Jurnal Bahasa, Sastra, Budaya, dan Pengajarannya (Protasis)
Vol.2, No.1 Juni 2023
e-ISSN :2829-727x p-ISSN :2829-5862, Hal 204-214
DOI: https://doi.org/10.55606/protasis.v2i1.90

High Frequency Words in The Reading Section of TOEFL PBT Exercises

Aghnia Dianah Anwar
Faculty of Humanities, Universitas Gadjah Mada

Tofan Dwi Hardjanto
Faculty of Humanities, Universitas Gadjah Mada

Jln. Nusantara No.1 Bulaksumur Yogyakarta 55281, Indonesia

Corresponding author: aghniaanwar@gmail,com

Abstract. *English profeciency is one of the most important prerequisites to be fulfilled in applying for educational degrees, scholarships, or even jobs nowadays. Test of English as a Foreign Language (TOEFL) is one of the most common english profeciency tests to be taken in Indonesia especially TOEFL PBT (paper-based), although most countries across the world now prefer to use TOEFL iBT. This study examines the high frequency words in TOEFL ITP exercises and their morphological processes using a corpus data taken from the reading practices of Longman TOEFL PBT Preparation book. This study is a mix between a quantitative and a qualitative research. In finding the HFWs, the author used a software called AntConc. The results of the study showed that from 100 HFWs found in the reading section of TOEFL PBT exercises, the ranks of the most frequent word class to appear from highest to lowest are Noun (30 words), Verb (17 words), Adverb (15 words), Preposition (13 word), Determiner (11 words), Adjective (5 words), Number (4 words), Conjunction (4 words) and 1 Contraction. Meanwhile, the 10 most frequent words to appear among those 100 words are “the”* (9,64%), “*of*” (4,24%), “*in*” (3,24%), “*to*” (2,56%), “*a*” (2,18%), “*is*” (1,77%), “*and*” (1,75%), “*that* (1,34%), “*passage*” (1,1%), and “*on*” (0,9%).

Keywords: *Corpus Linguistics, High Frequency Words, TOEFL PBT*

Abstrak. Kemahiran berbahasa Inggris merupakan salah satu prasyarat terpenting yang harus dipenuhi dalam melamar gelar pendidikan, beasiswa, atau bahkan pekerjaan saat ini. Tes Bahasa Inggris sebagai Bahasa Asing (TOEFL) adalah salah satu tes yang menguji kemahiran seseorang dalam berbahasa Inggris yang paling umum diambil di Indonesia, terutama TOEFL PBT (berbasis kertas), meskipun sebagian besar negara di dunia sekarang lebih suka menggunakan TOEFL iBT. Penelitian ini mengkaji kata-kata berfrekuensi tinggi dalam latihan TOEFL ITP dan proses morfologinya menggunakan korpus data yang diambil dari latihan membaca buku Longman TOEFL PBT Preparation. Penelitian ini merupakan campuran penelitian kuantitatif dan kualitatif. Dalam mencari HFW, penulis menggunakan perangkat lunak bernama AntConc. Hasil penelitian menunjukkan bahwa dari 100 HFW yang ditemukan pada bagian reading latihan TOEFL PBT, peringkat kelas kata yang paling sering muncul dari tertinggi ke terendah adalah *Noun* (30 kata), *Verb* (17 kata), *Adverb* (15 kata). kata), *Preposition* (13 kata), *Determiner* (11 kata), *Adjective* (5 kata), *Number* (4 kata), *Conjunction* (4 kata) dan 1 *Contraction*. Sedangkan 10 kata yang paling sering muncul diantara 100 kata

Received April 23, 2023; Revised Mei 13, 2023; Accepted Juni 15, 2023

* Aghnia Dianah Anwar, aghniaanwar@gmail,com

tersebut adalah *"the"* (9,64%), "*of*" (4,24%), "*in"* (3,24%), "*to"* (2,56%), "*a"* (2,18%), "*is"* (1,77%), "*and"* (1,75%), "*that"* (1,34%), "*passage"* (1,1%), dan "*on"* (0,9%).

