RESEARCH ARTICLE WWW.PEGEGOG.NET

Building a Corpus-Based Academic Vocabulary List of Four Languages

Mohammad Ahsanuddin*, Yusuf Hanafi, Yazid Basthomi, Febri Taufiqurrahman, Herri A. Bukhori, Joko Samodra, Utami Widiati, Primardiana H. Wijayati
Universitas Negeri Malang

Abstract

This study aims to establish and explore students' perception of a corpus in vocabulary learning. The corpus development was completed based on IDM ADDIE. This research was started by conducting a problem analysis that reveals students' obstacles in learning a language. The students' are identified to have a limited vocabulary of the language they learned. The corpus construction and development was begun by creating a script in PHP language. This research produces a corpus with 377880 tokens and five sub-corpora, namely Indonesian, English, German, Arabic, as well as art and design. The vocabularies are presented according to the highest frequency in the language and language teacher education field. The evaluation carried out by the experts of materials, language, and media discovers that the corpus is feasible to be integrated into the learning. Simultaneously, the assessment from students who have attended the corpus' implementation with data-driven learning (DDL) approach shows that this corpus helps students broaden their vocabulary, including the word meaning, form, and usage through observation on the concordance and collocation lines.

Keywords: Corpus linguistic, Data-Driven Learning, Language learning, Vocabulary.

Introduction

Vocabulary mastery does not only represent the word mere definition and form (Hiebert & Kamil, 2005; Strickland et al., 2003). It further deals with the way of associating and selecting various words to reflect an idea to be delivered (Schmitt et al., 2011). Students with limited vocabulary have a lower ability to communicate and comprehend an idea (Al-Kufaishi, 1988). In other words, students with inadequate vocabulary will face a predicament in learning a second or foreign language, especially in writing and reading using that language.

Studies have discovered a linear relationship between vocabulary mastery and reading comprehension. Research carried out by Schmitt et al. (2011) and Hu and Nation (2000) identify that the forecasted percentage of vocabulary required for a second language learner to comprehend a text is 98%. The level of vocabulary mastery can predict the number of text parts that can be understood (Hu & Nation, 2000). Vocabulary mastery has been acknowledged as an essential and fundamental element in reading comprehension (Nouri & Zerhouni, 2016; Sidek & Rahim, 2015). Simultaneously, Hinkel (2011) explains that vocabulary mastery also affects writing skills. A study carried out by Kiliç (2019) shows that vocabulary capacity carries a 26% contribution to the variance of writing performance. A greater vocabulary mastery gives students a broader knowledge to produce a well-structured text (Viera, 2017).

In Indonesia, Aziez et al. (2020) have identified that the prospective English teacher's college students only comprehend 2,800 out of 4,000 words that should be mastered to understand the text's ideas. Further, the results of research conducted by Novianti (2017) reveals that undergraduate students' receptive vocabulary mastery has not reached 2,000. It is also supported by other studies' findings that identify students' low vocabulary mastery, even far less than 2,000 (Sudarman & Chinokul, 2018). From those studies, vocabulary mastery can be concluded to be an essential issue in second or foreign language learning for Indonesian college students.

In addition, the lecturers also face a serious issue, especially in determining the vocabularies that should be taught to students. Vocabulary learning encompasses the number of words comprehended by the students and the determination of the substantial vocabularies to be mastered by the students (Youngblood & Folse, 2017). Ensuring that students learn to use the exact wording in a specific context became the authentic aspect that the lecturers should carry out. In relation to this, corpus implementation helps the lecturers to provide an original example and context-sensitive comparative learning (Carter & McCarty, 2006; O'Keeffe & McCarthy,

Corresponding Author e-mail: mohammad.ahsanuddin.fs@um.ac.id

https://orcid.org/0000-0002-3253-7721

How to cite this article: Ahsanuddin M, Hanafi Y, Yazid Basthomi, Taufiqurrahman F, Bukhori HA, Samodra J, Widiati U, Wijayati PH (2022). Building a Corpus-Based Academic Vocabulary List of Four Languages. Pegem Journal of Education and Instruction, Vol. 12, No. 1, 2022, 159-167

Source of support: Nil

Conflict of interest: None.

DOI: 10.47750/pegegog.12.01.15

Received: 29.08.2021

Accepted: 02.11.2021 **Publication:** 01.01.2022

2010). The teaching with this approach can be fundamental since corpus presents the vocabulary based on the frequency and context. This approach facilitates the lecturers to decide the vocabularies to be delivered to students (Coxhead, 2000). Meanwhile, the students can easily find the words required in a particular context (Laufer & Nation, 1995) and suitable for the actual needs (Coxhead, 2000).

The results of some studies confirm that computerized media aids students in learning vocabulary. Ma & Kelly (2006) explain in their findings that using software, students attain the vocabularies perceived to be challenging, both receptive and productive. Compared with textbook-based vocabulary learning, the use of software obtains higher effectiveness (Mouri & Rahimi, 2016). Similarly, Enayati & Gilakjani (2020) gain the same results after comparing two groups (with and without software learning). The results demonstrate that the experiment class using software attain a higher learning result and a positive attitude.

Based on the aforementioned explanation, one way to aid students in improving their vocabulary mastery is by establishing a corpus with a computerized program. This is also in accordance with the results of need analysis on the digital-based learning media for vocabulary learning that had been carried out earlier. It is found that most language program students involved as the respondents explicitly demand the development of this corpus.

