B for DBTechVoc. The lists are presented in ascending order of the tokens and lemmata respectively. Similarly, the data in Table VI and Table VIII is sorted on the frequency of the POS-tag. Notably, both the tables ended up with same order of the POS-tags though their frequencies were different for the lists corresponding to tokens and lemmata. Fig. 2 presents the share (in units of percentage of the total count) of POS type for tokens and lemmata. It can be observed that there is no much difference between the breakup of POS types for tokens and lemmata. Notably, the number of adverbs, verbs and adjectives are more in case of lemmata list compared to those in the list of tokens.

TABLE VIII. ANALYSIS OF FREQUENCIES OF POS TYPES OF LEMMATA

Sr. No.	Lemmata POS Type	Frequency	Share of POS Type (in %)
1	Noun	1129	60.73
2	Adjective	365	19.63
3	Verb	329	17.70
4	Adverb	31	1.67
5	Foreign Word (FW)	5	0.27
Total	5	1859	100

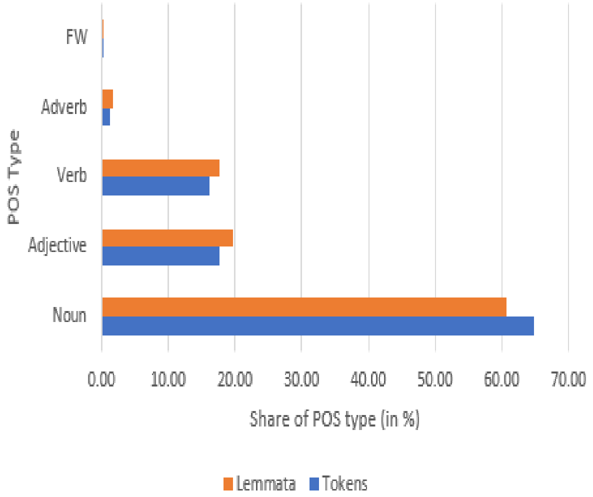

Fig. 2. Representation of Share (in %) of POS Types for Tokens and Lemmata.

In order to complement the vocabulary analysis of the database technical domain and to visualize the results of graphically, the word cloud, presented in Fig. 3, was generated. The word cloud was generated without stemming and lemmatization of the words though stop words were excluded from the list. Internal stop-word list was used during the execution of the code. The case of words was not considered and only unique words were considered for rendering through the cloud. In order to maintain the sanctity of data, the Multi-word Phrases (MWP) or the word formations with multiple words joined together with a hyphen like 'entity-relational', 'multi-valued' and 'grammar-based', just to name a few, were considered as-it-is without ignoring the hyphen. The number of words satisfying all these criteria was 1900 and out of these the cloud could accommodate the top 654 words. The frequency break-up of the remaining 1246 words which were not drawn through the cloud is given in Table IX.

As is clear from Fig. 3, the top most frequent word was 'database'. In order to further refine our analysis, the top 20 words were subjected to lemmatization. This resulted in the reduction of count and leading to 16 unique lemmata. These lemmata along with their corresponding Part-Of-Speech (POS) are shown in Table X. Notably, other than one verb and two adjectives, all other lemmata are nouns. This leads to an important inference that the authors tend to use more of nouns in the paper titles, at least for the research domain of databases.