Kata kunci: Kata berfrekuensi tinggi, linguistik korpus, TOEFL PBT

INTRODUCTION

In the 4.0 era, English is becoming increasingly important to master as a global communication tool, which allows humans to work together effectively and access a wider range of resources and information. English is currently used by almost 53 countries as their official language and is spoken by approximately 400 million people around the world. English is most commonly used in global communication, both in business, politics, education and everyday life. By mastering English, one can understand and communicate with people from different countries and different cultures, thereby opening up opportunities to broaden social and professional networks. As a result, English has become one of the competencies that must be mastered at the educational level as well as for pursuing a career path. One of the tests which is widely used to asses someone's proficiency in English is TOEFL (Test of English as a Foreign Language). It surveys one's capacity to both talk and comprehend English by breaking down their English capacity reading, speaking, listening and writing. These abilities are significant in completing their scholastic examinations. The test is frequently taken by understudies who are wanting to learn at a college abroad and those who want to get scholarship, alongside understudies and laborers who are applying for visas and English-language students following their English advancement.

There are three types of TOEFL tests which are TOEFL PBT (Paper-based Test), TOEFL CBT (Computer-based Test), and TOEFL iBT (Internet-based Test). Currently, TOEFL iBT is the one which is internationally accepted. However, in Indonesia, TOEFL PBT is still the most common type of TOEFL tests to be taken. One reason is due to its cheaper price compared to TOEFL iBT. TOEFL PBT consists of 3 sessions namely Listening, Structure and Written Expression, and Reading. The score of TOEFL PBT ranges from 310 to 677 as the maximum score. Local universities, companies, and institutions make it mandatory for applicants to have a TOEFL PBT certificate. The most common minimal score of TOEFL PBT as a prerequisite in applying for entry to a local education degree or for a job position is around 400-500. Some scholarships, for example Indonesia Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan* or LPDP), require their applicants to have a minimal score of 550 if they seek to enter international universities. Lots of test takers have difficulty achieving the desired score. Often, they repeatedly take the TOEFL test until their score meets the target. They are also willing to spend a lot of money to buy practice books and take TOEFL preparation courses.

One of the difficulties faced by test takers is related to the lack of vocabulary they know. Vocabulary learning is viewed as quite possibly of the most fundamental angle in learning a language (Jin, et al., 2012). It is particularly pivotal in the process of mastering a new or a second language. The significance of vocabulary learning emerged with regards to global English capability assessments such as the TOEFL test. Variations of methods in teaching and expanding vocabularies is an important matter to be taken into account in order to aid test takers with enough number of vocabularies. One way is to teach vocabularies through unplanned lessons, for example in students' assignments. Meanwhile, some teachers prefer to use the traditional way of teaching (Sutarsyah, Nation & Kennedy, 1994). Another way of teaching vocabularies can be done through recycling and repetition which would be beneficial in students' acquisition (Mukundan, 2006; 2009).

This study aims to discover and analyze the high frequency words (HFW) in TOEFL PBT questions, specifically in the reading section since it is the section in which many vocabularies are testes according to context. However, real TOEFL questions are confidential documents and it is quite a challenge to obtain them. Therefore, the author has chosen some of the most popular and widely used TOEFL preparation books containing exercises which, hopefully, will give a glimpse on the kinds of words usually tested in TOEFL PBT. Those books are Phillip's "Longman Complete Course for the TOEFL Test: Preparation for the Computer and Paper Tests and three books published by ETS namely "Official Guide to the TOEFL ITP Test", "TOEFL ITP Level 1 Practice Tests, Volume 1", and "TOEFL ITP Level 1 Practice Tests, Volume 3". The reading section of each practice found in those books are compiled into corpus data. In order to obtain the HFWs, the data are then inputted into a software called AntConc. The HFWs lists are then analyzed based on their word categories.