Context of Study and Rationale

Our institution, Faculty of Letter of Universitas Negeri Malang, which is located in Indonesia, offers a Language and Literature study program with a bachelor of art in language education and bachelor of art in language and literature. This program prepares excellent language and literature educators and scientists. The students of this program have various educational backgrounds. Most of them come from senior high schools with only English as their foreign language subject, with no other foreign language, such as Germany and Arabic. Thus, only a small percentage of the students have been habituated to use those two languages. Further, some students taking Indonesian and English study programs have low Indonesian and English language proficiency. Additionally, our faculty also has an art and design study program. The students in this program have lower Indonesian and foreign language proficiency than students in other programs. Consequently, establishing a corpus is perceived as the proper approach in vocabulary learning for our students, who have various initial abilities in both languages, as well as art and design programs. Several studies have revealed the importance of vocabulary mastery in the introductory level of second and foreign language learning (Deni & Fahriany, 2020).

In addition to students' ability, the corpus establishment was also based on previous research that discovers a relatively low number of lecturers and students who have been familiar with corpus implementation, especially in the learning environment. Only a small-scale language education program, especially a language teacher education program, has included corpus in their curriculum (Farr, 2008). Meanwhile, corpus provides great potential for the prospective language teaches students to promote language awareness (O’Keeffe & Farr, 2003), as the fundamental skill of a language teacher (Chambers, 2005; Farr, 2008). Besides, it is also an exceptional tool to accelerate more critical and reflective practices for prospective teachers (Aşık, 2017; Farr, 2008; O’Keeffe & Farr, 2003).To facilitate students’ vocabulary development, the corpus was developed based on the thesis and dissertation article in our institution to attain a language sample expected to explain how a community used languages, once analyzed (Gardner & Davies, 2014)derived from a 120-million-word academic subcorpus of the 425-million-word Corpus of Contemporary American English (COCA; Davies 2012. This corpus can generate a list of words frequently used in final projects. Therefore, from the gathered materials, the words in this corpus are classified as academic vocabulary commonly used in students’ final projects.

The developed corpus is dynamic (Davies, 2010)genre-balanced corpus of any language, which has been designed and constructed from the ground up as a ‘monitor corpus’, and which can be used to accurately track and study recent changes in the language. The 400 million words corpus is evenly divided between spoken, fiction, popular magazines, newspapers, and academic journals. Most importantly, the genre balance stays almost exactly the same from year to year, which allows it to accurately model changes in the ‘real world’. After discussing the corpus design, we provide a number of concrete examples of how the corpus can be used to look at recent changes in English, including morphology (new suffixes -friendly and -gate following the new thesis and dissertation produced by our students. This corpus will be continuously updated to investigate the transformation of students’ language use in their final projects. The developed corpus has 377,880 tokens (a number of total words), with almost 90% of the text are an annotation vocabulary list. However, some of the words in this corpus are not classified as academic vocabulary. Some of those general words with high frequency remain to be included due to their higher frequency (Gardner & Davies, 2014)derived from a 120-million-word academic subcorpus of the 425-million-word Corpus of Contemporary American English (COCA; Davies 2012.

Method

Research Design

This corpus was developed based on Instructional Design Model (IDM) ADDIE (Branch, 2009) consisting of five iterative stages, namely analysis, design, development, implementation, and evaluation (Trust & Pektas, 2018). ADDIE is usually used in learning media development, even if it was initially a framework to develop a learning (Peterson, 2003). IDM ADDIE was selected as the framework for this corpus development wue to various reasons. First, it could be implemented in various situations and allow a natural flow among its stages (Hanafi et al., 2020). Secons, each of its stages had an evaluation phase (Trust & Pektas, 2018). Third, it led to effective, efficient, and relevant learning (Gustafson & Branch, 2002). Thus, the validity and maximum pedagogical function aspects of this corpus establishment naturally and inherently occurred.

Analysis

The analysis process was carried out to define the issues which later used as the substantial consideration to develop the corpus. The obtained data were analyzed using different techniques, such as test, observation, and interviews with lecturers and students. First of all, we conducted an analysis on students' characteristics. The obtained data from students' writing and reading comprehension assessment carried out showed students' low skills. They required a support in the form of sufficient learning resource to get a better writing and reading skills. The results of students' writing analysis show that their primary issue was in the vocabulary mastery. As confirmed by many previous studies that vocabulary knowledge is the main aspect to improve the skills in writing (Guo et al., 2013; Laufer & Nation, 1995) and reading comprehension (Hu & Nation, 2000; Laufer & Nation, 1995).

In addition, the art and design materials were also included in the corpus due to a fundamental reason. The analysis results demonstrated that the art and design students needed a special attention related to their language proficiency, primarily in English, since it is relatively lower than students in other study programs (Wang, 2017). Liu (2010) mentions that this issue is caused by art and design students' low vocabulary mastery that inhibit them to properly understand a text and write.

Secondly, the results of learning activity analysis also show that the most common implemented approach in vocabulary learning was using dictionary. Even if the dictionary usage shows an effectivity (Chen, 2012), yet the vocabulary offered in dictionaries are limited and not ready to be used by the students (Achmad, 2013). In communicative perspective, this approach can result in the missuse of the words learnt apart from its context (Kilgarriff et al., 2014). The dictionary learning is even perceived incapable of transferring information from a piece of language, such as collocation. Therefore, this learning approach should be improved to ensure the context aspect, as offered by corpus.

Thirdly, the results of learning environment analysis demonstrated that students' vocabulary learning had not involved sufficient learning technology. Some of students and lecturers explicitly hoped for the development of a corpus to aid their learning activity to easily comprehend vocabularies. From the technology availability, all students have a proper access to computers and laptops since they own it and it is also provided by the institution. Thus, it enabled the establishment of this corpus.

