The 4th Proceeding International Conference on Arabic Language and Literature (ICALL) 2021
P-ISSN: 2809-364X | E-ISSN: 2808-8425
http://proceedings2.upi.edu/index.php/ical/index
Published by: Study Program of Arabic Language Education, Faculty of Language Education and Literature, The Education University.

The Corpus of Language, Literature, and Arts: Design and Construction

Mohammad Ahsanuddin*, Yazid Basthomi, Yusuf Hanafi, Edy Hidayat, Febri Taufiqurrohman, Muhammad Nurwiseso Wibisono
Universitas Negeri Malang, Indonesia
E-mail: mohammad.ahsanuddin.fs@um.ac.id

Abstract

This study examined the design and construction of the corpus of language, literature, and art at the Faculty of Letters, Universitas Negeri Malang. The materials compiled were theses from the Department of Indonesian Literature, English Literature, Arabic Literature, German Literature, and Art and Design. The study aimed to (1) design and develop the language, literature, and art corpus of the Faculty of Letters, Universitas Negeri Malang and (2) describe the results of the feasibility test for the corpus of language, literature, and art of the Faculty of Letters, Universitas Negeri Malang. This study employed a Research and Development design. The development model used in this study was the ADDIE (Analysis, Design, Development, Implementation, and Evaluations) research model developed by Reiser and Molenda. The research stages carried out based on the ADDIE research model are Analysis Stage, Design Stage, Development Stage, Trial Stage, and Evaluation Stage. The results showed several stages in designing and developing the language, literature, and art corpus of the Faculty of Letters, Universitas Negeri Malang, namely analysis, design, development, implementation, and evaluation. The feasibility test for the language, literature, and art corpus of the Faculty of Letters, Universitas Negeri Malang, showed valid results with a value of 91%. This means that the web corpus developed is feasible to use.

Keywords: Art, Concordance, Corpus, Language concordance

Introduction

The term corpus describes a collection of written or spoken documents stored and processed on a computer for linguistic investigation. In detail, a corpus can be used in Teaching or Learning, Statistics, Dialectology, to Historical Linguistics.

Documents used as material for the corpus are undergraduate theses, theses, dissertations, papers, and research reports in writing. It can also be in the form of utterances such as expressions of lecturers when teaching, students when practicing speaking, dancing, painting, and so on. All of that can be used as a data source for language studies.

Abstract collections have now been digitized with various versions, formats, and models as a data source for language studies. The conversion or adjustment into digital form is made by utilizing information technology and computers.

The standard format is adequate for some purposes. However, for corpus linguistic research with special applications for corpus processing, a special method is needed to create a corpus the method is called a parallel corpus. Parallel corpus means a corpus of more than one language it can be two languages or even more. Generally, the parallel corpus format is created in translation research and contrastive analysis in two or more languages. So far, this research will be the first in Indonesia. This is motivated by the lack of introduction to the studies and linguistic products of the Arabic corpus among Arabic language circles in Indonesia. The study of the Arabic of Al-Qur'an in general still uses conventional

approaches, and the use of digital models is still limited. Therefore, we hope that this research will become a breakthrough in Arabic studies in Indonesia in the form of a special methodology in preparing a bilingual parallel corpus, especially the Al-Quran model and its translation in Indonesian.

A previous study underlie this research was conducted by Sasongko (2010) entitled "*Aplikasi untuk Membangun Corpus dari Data Hasil Crawling dengan Berbagai Format Data Secara Otomatis*". The result of this research is that the application can display the search results of documents and sort them based on the order of discovery of the searched data file, in the sense that the data document found first will be placed on top, while the data document found last will be placed at the bottom. The application can also convert text documents with various data formats into txt text documents and convert all image file formats into bmp format. Conversion is done to equalize the form to make storage in the database easier.