TABLE IX. FREQUENCY BREAK-UP OF WORDS NOT DRAWN THROUGH THE CLOUD

Frequency of words	Number of words
1	940
2	175
3	111
4	7
5	10
6	3
Total	1246

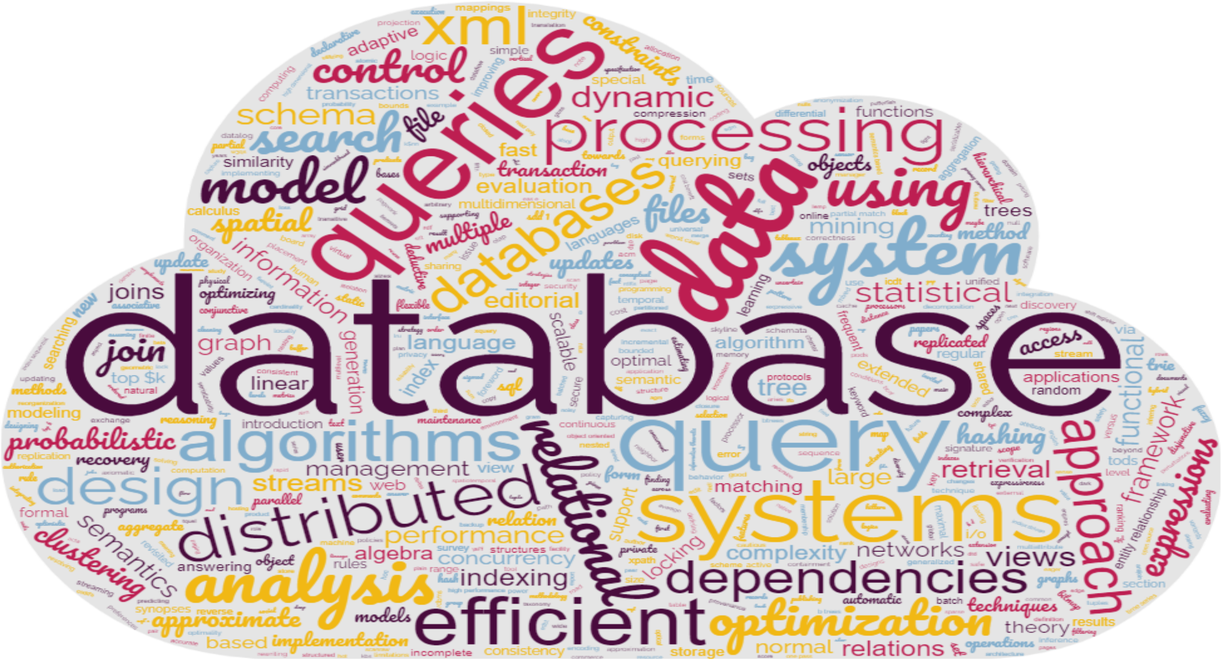

Fig. 3. Word Cloud of the Most Frequent Terms in the Database Vocabulary Corpus Created for 45 Years.

TABLE X. TOP 16 LEMMATA AND CORRESPONDING POS

Sr. No.	Lemmata	POS
1	algorithm	Noun
2	analysis	Noun
3	approach	Noun
4	data	Noun
5	database	Noun
6	design	Noun
7	distribute	Verb
8	efficient	Adjective
9	information	Noun
10	language	Noun
11	model	Noun
12	processing	Noun
13	query	Noun
14	relational	Adjective
15	system	Noun
16	xml	Noun

As the proposed work is unique and first of its kind, its comparison as well as performance evaluation with respect to existing works is not feasible. However, the proposed work is better than the existing ones in terms of presenting a more specific vocabulary as well as better annotated vocabulary of the technical words of the database sub-domain of the computer science domain.

V. Conclusion, Limitations and Future Work

The present research work is the first formal attempt to create a technical vocabulary for the domain of databases. This vocabulary called DBTechVoc consists of a POS-tagged token list having 1758 multi-word phrase unigrams and a POS-tagged lemmata list having 1530 multi-word phrase unigrams. It is noteworthy that most of the Natural Language Processing (NLP) applications for generation of various word lists generally do not consider the multi-word phrases owing to the ease of processing that way. It is remarkable that as the present research work intended to create a technical vocabulary without the loss of semantic information of the technical phrases, the multi-word phrases have been well considered.