Those analysis results confirm that students vocabulary learning should be supported by technology based learning resources that can fullfiled by establishing a corpus to be used by the students to accelerate their vocabulary mastery (Koosha & Jafarpour, 2006; Paker & Özcan, 2017).

Design

In this stage, all of the corpus development preparations were completed. The essential aspect in this stage was considering the learning context, students' needs, and the application of developed corpus. This stage facilitated the selection of componnets and materials, as well as the software needed to develop the corpus. Besides, in this stage, the evaluation instrument was also selected to measure the developed corpus' validity.

First of all, the materials included in the corpus were the undergradate and graduate thesis, as well as dissertation articles from students in five study programs in Faculty of Letter, Universitas Negeri Malang. Those programs consisted of Indonesian, Arabic, English, and German Literature, as well as art and design study programs. The inclusion of these materials was aimed to give example of the vocabulary use in a proper context of academic writing. Besides, these materials were selected due to its accessibility as the data with no permission request since it had been owned by the institution and could be used as a research database. After the material selection, the article softfiles from the library was still in the DOC format, so that they were transformed into word processing documents. For the material preparation, they were uploaded in the PDF format. The DOC to PDF conversion was completed using conversion application, such as Nitro PDF, do PDF, and Microsoft Word 2013. The conversed document were reevaluated to ensure their accuracy.

The second corpus preparation stage was word deletion process from the corpus. If the words were not categorized as content word, then they were erased. It was completed so that the corpus could be analyzed. The example of words being deleted were the repeated words in the title, authors' name, and institution or organization names.The thisd stage was determining the software used to analyze the text. Many softwares can be used to analyze texts, but this corpus was specifically designed by a script using PHP language to decide the text frequency, concordance, and collocation (Ahsanuddin et al., 2020)fi'il (verb This script with PHP language use produced list of words based on their frequency (Rafatbakhsh & Ahmadi, 2019)one-by-one and quite incidentally; and the existing teaching materials and references for idioms are mostly intuition-based. However, a more recent approach to better teaching and learning idioms is to present them under categories of their common themes and topics. Corpus linguistics can be of much contribution through helping the design and development of more authentic and systematic materials using comprehensive corpora which are typically the best representatives of the target language. In this connection, the present study aimed at searching for the thematic index of 1506 idioms under 81 categories at the end of the Oxford Dictionary of Idioms in the largest freely available corpus, i.e. the Corpus of Contemporary American English (COCA.

Lastly, a corpus storyboard was developed, along with evaluation instruments to test the developed product's feasibility. The instrument was an expert validation sheet and questionnaire for the students using 1-5 Likert scale with criteria 1 representing a strong disagreement and criteria 5 represents strong agreement. The assessed components were from learning materials and language perspectives covering the conformity aspect of language structure analysis results and its simplicity to be comprehended, related to the frequency, vocabulary, concordance, and collocation. Meanwhile, the learning media aspect covers visual communication and software engineering.

Development

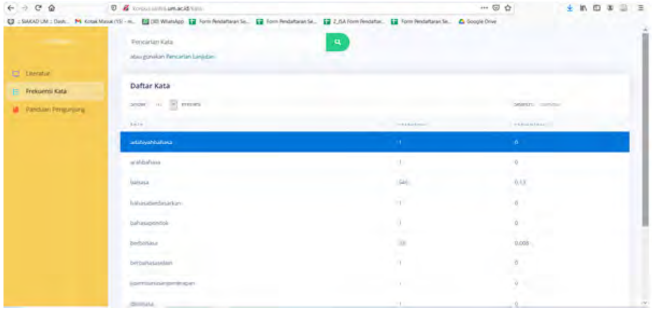
This stage involved the corpus establishment based on the development plan generated in the design phase. This stage consisted of two phases of development and evaluation. In the first phase, the required software was installed. In this particular corpus development, the server used was a Linux operation system. The server was also equipped with a web server application (Apache and PHP 7) and server database (MAriaDB version). The application was then supplied with an established program code. Lastly, the basis data adjustment was carried out so that it could be accessed through the internet network. In the second phase, the vocabulary, concordance, and collocation that facilitated students to learn based on the frequency, focusing on the word usage was developed. Consequently, it created more efficient vocabulary learning.

In creating the frequency-based words list, the files were uploaded into the server in PDF format. After they were uploaded, once the 'analysis' button is clicked, the server analyzed the uploaded documents. In the beginning, all the sentences in the documents uploaded by the program are broken down (based on the space) into a list and a number of words in every document (with an assumption that each document had 100 words). After that, the program completely checks the documents to estimate the number of each word appearance (word frequency) in each document. For instance, if it analyzed the word '*bahasa*', then the program searched the number of that word in the word list of that document. If that word is estimated to appear 546 times, then it becomes its frequency number. To find the percentage of word frequency, the total word's number of frequency (e.g. 3) is divided by the total word number in that document (e.g. 4,200 words). Thus, if calculated with the frequency formula, it becomes 546/420.000 x 100%, and the obtained occurrence rate is 0.13% (Figure 1).

The word '*bahasa*' can be used as a noun and verb in the corpus. In that case, the sentences that use the word should be seen. This corpus enables the user to access all the sentences that use that specific word (Figure 2). It is called the concordance line.

For the collocation identification process, the collocation list from each corpus category was added. Once the 'analysis' button is clicked, the program automatically searches the document's collocation. For instance, for collocation of 'negeri jiran', if it appears in the uploaded document, the program informs that the collocation is found in that document. However, if the collocation is not found in the database of uploaded documents, then the program tells that the collocation data is not found.