Another study by Ahsanuddin (2018) entitled "*Tashmim Al-mudawwanah Al-Mutawaziyah Li mustakhlash Al-buhuts Al-Ilmiah Al-Indunisiya Al-Arabiyah 'Ala Dhaui Nadzariyah Mona Baker Li Al-Takafu' Al-Lughawi Fi Al-Tarjamah.*" The study has confirmed the following results. *First*, it has resulted in a parallel corpus of Indonesian-Arabic research abstracts. Before making a parallel corpus, the following must be done: (a) equivalence analysis and (b) parallel corpus design. The equivalence analysis of the translated abstract used Mona Baker's perspective with three levels, namely word level, grammatical level, and text level. The word-level equivalence is the translation of the abstract and keywords. The grammatical-level equivalence includes numbers (*adad*), pronouns (*dhomir*), personal (*syakhsyiah*), and verbs (*af'al*). The types of *adad* in the dissertation abstract consist of *mutsana-mutsana, mufrad-jama', jama'-mufrad,* and *jama'-jama'. Dhomir* in the abstract is a third-person pronoun (*dhomir ghiyab*). The text-level equivalence includes references (*ihalah*), conjunctions (*adawat rabth*), and cohesion (*al-ittisaq al-mu'jami*). The references found in the abstract are persona and gesture, while the most widely used conjunctions are *wawu,* and the cohesion is repetition (*tikrar*). To determine the parallel corpus, we examine several previous studies related to the corpus. After the preliminary study was conducted, we designed the parallel corpus using the Dreamweaver program with the PHP (Hypertext Preprocessor) language. The design contains a corpus and functions as a search engine, vocabulary list, concordance, and word frequency. When expert validation was carried out, we got a score of 85%, which means the parallel corpus was valid to use. *Second*, it has resulted in the use of a parallel corpus of Indonesian-Arabic research abstracts. The parallel corpus functions as a search engine, vocabulary list, concordance, and word frequency. *Third*, it has resulted in a parallel corpus user satisfaction. This corpus is effective for searching for previous research and as translation material.

A similar study was also conducted by Ahsanuddin, Ma'sum, and Ridwan (2020) with the title "*Investigating Arabic Corpus (Korsa) Of Indonesian Undergraduate Thesis Abstracts.*" The results show that the researchers go through certain stages to produce a web corpus: (1) reviewing various literature or reference books related to the linguistic corpus, (2) collecting thesis from Universitas Negeri Malang and UIN Maulana Malik Ibrahim Malang, (3) designing and developing the corpus web using the PHP program, (4) inserting thesis material (abstract) into the web corpus, and (5) product implementation. The product of this research is an online corpus web named KorSA (*Korpus Skripsi Berbahasa* Arab).

Method

This study was field research using Research and Development method. Sugiyono (2009:297) states that Research and Development are used to produce certain products and test the effectiveness and efficiency of products.

This research produces a web corpus product through certain stages tested for validation, effectiveness, and efficiency against needs.

The development model used in this study is the ADDIE (Analysis, Design, Development, Implementation, and Evaluations) research model developed by Reiser and Molenda (Molenda, 2018). We chose this research method because it is considered more rational and suitable for our research needs. In addition, this model can be used for various development such as models, learning strategies, learning methods, media, and teaching materials. The research stages based on the ADDIE research model are Analysis, Design, Development, Trial, and Evaluation. The five stages can be seen in Figure 4.1.

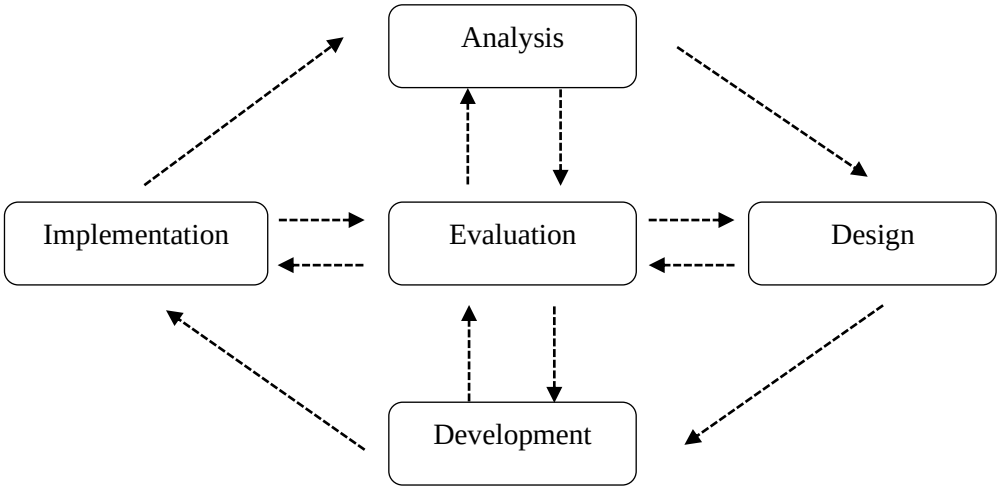


Figure 4.1 ADDIE Development Model According to Reiser and Molenda

1. Research Procedure

The following stages were done based on the ADDIE model:

a. Analysis

The first step was analyzing the needs for new learning media and the feasibility and requirements for the development.