The various results and findings of the present research work are bound to have a good ripple effect for the researchers working in the same and similar fields. From the processing of more than 1000 research papers of last 45 years, it is concluded that the authors use 6% stop words in the titles of the research papers. Also, 13% of the words used for the research papers titles are inflectional forms of lemmata from a set consisting of tokens from the technical domain. There is a negligible overlap between the technical word lists for computer science domain and database domain. Also, based on perhaps first of its kind comparison between the frequency break-up of POS categories for tokens and frequency break-up of POS-categories of the corresponding lemmata of the tokens, it is concluded that the lemmatization results in increase in the number of adverbs, verbs and adjectives while reducing the number of nouns. Though the results reported here are for the technical domain of databases, they could be applied to other technical domains also as the period of 45 years and more than 1000 research papers is believed to be enough to normalize the values. All these results could be applied for analysis of and investigation on various works including the usage of these results as an aid to solve cases dealing with plagiarism as well as author attribution. This application may include research papers as well as other literary works, including touching on the areas of violation of copyrights and other intellectual property rights.

DBTechVoc itself could be used for various downstream tasks dealing with NLP of technical domains. DBTechVoc lists or the derived ones could also be used as a source of stop-words for some advanced applications dealing with the processing of this technical domain specific textual data, for instance, processing of social media reviews or opinions or comments on a particular topic dealing with the field of databases. The presented lists could also be used for generation of artificial language specific to the database domain. The lists could also be used for the readability analysis of the technical domains, particularly the databases. The proposed lists could also be used alongside the technical word list of the field of computer science in general as it happens to be the parent field of the domain of databases. Additionally, the lists could also be used for word-embeddings, Machine Translation Systems (MTS) and generation of a domain-specific technical WordNet.

One of the limitations of the present research work is that it presents the technical vocabulary of only the database domain. Also, though standard stop word list, technical word list, Lemmatizers and POS-taggers have been used, the results may differ if a different combination of these items is used. The findings, results and technical vocabulary presented here are all best reported as per the context and scope of the present research work. Though we believe that the proposed list tends to be exhaustive as on moment, it is notable that the field of database, like any other technical field, keeps on evolving and with passage of time, new words could be added to the domain. As future work, in addition to keeping the lists updated with appropriate versioning, we plan to consider the other parts of the published research papers like abstract, keywords, manuscript body, etc. for further fortifying the research methodology. Also, in addition to just the unigrams, bigrams, trigrams, etc. could also be considered for the vocabulary creation. Most importantly, with the measurement of semantic similarity between the tokens, we are working to generate a technical wordnet specifically for the database domain.

References

[1] Liu N., Nation I.S.P. (1985), "Factors affecting guessing vocabulary in context", RELC Journal, 16(1):33–42.

[2] Dewan P., Gupta P. (2016), "Writing the Title, Abstract and Introduction: Looks Matter!", Indian Pediatrics, 53:235-241. Online: https://www.indianpediatrics.net/mar2016/mar-235-241.htm.

[3] Tullu M.S. (2019), "Writing the title and abstract for a research paper: Being concise, precise, and meticulous is the key", Saudi Journal of Anaesthesia, 13(1):S12-S17. doi: 10.4103/sja.SJA_685_18.

[4] Mack C. (2012), "How to write a good scientific paper:title, abstract, and keywords", Journal of Micro/Nanolithography, MEMS and MOEMS, 11(2):020101-1--020101-4.

[5] Karagel H., Karagel D.U. (2014), "Identification and importance of headings and key words in research in the framework of geography methodology", Procedia - Social and Behavioral Science, 120:356-364. doi: 10.1016/j.sbspro.2014.02.113.

[6] Soler V. (2007), "Writing titles in science: An exploratory study", English for Specific Purposes, 26:90–102. doi:10.1016/j.esp.2006.08.001.

[7] Hengl T., Gould M. (2002), "Rules of thumb for writing research articles", Enschede, pp.1-9. Online: https://webapps.itc.utwente.nl/librarywww/papers/hengl_rules.pdf.

[8] Webster J.J., Kit C. (1992), “Tokenization as the Initial Phase in NLP”, in proceedings of COLING-92, pp. 1106-1110. Online: https://aclanthology.org/C92-4173.pdf.

[9] Kyle C. (1989), "Double, Triple, and Quadruple Bigrams," Word Ways, 22(3), art. 8. Online: https://digitalcommons.butler.edu/wordways/vol22/iss3/8.