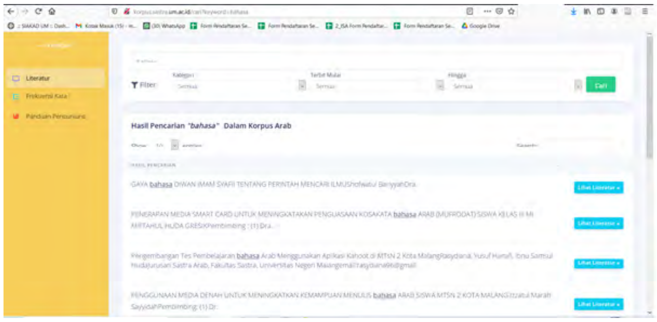
The last phase in this stage is the testing to ensure that the developed corpus is in accordance with the applied specification. The testing was carried out through an expert validation. The validation involved the experts from three fields, namely learning material, language, and media. This stage was aimed to attain evaluation, critics, and suggestions from experts as the fundamental of the improvement so that the developed corpus is feasible to be implemented in second



Daftar Kata

adabiyahbahasa	1	0
arabbahasa	1	0
bahasa	546	0.13
bahasaberdasarkan	1	0
bahasapondok	1	0
berbahasa	33	0.008
berbahasadan	1	0
[illegible]	1	0
[illegible]	1	0

Fig. 1: The word frequency in the developed corpus



Kategori: Semua
Terbit Mulai: Semua
Hingga: Semua

Hasil Pencarian "*bahasa*" Dalam Korpus Arab

GAYA bahasa DIWAN IMAM SYAFII TENTANG PERINTAH MENCARI ILMUShofwatul BariyyahDra.

PENERAPAN MEDIA SMART CARD UNTUK MENINGKATAKAN PENGUASAAN KOSAKATA bahasa ARAB (MUFRODAT) SISWA KELAS III MI MIFTAHUL HUDA GRESIKPembimbing : (1) Dra.

Pengembangan Tes Pembelajaran bahasa Arab Menggunakan Aplikasi Kahoot di MTsN 2 Kota MalangRasydiana, Yusuf Hanafi, Ibnu Samsul HudaJurusan Sastra Arab, Fakultas Sastra, Universitas Negeri Malangemail:rasydiana96@gmail

PENGGUNAAN MEDIA DENAH UNTUK MENINGKATKAN KEMAMPUAN MENULIS bahasa ARAB SISWA MTSN 2 KOTA MALANGIzzatul Marah SayyidahPembimbing: (1) Dr.

Fig. 2: Concordance on the developed corpus

or foreign language vocabulary learning at the undergraduate level in Faculty of Letter, Universitas Negeri Malang.

Implementation

The implementation stage was the teaching process using the developed corpus. This stage was completed in a class consisting of 16 prospective English teacher students, with the classical method. English program students were selected due to two main reasons. First, they have been familiar with the corpus. Second, all of the uploaded documents have the English version.

Before attending the learning process, the students were asked to install the required programs. They prepared a web browser (Google Chrome, Mozilla Firefox, Microsoft Edge). Then, they were asked to pen the corpus link to access the corpus website. In the initial learning phase, they were asked to write a most frequently used learning technique to broadening their vocabulary. Then, they followed the learning process using Data-Driven learning (DDL). In this approach, the students used a corpus with concordance as the source to find the answers to linguistic questions. This approach encourages student-centered language learning (Allan, 2009) graded reader texts can be made into a corpus appropriate for use with lower-level learners. Here I consider using such a corpus for data-driven learning (DDL. In this DDL, the linguistic questions were generated by the students, while the corpus and the lecturer acted as the informant, and facilitator, respectively (Johns, 2012).

At the end of the learning, the students were asked to fill a questionnaire consisting of two types of questions. The close-ended questions were used to score the feasibility. The open-ended question was used to reveal their learning experience using corpus, compared to their usual learning without digital corpus. Generally, this trial phase was used to obtain detailed critics and suggestions from students to improve the quality of the corpus.

Evaluation

The evaluation stage covers all of the improvements toward the developed corpus with two phases of formative and summative assessment. The formative evaluation included the evaluation from the experts' validation results. Meanwhile, the summative evaluation only covered the first level, according to the Kirkpatrick model (Kirkpatrick & Kirkpatrick, 1994), the students' opinion on vocabulary learning using the corpus.

The quantitative data were analyzed by calculating the average scores. The corpus is categorized great and feasible if the average score of each primary component of the assessment aspects attains more than four score (Akbar, 2013). Additionally, the reliability was obtained from the expert's and users' evaluation results, which was calculated using Cronbach Alpha (CA). Simultaneously, the qualitative data, in the form of statements, were processed using content analysis. The students' identity was concealed, and each of them was given a number (e.g., Student 1: S1). The content analysis was carried out following the stages brought up by Bengston (2016). The analysis was completed in the manifest level. The code was generated inductively through the repeated filtering process from the descriptive category into themes. They were then tabulated as frequency supported by the direct quotation from the students. To ensure that the coding was properly completed, in this research, it was also calculated using Cohen's Kappa, which resulted in a more than 0.75 score. To assure the finding's reliability, in the end of analysis stage, a member check was carried out by distributing the coding results to the participants to attain their consents.

FINDINGS

Results of Experts Validation

The corpus validation in the term of material and language was completed by five experts. The material and language validation sheet consists of ten items with two primary components. The first component involves five items, with indicators of the compatibility of language structure analysis results with the applicable rules and its relation with word frequency, concordance, and collocation. A score of 1-5 was used, with scale 1 representing great incompatibility and scale 5 means high suitability. Meanwhile, the second component has five items, with assessment indicators of language simplicity and ease of comprehension of the frequency, concordance, and collocation analysis results. It uses the same 1-5 score, with scale 1 covers the very hard to be understood, and scale 5 means easily understood. In this evaluation, the obtained average score of expert validation on the first and second components is 4.48 and 2.28, respectively, from a maximum of 5 (Figure 3). This result indicates that the developed corpus has an accurate language structure that can be easily comprehended. Additionally, the reliability calculation from the material and language experts evaluation results in a 0.84 score. It shows that the instruments used in this study are reliable, and the developed corpus has excellent quality from both materials and language aspects.