There have been several studies regarding HFWs in TOEFL or in other English modules. In 2012, Jin, et al. examined HFWs in TOEFL course books in their paper entitled "Corpus Based Analysis of the TOEFL Course Books: What are the Words We Should Teach Our Students?". 100 technical content words in TOEFL course books according to their significant frequency in the passages found in the reading section of TOEFL iBT are presented in the study. Kaneko (2020) did research entitled "Lexical Frequency Profiling of High-Stakes English Tests: Text Coverage of Cambridge First, EIKEN, GTEC, IELTS, TEAP, TOEFL, and TOEIC". The study aimed to determine the size targets of vocabularies for English tests used in university admission examinations in Japan from 2020. The results showed that in the reading passages, the most frequent 2,000 to 5,000 word families were necessary to yield 95% lexical coverage. In 2016, Pribakht and Webb published an article entitled "The Relationship between Academic Vocabulary Coverage and Scores on a Standardized English Proficiency Test". This study examined the occurrence of items from the Academic Word List (AWL) in 12 differen English proficiency test which are mandatory to be taken to be admitted to universities in Canda. The results indicated that AWL coverage in the passages was consistently present and substantial although below the established level of coverage found in academic texts. The novelty that this present study offers that it examines the HFWs in the reading section of TOEFL PBT using a corpus-based analysis. This study also uses a different software, which is AntConc, to obtain the HFWs found in the texts.

Words that teachers-students encounter in teaching materials can experience repetition several times or even appear only once in the teaching material. According to Jones and Durrant (2010: 390), frequency of occurrence is the main identification criterion in corpus linguistics. Shaw (2011: 16) stated that lists of HFWs are beneficial for educators in building a strong foundation of vocabularies to be taught and they can utilize corpus data to pick vocabularies from specific units or certain books to teach according to how often those words appear. Jones and Durrant (2010: 387) also mentioned that the argument for prioritizing vocabulary learning based on frequency information is based on the principle that the more often a word is, the more important it is to learn. Similarly, Criado and Sánchez (2012: 78) argued that repetition of forms has become one of the most important strategies in language education to encourage learning and consolidate knowledge.

One area of linguistics related to vocabulary knowledge is morphology. According to Verhaar (1992: 52), morphology is a field of linguistics that studies the grammatical arrangement of word parts and morphology. Chaer (2014: 15) states that morphology investigates the structure of words, their parts, and how they are formed. According to Ramlan in Kesuma (2020), morphology is a part of linguistics that discusses or studies the intricacies of word forms and the effect of changes in word forms on word class and meaning. Kridalaksana in Kesuma (2020) reveals that morphology is a field of linguistics that studies morphemes and their combinations. From this definition it can be seen that morphology examines questions related to word forms (Kesuma, 2020). According to Muhadjir (1996: 2) in Sukarto (2006: 10), one of four ways to show the emergence of vocabulary is by listing the frequency of vocabulary which can be arranged in which morphological form a basic form appears most often.

METHODS

According to Sudaryanto (2015: 3), language research activities are divided into two sessions: problem finding and problem solving. The search for problems is manifested in the formulation of the problem. Then these problems were solved through three stages, namely providing data, analyzing data, and presenting the results of data analysis. Empirical data in the data preparation stage are sourced from texts from the reading sections of three TOEFL PBT practice books, namely Phillip's "Longman Complete Course for the TOEFL Test: Preparation for the Computer and Paper Tests and three books published by ETS namely "Official Guide to the TOEFL ITP Test", "TOEFL ITP Level 1 Practice Tests, Volume 1", and "TOEFL ITP Level 1 Practice Tests, Volume 3".

The data is stored in .docx format and then uploaded to a concordance software called Ant.Conc. The list of HFWs is obtained through the Word List feature. Crawford (2016: 91) stated that word list analysis can create a simple word list with the number of frequencies and ratings from your own data set. Weisser (2016: 151) states that to generate a word frequency list in AntConc is to load the corpus, select the 'Word List' tab, and click Start (Weisser, 2016: 151). Next, AntConc displays a list of words in the data according to the frequency of their occurrence. Based on the process of the Word List feature at AntConc, the number of word types and tokens was found from the data as well as the word list which is equipped with the ranking and frequency of use of each word in the word list. The data are then analyzed based on their word categories according to the theory of morphology.