[10] Akhmetov I., Pak A. Ualiyeva I., Gelbukh A. (2020), "Highly Language-Independent Word Lemmatization Using a Machine-Learning Classifier", Computación y Sistemas, 24(3):1353–1364. doi: 10.13053/CyS-24-3-3775.

[11] Maggini M. (n.d.), "Natural Language Processing Part 2: Part of Speech Tagging ", Teaching Slides, Department of Information Engineering and Mathematical Sciences, University of Siena, Italy. Online: https://www3.diism.unisi.it/~maggini/Teaching/TEL/slides%20EN/06%20-%20NLP%20-%20PoS%20Tagging.pdf.

[12] Stanford University (n.d.), "Part-Of-Speech (POS) Tagger", The Stanford Natural Language Processing Group. Online: http://nlp.stanford.edu:8080/parser/index.jsp.

[13] University of Copenhagen (n.d.), "CST's Part-Of-Speech tagger", Center for Language Technology. Online: https://cst.dk/online/pos_tagger/uk/.

[14] Princeton University (n.d.), "WordNet Search-3.1", WordNet: A Lexical Database for English. Online: http://wordnetweb.princeton.edu/perl/webwn.

[15] Smith S. (2019). (n.d.). Online: https://www.eapfoundation.com/vocab/other/lists/#thetable.

[16] West M. (1953), "A General Service List of English Words", London: Longman, Green and Co.

[17] Brezina V., Gablasova D. (2015), "Is There a Core General Vocabulary? Introducing the New General Service List", Applied Linguistics, 36(1):1–22. doi: 10.1093/applin/amt018.

[18] Browne C. (2013), "The New General Service List: Celebrating 60 years of Vocabulary Learning", The Language Teacher, 4(37):13–16.

[19] Gilner L., Morales F. (2008), "Elicitation and application of a phonetic description of the General Service List", System, 36(4):517–533.

[20] Gilner L. (2011), "A primer on the General Service List", Reading in a Foreign Language, 23(1):65–83. Online: https://files.eric.ed.gov/fulltext/EJ926367.pdf.

[21] Nation P., Waring R. (2004), "Vocabulary size, text coverage and word lists". Online: https://web.archive.org/web/20080111133710/http://www.wordhacker.com/.

[22] Kaur J., Saini J.R. (2015), "POS Word Class based Categorization of Gurmukhi Language Stemmed Stop Words", in proceedings of Springer International Conference on ICT for Intelligent Systems (ICTIS-2015), Ahmedabad, India, 51(2):3-10. doi: 10.1007/978-3-319-30927-9_1.

[23] Kaur J., Saini J.R. (2016), "Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle ", in proceedings of Symposium on ACM Women in Research (ACM-WIR-2016), Indore, India, 01188:32-37. doi: 10.1145/2909067.2909073.

[24] Rakholia R.M., Saini J.R. (2017), "A Rule-based Approach to Identify Stop Words for Gujarati Language", in proceedings of The 5th Springer International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA-2016), Bhubaneshwar, India, 515:797-806. doi: 10.1007/978-981-10-3153-3_79.

[25] Rakholia R.M., Saini J.R. (2016), "Lexical Classes Based Stop Words Categorization for Gujarati Language", in proceedings of 2nd IEEE International Conference on Advances in Computing, Communication & Automation (ICACCA-2016), Bareilly, India, pp. 1-5. doi: 10.1109/ICACCAF.2016.7749005.

[26] Hancioglu N., Neufeld S., Eldridge J. (2008), "Through the looking glass and into the land of lexico-grammar", English for Specific Purposes, 27(4):459-479. doi: 10.1016/j.esp.2008.08.001.

[27] Coxhead A. (2000), "A new Academic Word List", TESOL Quarterly, 34(2):213–238.

[28] Billurog˘lu A., Neufeld S. (2005), "The Bare Necessities in Lexis: A new perspective on vocabulary profiling". Online: http://lextutor.ca/vp/BNL_Rationale.doc.