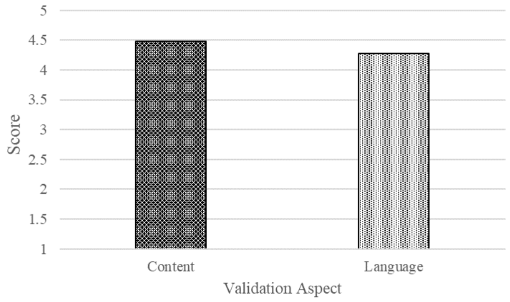


Fig. 3. The average score of materials and language experts' validation

The validation of the media aspect was carried out by five experts. The media validation sheet subsists of ten items with two main components of visual communication and software engineering. In the first component, there are five items with indicators of layout attractiveness, font types and size selection, color, and icon navigation. It uses 1-5 scores, with score 1 represents highly not attractive, while score 5 means very engaging. The second component also has five items with indicators of ease of maintenance, installation, and operation. It also applies 1-5 score, the score 1 means very difficult to be operated, and score 5 represents very easily operated. From this evaluation, the obtained media experts' validation average scores for the first and second components are 4.44 and 4.20, respectively, from the maximum score of 5 (Figure 4). These results demonstrate that the developed corpus has an engaging display and can be easily operated. The reliability estimation from the media experts attained a 0.88 score. Therefore, the instruments used in this research are great, and the established corpus has excellent quality in the media aspect.

Results of Product Implementation on Students

The questionnaire distributed to students has 12 items of three primary components, namely software engineering, visual communication, and language. Each component consists of four items. The obtained average scores for the first, second, and third components are 4.55, 4.25, and 4.60, respectively, from the maximum score of 5 (Figure 5). These results indicate that the developed corpus has an attractive design, can be easily operated and comprehended in vocabulary learning. The obtained reliability score from the students' evaluation is 0.86. consequently, the instrument used in this study is categorized as excellent. At the same time, the established corpus has an excellent quality, feasible to be implemented in vocabulary learning for students in the Faculty of Letter.

In the initial learning stage, the students were asked to mention their frequently used vocabulary learning techniques. The results show that all of the students use dictionaries. After they attended the learning using the classical method, they were asked to list the convenience and usefulness of corpus in vocabulary learning. They mention that vocabulary learning using corpus is better than the one using a dictionary. They acknowledge that corpus use is exceptionally essential to comprehend the words' form, meaning, and usage in the right context. These students suggest that the corpus should add much more literature to increase its token. Students' opinions are summarized in Table 1.

In addition, some direct excerpts from selected students' responses. The underlined words represent the description code categorized in the theme.

> *"I felt that this corpus helps me a lot, compared to my previous learning technique that uses dictionaries; it better facilitates me to improve my vocabulary mastery. Using a dictionary, I cannot easily comprehend the words' form and meaning, while this corpus helps me understand them easily. In the future, the literature should be increased so that it has a larger number of vocabulary" (S2).*
>
> *"I give a satisfactory score for the corpus' display since I think it is too plain that left a useless impression. However, once it is used, I felt its benefits, even if I need to use it more frequently to be familiar with this corpus. At least, by looking at its collocation and concordance, I can find the relation among part of speech and words' usage' (S6).*
>
> "I obtain many words from this corpus. In some words, corpus helps me better understand their meaning than a dictionary. I feel I can better understand the words' function if I see the concordance and collocation list. It sufficiently helps me to form a sentence using various forms of words I found in this corpus" (S11).
>
> "I feel that this corpus has some weaknesses. Besides, it also has many words that I don't understand the meaning so that I have to guess their meaning. Thus, I do not like this corpus. I still need help from friends and lecturer to sufficiently understand the concordance" (S12).

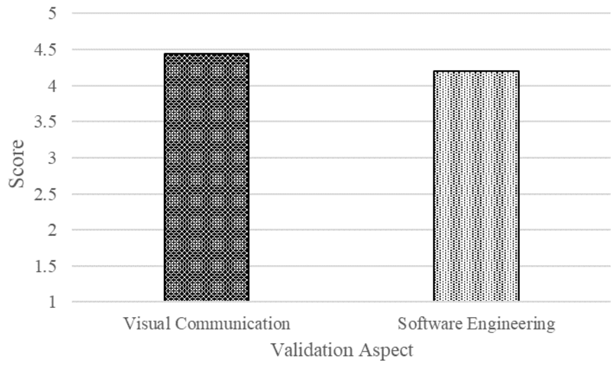


Fig. 4. The average score of media experts' validation

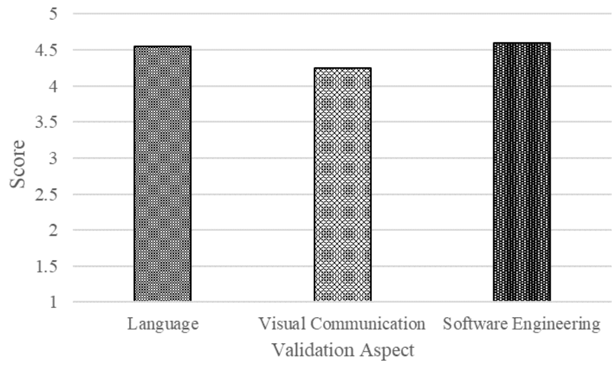


Fig. 5. The average score of students' evaluation

Table 1: Students' opinion on corpus

Category	*Theme*	*f*
Students' Opinion	Enlarge the vocabulary mastery	14
	Help to comprehend words' form	12
	Aid to understand words' meaning	12
	Facilitate to understand the words' usage	11
	Help to know new words	10
Students' Suggestion	Include more literature	12
	Make a more attractive display	6
	Reduce non annotated words	3