RESULTS AND DISCUSSION

1.1. High-Frequency Words in the Reading Sections of TOEFL PBT exercises

After inputting the corpus data into AntConc, the author first performs data reduction before determining the HFWs. Arum and Winarti (2020: 64) carried out the reduction stage to determine the degree of relevance to the research objectives. In this study data reduction was carried out to obtain data that could be further analyzed according to the research objectives. The reduced data are in the form of: (1) single letters (except articles); (2) syllables or word fragments; (3) words that contain hyphens but are not repeated words; (4) personal name; and (5) foreign words. After data reduction, it was discovered that there were 2813 word types and

23483 tokens in the data. Desagulier (2017: 108) stated that the principle of rooted frequency lists is the difference between types and tokens. Crawford (2016: 93) explained that word tokens are every word in a text, while word types are every type of word in a text. Among those words, the list of 100 HFWs in the reading sections of TOEFL PBT exercises were obtained which are shown in the following table.

Table 1. The List of 100 HFWs in the reading sections of TOEFL PBT exercises

No	Word	Frequency	No	Word	Frequency
1	the	2266	51	when	48
2	of	996	52	were	46
3	in	761	53	where	46
4	to	602	54	closest	45
5	a	514	55	american	42
6	is	416	56	each	42
7	and	411	57	look	41
8	that	315	58	moon	40
9	passage	260	59	probably	40
10	on	214	60	lines	39
11	it	190	61	united	39
12	was	176	62	into	38
13	be	155	63	according	37
14	from	153	64	other	37
15	line	146	65	used	37
16	by	144	66	what	37
17	for	143	67	would	37
18	word	137	68	earth	36
19	as	136	69	time	36
20	are	131	70	only	33
21	this	129	71	all	32
22	paragraph	123	72	sentence	32
23	which	122	73	three	32
24	s	119	74	blood	29
25	questions	116	75	during	29
26	following	105	76	first	29
27	at	104	77	how	29
28	an	103	78	but	28
29	not	102	79	replaced	28
30	with	100	80	age	27
31	one	90	81	type	27

32	he	84	82	paper	26
33	his	82	83	refers	26
34	can	80	84	stars	26
35	best	78	85	these	26
36	most	77	86	war	26
37	states	71	87	because	25
38	or	67	88	had	25
39	than	65	89	president	25
40	two	64	90	their	25
41	have	62	91	music	24
42	been	60	92	sun	24
43	its	59	93	water	24
44	reading	59	94	who	24
45	they	58	95	year	24
46	about	57	96	century	23
47	could	56	97	fever	23
48	has	56	98	however	23
49	meaning	54	99	period	23
50	more	48	100	such	23

According to the HFW list, there are ten words which appear most frequently in the reading section of TOEFL PBT exercises namely *the* (9,64%), *of* (4,24%), *in* (3,24%), *to* (2,56%), *a* (2,18%), *is* (1,77%), *and* (1,75%), *that* (1,34%), *passage* (1,1%), and *on* (0,9%).

1.2 Word Classification of HFWs in the Reading Section of TOEFL PBT Exercises

The HFWs found in the reading section of the TOEFL PBT exercises are classified based on their word class. These words are classified based on the theory of Gelderen (2010) which explains that lexical categories (word classes) in English are syntactically divided into 5 categories, namely Noun (N), Verb (V), Adjective (Adj), Adverb (Adv), and Prepositions (P).

Table 2. Word Classification of 100 HFWs in the Reading Section of the TOEFL PBT Exercises

No	Word Class	Word	Number
1	Noun (N)	*passage, line, word, paragraph, questions, following, states, american, moon, lines, earth, time, sentence, blood, age, type, paper, stars, war, president, music, sun, water, year, century, fever, period* ***Pronouns**: they, it, he*	30
2	Verb (V)	*is, was, be, were, are, have, been, reading, could, has, look, used, replaced, refers, would, had, can*	*17*
3	Adjective (Adj)	*best, meaning, closest, united, other*	*5*
4	Adverb (Adv)	*as, that, which, not, most, when, where, probably, according, what, only, all, how, who, however*	*15*
5	Preposition (P)	*of, in, to, on, from, by, for, at, with, than, about, into, during 13*	*13*
6	Others	***Number (Num)**: one, two, three, first (4)* ***Determine (Det)r:** the, a, this, an, his, its, more, each, their, such, these (11)* ***Conjunction (Conj):** and, or, but, because (4)* ***Contraction (Cont):** 's (1)*	*20*

Table 2 shows the classification of 100 HFWs in the reading section of the TOEFL PBT exercises. Based on the table the ranks of word classes from highest to lowest are N (30 words), V (17 words), Adv (15 words), P (13 word), Det (11 words), Adj (5 words), Num (4 words), Conj (4 words) and 1 contraction.