A corpus is a collection of data for various purposes, especially language analysis. Based on our observations, there was still no corpus of language, literature, and art, especially in the Faculty of Letters, Universitas Negeri Malang. Another reason was that students often write theses and dissertations only to change the research object because of the lack of a database in writing scientific papers.

b. Design

In designing the corpus, we first illustrated the main model to develop. The activities in the design stage included: (a) designing the corpus and (b) making the corpus, including the main menu, word frequency, concordance, and collocation.

We also arranged a validation instrument to be used throughout the research process to assess the product developed so the corpus developed would be valid. The instrument included a validation questionnaire for material experts, media experts, and the use of the developed product. The instrument was further validated to obtain valid instrument results.

c. Development

We made the product designed in the previous stage by considering its feasibility. The development was carried out using the PHP language. The menus made were: (a) the main web page, (b) the word frequency, (c) concordance, and (d) collocation.

d. Implementation

At this stage the developed design and method were implemented to users. During the implementation phase, the product was applied to the actual conditions. We implemented a corpus web for a team of eight IMLA Indonesia's Arabic corpus. Team eight team consisted of Universitas Negeri Malang, UIN Maulana Malik Ibrahim Malang, UIN Syarif Hidayatullah Jakarta, Universitas Al-Azhar Indonesia Jakarta, Universitas Darussalam Gontor, Universitas Negeri Surakarta, Universitas Muhammadiyah Yogyakarta, and Pondok Modern Tazakkah Batang.

The team of eight IMLA Indonesia's Arabic corpus members used the corpus web and provided input related to its development.

e. Evaluation

After implementation, the product was evaluated. The evaluation was done after completing each step in the ADDIE model research and development procedure. The evaluation was also done after the end of activities to measure the final results of the developed product. The results of the evaluation would be used as feedback to the users of the developed corpus.

At this stage, we also made a final revision according to the analysis results or needs that were not met. Product revision was carried out based on the results of the validation questionnaire and the response to the product. This step was done to ensure product validity and feasibility.

2. Product Trials

Product trials were carried out through data collection used to determine the feasibility and validity levels of the developed product. This stage included (1) trial design, (2) research objects and subjects, (3) data collection techniques, (4) data collection instruments, and (5) data analysis techniques.

1) Trial Design

This study involved two trial types:

a. Expert judgment

This trial was conducted to strengthen and review the initial product and obtain suggestions for improvement. This expert trial was done by corpus experts.

b. Field trial

This trial aimed to determine the user's response to the product being developed. This trial was carried out by distributing questionnaires (with alternatives developed using a Likert scale) and addressed to a team of eight IMLA Indonesia's Arabic corpus.

The test flow of this product is presented in Figure 4.2.

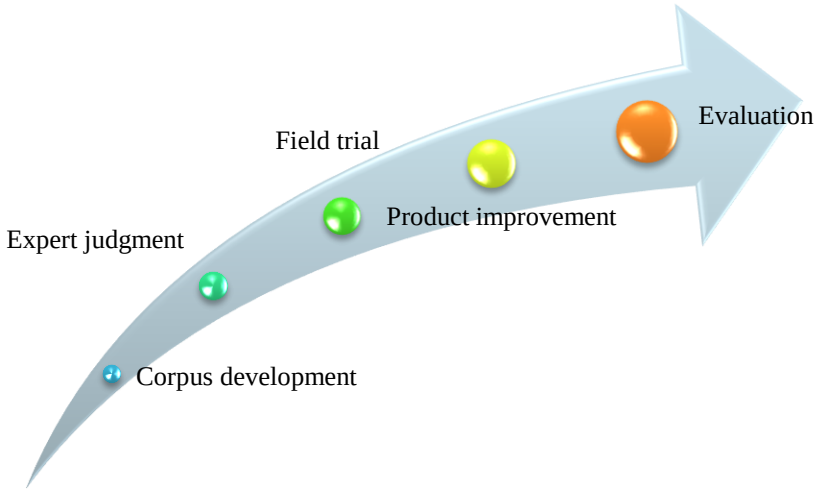


Figure 4.2. Steps in Product Trial

3. Research Object and Subject

The research object investigated was the web corpus of language, literature, and art used for various research purposes and linguistic analysis. The research subjects or targets in this study were students and lecturers.