Discussion and Conclusion

In this research, the corpus was developed to improve students' vocabulary mastery to enhance their reading comprehension and writing skills. To investigate the attainment of this purpose, the experts' results validation was evaluated, along with the students' opinion. Through the implementation of the ADDIE model, this corpus was developed and each of its phases was evaluated. The summative evaluation results show that the students can easily use it and attains its benefits in identifying words' form, finding words' meaning, and using the words in the right context, which finally improves their vocabulary mastery, reading comprehension, and writing skills simultaneously. Generally, students demonstrate a positive attitude toward vocabulary learning using the corpus. It is consistent with some previous studies that software use aids vocabulary learning and creates students' positive attitudes (Enayati & Gilakjani, 2020; Koosha & Jafarpour, 2006; Ma & Kelly, 2006; Mouri & Rahimi, 2016).

This finding becomes fundamental that corpus can be developed using ADDIE to provide additional technology-based facilities in vocabulary learning. The presence of corpus helps lecturers to decide the words that should be learned by the students and create a more attractive education. This corpus carries benefits for the students. They can easily and infinitely access the authentic materials in the corpus online for free (Basanta & Martín, 2006). The students can further use it to comprehend a language use contextually [53] more effectively. As explained by the students, they do not need a long period to understand words' meaning and usage. These results indicate that simplification and adaptation on the corpus' concordance line are in accordance with the students' level. This success is also connected to the guidance provided by the lecturers in finding the language usage pattern while looking at the concordance line (Reppen, 2012) in a context.

This research also confirms that the implementation of the DDL approach in corpus-based vocabulary learning can be used at the university level since its principles correspond with current students' expectations (Fuster Márquez & Clavel Arroitia, 2010). This study also adds that the DDL approach can facilitate vocabulary learning (Corino & Onesti, 2019; Koosha & Jafarpour, 2006). This approach with corpus implementation has shifted students' and lecturers' roles (Paker & Özcan, 2017). With this approach, the students are conditioned to be more active in analyzing the structures in the concordance line and constructing the correct form by rewriting their sentences (Girgin, 2019). It is profoundly useful for students since they get a wider opportunity to learn the words' meaning, function, and usage in a context (Paker & Özcan, 2017), which at the end provokes awareness of language use (Krieger, 2003). Additionally, the lecturers act as the source of knowledge, guide, and facilitator (Gabrielatos, 2005). This learning by analyzing corpus and with DDL approach is reciprocal with frequently implemented language learning theories (Gavioli & Aston, 2001), namely constructivism theory and student-centered approach (Corino & Onesti, 2019).

According to the findings in this research, some theoretical and pedagogical potential implications have been discovered. First, as shown from the students' response, they have not entirely convinced by the corpus' benefits, even if the corpus usage offers some excellences in language learning. In other words, ideally, a training session should be carried out before the corpus application so that its advantages can be maximized. Thus, students can also obtain the fullest benefits. It should be noted that some of this corpus' words' meaning has not been appropriately comprehended by the students, as explained by students S12. Therefore, this corpus's concordance line can delude the students, even if this concordance has been identified to match students' level. Consequently, the lecturer has to be there to guide students to interpret the concordance accurately.

Second, the corpus implementation in this study was limited to 16 students in a relatively short period. Besides, the evaluation only involves summative assessment in the form of students' opinions. A further study has to involve many more participants and a longer period of intervention. The evaluation system can also be made in a more detailed version to measure the improvement in students' vocabulary mastery. It will provide a more transparent and satisfying conclusion on the effect of corpus implementation using DDL approach in the students' vocabulary mastery and attitude.

Third, the limitation of this developed corpus is in its imbalance token in every subcorpora. This should be avoided in future research, even if this token amount has fulfilled the corpus' requirement. Additionally, due to the specific use of vocabularies in the language and education science, future researchers have to examine the vocabulary used in those fields. The words in this corpus are presented based on their frequency, including the general words which are not explicitly related to the language and education fields. Therefore, this frequency helps the lecturers to direct students' attention to those high-frequency words. Once the students comprehend those words, they can move to the less frequency words, which are rarely used. In the future, the researchers can also maximize the frequency of the collocation identification by adding the n-gram algorithm to show the collocation estimation's statistical process.

From those limitations, this corpus development effort in this institution is still in the initial stage. However, this research results carry a significant explanation of the integration of corpus within the prospective language teachers' education curriculum (Farr, 2008; Heater & Helt, 2012; Lenko-Szymanska, 2014). This is the first corpus that incorporates four languages with art, and design subcorpora in Indonesia integrated into the prospective language teachers' education. In the end, this research has identified the benefits that can be obtained by the language program students, especially those who take the prospective language teacher program. This corpus stands as an answer to the challenge that the future prospective teacher education program should provide a more excellent opportunity to integrate corpus (Farr, 2008). This corpus also offers materials that follow students' needs in the future (O'Keeffe & Farr, 2003).

Acknowledgments

We address our most sincere gratitude to the Faculty of Letter, Universitas Negeri Malang. We also thank all experts involved in the validation process and students who have attended the trial and assessment process, while also given their opinion on the developed corpus in this study.