After all the data has been classified based on word class, an analysis is performed on the vocabulary included in the open classes (nouns, adjectives, verbs, and adverbs) based on Gelderen (2010) and Quirk. at al (1980). Besides being seen from their meanings, the use of the ending -s as a singular noun marker and plural, the mention of the name of a person or place, the noun features found in the noun class are as follows.

Countable noun, such as *word, line, paragraph, war*

Uncountable noun, such as *water, blood*

Proper Nouns, such as *American*

Common Noun, such as *music, year, century*

Concrete Noun, such as *moon, stars, sun*

Abstract Noun, such as *fever, age*

There are also some pronouns found in the data such as *they, he, it.*

Meanwhile, verb features which are found in the 100 HFWs in the reading section of

TOEFL PBT exercises are as follows.

Main verbs such as *reading, replaced, refers*

Auxiliary verbs such as *had, were, is, has, have*

Modal auxiliary verbs: *could, would, can*

For adjectives, there is one form of ***superlative adjective*** which is *best.* Furthermore, there are several types of adverbs found in the data as follows

Adverb of degree such as *only, all*

Adverb of probability such as *probably*

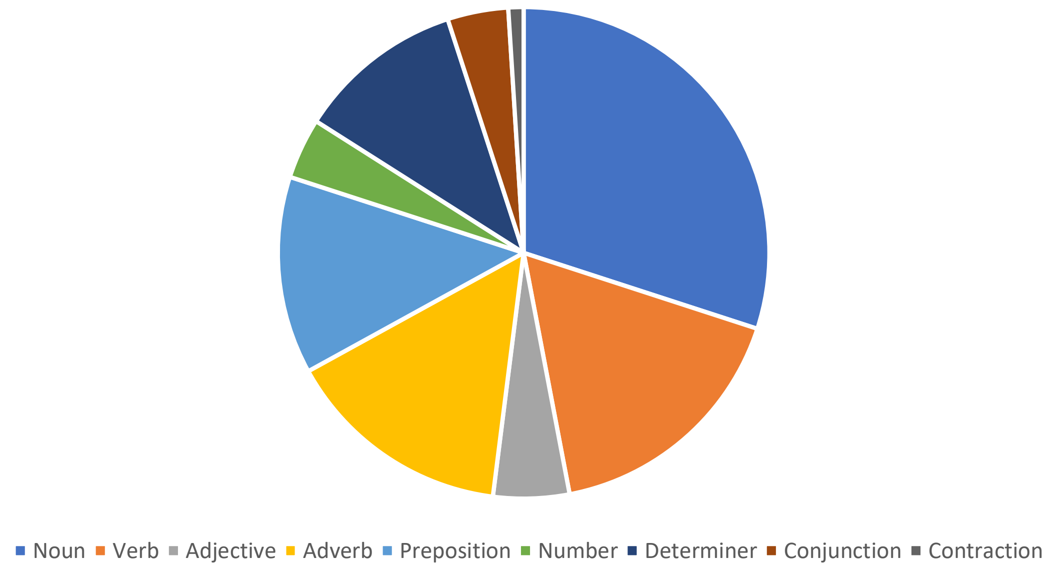
Conjunctive adverb such as *however*

There are also several other word classes besides those five ones which are ***Number*** **(***one, two, three, first*), ***Determiner (*** *the, a, this, an, his, its, more, each, their, such, these),* ***Conjunction*** **(***and, or, but, because)* and ***Contraction (****'s).*

CONCLUSIONS

Based on the results of this study, the percentages of each syntactical category found in the list of 100 HFWs in the reading section of TOEFL PBT exercises are shown in the following figure.