[29] Bakaric M.B., Babic N., Matetic M. (2021), "Application-based Evaluation of Automatic Terminology Extraction", International Journal of Advanced Computer Science and Applications, 12(1):18-27. doi: 10.14569/IJACSA.2021.0120103.

[30] Choy M. (2012), "Effective Listings of Function Stop words for Twitter", International Journal of Advanced Computer Science and Applications, 3(6):8-11. doi: 10.14569/IJACSA.2012.030602.

[31] Raulji J.K., Saini J.R. (2017), "Generating Stopword List for Sanskrit Language", in proceedings of 7th IEEE International Advance Computing Conference (IACC-2017), Hyderabad, India, pp. 799-802. doi: 10.1109/IACC.2017.0164.

[32] Raulji J.K., Saini J.R. (2020), "Sanskrit Stopword Analysis through Morphological Analyzer and its Gujarati Equivalent for MT System", in proceedings of International Conference on ICT for Sustainable Development (ICT4SD-2019), Panaji, India, 93:427-433. doi: 10.1007/978-981-15-0630-7_42.

[33] Venugopal G., Saini J.R., Dhanya P. (2020), "Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List", International Journal of Advanced Computer Science and Applications, 11(1):233-239. doi: 10.14569/IJACSA.2020.0110130.

[34] Saini J.R., Rakholia R.M. (2016), "On Continent and Script-wise Divisions-based Statistical Measures for Stop-words Lists of International Languages", Procedia Computer Science, 89:313-319. doi: 10.1016/j.procs.2016.06.076.

[35] (n.d.) "Database Terminology – A Dictionary of the Top Database Terms". Online: https://raima.com/database-terminology/.

[36] Fayaza M.S.F., Farhath F.F. (2021), "Towards Stopwords Identification in Tamil Text Clustering", International Journal of Advanced Computer Science and Applications, 12(12):524-529. doi: 10.14569/IJACSA.2021.0121267.

[37] The ACM Digital Library (2022), “ACM Transactions on Database Systems”, The Association for Computing Machinery. Online: https://dl.acm.org/journal/tods.

[38] Shuson N. (2022), “StopWords list based on Rainbow statistical text”. Online: https://gist.github.com/shuson/b3051fae05b312360a18.

[39] Saed H., Hussein R.F., Haider A.S., Al-Salman S., Odeh I.M. (2022), "Establishing a COVID-19 lemmatized word list for journalists and ESP learners", Indonesian Journal of Applied Linguistics, 11(3): 577-588. doi: 10.17509/ijal.v11i3.37103.

[40] Das S.R., Donini M., Zafar M.B., He J., Kenthapadi K. (2022), "FinLex: An effective use of word embeddings for financial lexicon generation", The Journal of Finance and Data Science, 8:1-11. doi: 10.1016/j.jfds.2021.10.001.

[41] Ahsanuddin M., Hanafi Y., Basthomi Y., Taufiqurrahman F., Bukhori H.A., Samodra J., Widiati U., Wijayati P.H. (2022), "Building a corpus-based academic vocabulary list of four languages", Pegem Journal of Education and Instruction, 12(1):159–167. doi: 10.47750/pegegog.12.01.15.

[42] Joensuu J. (2022), "Culinary List Form in the Experimental Poetry of 1960s Finland: Literary Menus and Recipes", in: Barton R.A., Böckling J., Link S., Rüggemeier A. (eds) Forms of List-Making: Epistemic, Literary, and Visual Enumeration, Palgrave Macmillan, Cham. doi: 10.1007/978-3-030-76970-3_9.

[43] Wingrove P. (2022), "Academic lexical coverage in TED talks and academic lectures", English for Specific Purposes, 65:79-94. doi: 10.1016/j.esp.2021.09.004.

[44] Alasmary A. (2022), "Academic lexical bundles in graduate-level math texts: A corpus-based expert-approved list", Language Teaching Research, 26(1):99-123. doi: 10.1177/1362168819877306.