References

Achmad, S. (2013). Developing English Vocabulary Mastery through Meaningful Learning Approach. International Journal of Linguistics, 5(5), 75–97. https://doi.org/10.5296/ijl.v5i5.4454

Ahsanuddin, M., Ma'sum, A., & Ridwan, N. A. (2020). Investıgatıng Arabıc Corpus (KorSA) of Indonesıan Undergraduate Thesıs Abstracts. Humanities & Social Sciences Reviews, 8(3), 920–927. https://doi.org/10.18510/hssr.2020.8396

Akbar, G. (2013). Metode Pembelajaran Alquran Melalui Media Online. Indonesian Jurnal on Networking and Security (IJNS), 2(1), 65–68.

Al-Kufaishi, A. (1988). Iraq: A Vocabulary-Building Program is a Necessity Not a Luxury. English Teaching Forum, 26(2), 42.

Allan, R. (2009). Can a Graded Reader Corpus Provide "Authentic" Input? ELT Journal, 63(1), 23–32. https://doi.org/10.1093/elt/ccn011

Aşık, A. (2017). A sample corpus integration in language teacher education through coursebook evaluation. Dil ve Dilbilimi Çalışmaları Dergisi, 13(2), 728–740.

Aziez, F., Aziez, F., & Purwanto, B. E. (2020). Receptive vocabulary knowledge and reading skills of Indonesian prospective EFL teachers. Universal Journal of Educational Research, 8(5), 2005–2011. https://doi.org/10.13189/ujer.2020.080538

Basanta, C. P., & Martín, M. E. R. (2006). The Application of Data-Driven Learning to a Small-Scale Corpus of Conversational Texts from the BNC –British National Corpus: Teaching Speech Acts. The International Journal of Learning: Annual Review, 12(8), 183–192.

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. NursingPlus Open, 2, 8–14. https://doi.org/10.1016/j.npls.2016.01.001

Branch, R. M. (2009). Instructional Design: The ADDIE Approach. Springer New York Dordrecht Heidelberg. https://doi.org/10.1007/978-0-387-09506-6

Carter, R., & McCarty, M. (2006). Cambridge Grammar of English: A Comprehensive Guide. Cambridge University Press.

Chambers, A. (2005). Integrating Corpus Consultation in Language Studies. Language Learning and Technology, 9(2), 111–125.

Chen, Y. (2012). Dictionary Use and Vocabulary Learning in The Context of Reading. International Journal of Lexicography, 25(2), 216–247. https://doi.org/10.1093/ijl/ecr031

Corino, E., & Onesti, C. (2019). Data-Driven Learning: A Scaffolding Methodology for CLIL and LSP Teaching and Learning. Frontiers in Education, 4(7), 1–12. https://doi.org/10.3389/feduc.2019.00007

Coxhead, A. (2000). A New Academic Word List. TESOL Quarterly, 34(2), 213–238.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. Literary and Linguistic Computing, 25(4), 447–464. https://doi.org/10.1093/llc/fqq018

Deni, R., & Fahriany, F. (2020). Teachers' Perspective on Strategy for Teaching English Vocabulary to Young Learners. Vision: Journal for Language and Foreign Language Learning, 9(1), 48–61. https://doi.org/10.21580/vjv9i14862

Enayati, F., & Gilakjani, A. P. (2020). The impact of computer assisted language learning (CALL) on improving intermediate EFL learners' vocabulary learning. International Journal of Language Education, 4(1), 96–112. https://doi.org/10.26858/ijole.v4i2.10560

Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. Language Awareness, 17(1), 25–43. https://doi.org/10.2167/la414.0

Fuster Márquez, M., & Clavel Arroitia, B. (2010). Corpus linguistics and its aplications in higher education. Revista Alicantina de Estudios Ingleses, 23, 51. https://doi.org/10.14198/raei.2010.23.04

Gabrielatos, C. (2005). Corpora and Language Teaching: Just a Fling or Wedding Bells?. Tesl-Ej, 8(4), 1–37.

Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List. Applied Linguistics, 35(3), 305–327. https://doi.org/10.1093/applin/amt015

Gavioli, L., & Aston, G. (2001). Enriching Reality: Language Corpora in Language Pedagogy. ELT Journal, 55(3), 238–246. https://doi.org/10.1093/elt/55.3.238

Girgin, U. (2019). The Effectiveness of Using Corpus-based Activities on The Learning of Some Phrasal-prepositional Verbs. TOJET: The Turkish Online Journal of Educational Technology, 18(1), 118–125. http://files.eric.ed.gov/ fulltext/EJ1201799.pdf

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting Human Judgments of Essay Quality in Both Integrated and Independent Second Language Writing Samples: A Comparison Study. Assessing Writing, 18(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Gustafson, K. L., & Branch, R. M. (2002). Survey of Instructional Development Models (4th ed.). Eric Publications.

Hanafi, Y., Murtadho, N. M., Ikhsan, A., & Diyana, T. N. (2020). Reinforcing Public University Student's Worship Education by Developing and Implementing Mobile-learning Management System in The ADDIE Instructional Design Model. International Journal of Interactive Mobile Technologies, 14(2), 215–241. https://doi.org/10.3991/ijim.v14i02.11380

Heater, J., & Helt, M. (2012). Evaluating Corpus Literacy Training for Pre-Service Language Teachers: Six Case Studies. Journal of Technology and Teacher Education, 20(4), 415–440. https://www.learntechlib.org/primary/p/39324/

Hiebert, E. H., & Kamil, M. L. (2005). Teaching and Learning Vocabulary: Bringing Research to Practice. In Teaching and Learning Vocabulary: Bringing Research to Practice (Issue August). Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9781410612922

Hinkel, E. (2011). What research on second language writing Tells Us and what it doesn't. Handbook of Research in Second Language Teaching and Learning, 2, 523–538. https://doi.org/10.4324/9780203836507.ch32

Hu, H.-C., & Nation, P. (2000). Unknown Vocabulary Density and Reading Comprehension. Reading in a Foreign Language, 13(1), 403–430.