Figure 1.



The diagram shown in Figure 1 depicts the percentage of each syntactic category found in the data. It can be concluded that noun is the most frequent word class to appear, followed by verb, adverb, preposition, determiner, adjective, number, conjunction, and contraction.

Further elaboration regarding the word classes is necessary to enhance the vocabularies of test-takers. The vocabulary obtained can be used to make dictionaries or vocabulary enrichment teaching materials that are useful for helping test takers in the TOEFL PBT reading section. Corpus linguistics is also proven to be effective in finding the HFWs found in TOEFL PBT exercises. Hopefully, these findings can also help teachers and educators to aid their students with enough vocabularies before taking TOEFL PBT examinations.

REFERENCES

Arum, E.R. & Winarti, W. 2020. "Penggunaan Linguistik Korpus dalam Mempersiapkan Bahan Ajar English for Specific Purpose di Bidang Radiologi". *Jurnal Teras Kesehatan*, Vol.2 (2): 58-69

Chaer, A. (2015). *Morfologi Bahasa Indonesia (Pendekatan Proses)*. Jakarta: Rineka Cipta.

Crawford, W. J. & Eniko, C. (2016). *Doing Corpus Linguistics*. New York: Routledge.

Criado, R. & A. Sánchez. (2012). "Lexical Frequency, Textbooks and Learning from a Cognitive Perspective. Corpus-Based Sampel Analysis of ELT Materials". *Journal Volumen Monográficoa*: 77-94.

Desagulier, G. (2017). *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics (Quantitative Methods in the Humanities and Social Sciences).* Cham: Springer.

Jin, N.Y., Tong, C.S., Nor, M.B.M., Tarmizi, M.A.A., Mohamad, A.F.N. (2012). "Corpus Based Analysis of the TOEFL Course Books: What are the Words We Should Teach Our Students?". *International Review of Social Sciences and Humanities*. Vol 3: 152-160.

Jones, M. & Durrant, P. (2010). "What Can a Corpus Tell As about Vocabulary Teaching Materials?". *The Routledge Handbook of Corpus Linguistics*. London: Routledge: 387-400.

Kaneko, M. (2020) "Lexical Frequency Profiling of High-Stakes English Tests: Text Coverage of Cambridge First, EIKEN, GTEC, IELTS, TEAP, TOEFL, and TOEIC". *JACET Journal 64 (2020): 79–93.*

Kesuma, T. M. J. (2020). "Morfologi Bahasa Indonesia". In Module.

Kridalaksana, H. (1989). *Pembentukan Kata dalam Bahasa Indonesia*. Jakarta: PT Gramedia.

Mukundan, J. (2006). "Are There New Ways of Evaluating ELT Textbooks?". *Readings on ELT Materials II*: 170-180.

Pribakht, T.S. & Webb, S. (2016). “The Relationship between Academic Vocabulary Coverage and Scores on a Standardized English Proficiency”. *Journal of English for Academic Purposes*: 121-132.

Shaw, E. M. (2011). “Teaching Vocabulary Through Data-Driven Learning”. *Thesis*. Utah: Brigham Young University.

Sudaryanto. (2015). *Metode dan Aneka Teknik Analisis Bahasa*. Yogyakarta: Sanata Dharma University Press.

Sukarto, K. A. (2006). “Pengajaran Kosakata dan Kalimat dalam Bahasa Indonesia bagi Penutur Asing (BIPA) Tingkat Pemula (Sebuah Analisis Isi)”. *Jurnal Bahasa dan Sastra Sawomanila*, Vol. 1:2.

Sutarsyah, C., Nation, P., Kennedy, G. (1994). “How Useful is EAP Vocabulary for ESP? A corpus Based Case Study”. *RELC Journal*, 25(2): 34-50.

Verhaar, J. W. M., (1992). *Pengantar Lingguistik*. Yogyakarta: Gadjah Mada University Press.

Weisser, M. (2016). *Practical Corpus Linguistics An Introduction to Corpus-Based Language Analysis*. United Kingdom: Wiley Blackwell.