Johns, T. (2012). From Printout to Handout: Grammar and Vocabulary Teaching in The Context of Data-driven Learning. Perspectives on Pedagogical Grammar, 4(July 1990), 293–313. https://doi.org/10.1017/cbo9781139524605.014

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Johansson Kokkinakis, S., Lew, R., Sharoff, S., Vadlapudi, R., & Volodina, E. (2014). Corpus-based Vocabulary Lists for Language Learners for Nine Languages. Language Resources and Evaluation, 48(1), 121–163. https://doi.org/10.1007/s10579-013-9251-2

Kiliç, M. (2019). Vocabulary knowledge as a predictor of performance in writing and speaking: A case of turkish efl learners. Pasaa, 57(March), 133–164.

Kirkpatrick, D., & Kirkpatrick, J. (1994). Evaluating Training Programs: The Four Levels (1st Edition). Berrett-Koehler Publishers.

Koosha, M., & Jafarpour, A. A. (2006). Data-driven Learning and Teaching collocation of prepositions: The Case of Iranian EFL Adult Learners. The ASIAN EFL Jorunal, 8(4), 192–209.

Krieger, D. (2003). Corpus Linguistics: What It Is and How It Can Be Applied to Teaching. The Internet TESL Journal, 9(3).

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. Applied Linguistics, 16(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Lenko-Szymanska, A. (2014). Is This Enough? A Qualitative Evaluation of the Effectiveness of a Teacher-Training Course on the Use of Corpora in Language Education. ReCALL, 26.

Liu, J. T. (2010). Analysis on the Status Quo and Countermeasures of Art Majors' English Studying. Hundred Schools in Arts, 117(8), 429–432.

Ma, Q., & Kelly, P. (2006). Computer Assisted Vocabulary Learning: Design and Evaluation. Computer Assisted Language Learning, 19(1), 15–45. https://doi.org/10.1080/09588220600803998

Mouri, S., & Rahimi, A. (2016). The impact of computer-assisted language learning on Iranian EFL students' vocabulary learning. Global Journal of Foreign Language Teaching, 06(4), 210–217.

Nouri, N., & Zerhouni, B. (2016). The relationship between vocabulary knowledge and reading comprehension among Moroccan EFL learners. IOSR Journal of Humanities And Social Science (IOSR-JHSS, 21(10), 19–26. https://doi.org/10.9790/0837-2110051926

Novianti, R. R. (2017). A study of Indonesian university students' vocabulary mastery with vocabulary level test. Global Journal of Foreign Language Teaching, 6(4), 187–195. https://doi.org/10.18844/gjflt.v6i4.2685

O'Keeffe, A., & Farr, F. (2003). Using Language Corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications. TESOL Quarterly, 37(3), 389. https://doi.org/10.2307/3588397

O'Keeffe, A., & McCarthy, M. (2010). The Routledge handbook of corpus linguistics. Routledge.

Paker, T., & Özcan, Y. E. (2017). The Effectiveness of Using Corpus-Based Materials in Vocabulary Teaching. Online Submission, 5(1), 62–81.

Peterson, C. (2003). Bringing ADDIE to life: instructional design at its best - learning & technology library (LearnTechLib). Journal of Educatioanal Multimedia and Hypermedia, 12(3), 227–241. http://www.learntechlib.org/p/2074/

Rafatbakhsh, E., & Ahmadi, A. (2019). A Thematic Corpus-based Study of Idioms in The Corpus of Contemporary American English. Asian-Pacific Journal of Second and Foreign Language Education, 4(1), 1–21. https://doi.org/10.1186/s40862-019-0076-4

Reppen, R. (2012). Using Corpora in the Language Classroom (Cambridge Language Education) 1st Edition.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. Modern Language Journal, 95(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Sidek, H. M., & Rahim, H. A. (2015). The Role of Vocabulary Knowledge in Reading Comprehension: A Cross-Linguistic Study. Procedia - Social and Behavioral Sciences, 197, 50–56. https://doi.org/10.1016/j.sbspro.2015.07.046

Strickland, D. S., Galda, L., & Cullinan, B. E. (2003). Language arts : learning and teaching. Wadsworth Pub Co.

Sudarman, S., & Chinokul, S. (2018). the English Vocabulary Size and Level of English Department Students At Kutai Kartanegara University. ETERNAL (English, Teaching, Learning and Research Journal), 4(1), 1–15. https://doi.org/10.24252/eternal.v41.2018.a1

Trust, T., & Pektas, E. (2018). Using the ADDIE Model and Universal Design for Learning Principles to Develop an Open Online Course for Teacher Professional Development. Journal of Digital Learning in Teacher Education, 34(4), 219–233. https://doi.org/10.1080/21532974.2018.1494521

Viera, R. T. (2017). Vocabulary knowledge in the production of written texts : a case study on EFL language learners. Revista Technologica ESPOL - RTE, 30(3), 89–105.

Wang, P. (2017). A Corpus-based Study of English Vocabulary in Art Research Articles. Journal of Arts and Humanities, 6(8), 47. https://doi.org/10.18533/journal.v6i8.1255

Youngblood, A. M., & Folse, K. S. (2017). Survey of Corpus-Based Vocabulary Lists for TESOL Classes 1. MEXTESOL Journal, 41(1), 1–15.