4. Data Collection Technique

Data collection methods are the methods used to obtain accurate and accountable data as needed. The data collection methods used in this study are as follows:

a. Interviews

Interviews were conducted to find out the advantages and disadvantages of the developed corpus web. Interviews were conducted directly between validators and corpus developers.

b. Literature review

This study aimed to collect research findings and other information related to the planned product development through books, articles, journals, and relevant studies as follows:

1) Books: linguistics, corpus, language learning, and so on
2) Article: linguistic corpus.

5. Questionnaire Distribution

Questionnaires were used for the validation stage of instruments and corpus products. The questionnaire was a non-test questionnaire, in which the answers had been provided and arranged in a checklist (√). This questionnaire was provided for media experts and corpus users.

6. Research Instruments

This study used a questionnaire for media experts and respondents (product users) to find their responses. The questionnaire included:

1. Interview guide

Before interviews, we prepared a list of questions in advance as a reference. This interview guide helped us to ask questions for a more structured interview. The interview guide also prevented questions that were out of the research context.

2. Product validation questionnaire

This questionnaire was used to get an assessment from media experts regarding the suitability and display of the product. The media expert as the validator was a corpus expert

from Universitas Indonesia. The results of this questionnaire were validation scores of the product in terms of language, literature, and art corpus web products.

Two types of data were collected using this questionnaire, quantitative and qualitative data. We used the questionnaire by placing a checkmark in the column provided. This checkmark dealt with the suitability of the indicator with the product and was represented in the appropriate scale. The qualitative data were obtained from the comment and suggestion column at the bottom after the indicators.

3. Product user questionnaire

This questionnaire aimed to see the product user's response to the developed product by ticking the column provided in the google form. The thick mark helped to see the suitability of the indicators listed. The thick mark was given based on a predetermined score.

This questionnaire also collected two types of data, quantitative and qualitative data. Quantitative data came from the score of each indicator, while qualitative data came from the comment column at the bottom of the questionnaire. The questionnaire used the Likert scale interpreted as follows.

Table 4.1. The Likert Scale Used

SCORE	DESCRIPTION
4	Very interesting/very clear/very good/very easy/very accurate
3	Interesting/clear/good/easy/accurate
2	Quite interesting/quite clear/quite good/quite easy/quite accurate
1	Less interesting/less clear/less good/less easy/less accurate

7. Data Analysis

Data were analyzed qualitatively and quantitatively.

a. Qualitative data

Qualitative data came from interviews with validators and validation results from media experts in the form of comments and suggestions for product improvement. The analysis to obtain qualitative data involved data identification, data grouping, data display, and conclusion drawing.

b. Quantitative data

Data were analyzed by changing them into a percentage. This percentage was used to process the results of the expert judgment and field trials. The analysis employed the following formula:

$$P = \frac{\Sigma x}{\Sigma xi} \times 100\%$$

P = Validation results

Σ x = Total score from validators

Σ xi = Maximum total score

After the percentage for all indicators was obtained, the validation results were described, and a decision was made for each indicator based on Arikunto's table (2002).

Table 4.2 Percentage and Validation Criteria

Percentage (%)	Validation Criteria
76-100	Valid
56-75	QuiteValid
40-55	Less Valid (Revision)
0-39	Not Valid (Revision)

Results

1. Design and develop the language, literature, and art corpus of the Faculty of Letters, Universitas Negeri Malang

At this stage, the product development process for audiovisual comics follows the ADDIE (Analysis, Design, Development, Implementation, and Evaluation) method. The stages are:

a. Analysis

(1) Content Analysis

Corpus means a collection of texts. According to Baker (2010), a corpus is a collection of written and spoken texts stored on a computer. Baker defines the corpus in electronic media only. According to Setiawan (2017), a corpus is a collection of writings written by someone in hard copy and soft copy. Corpus in hard copy can be exemplified in books, magazines, dictionaries, and newspapers. Examples of the soft copy are applications, websites, online dictionaries, and so on.

Text is a sequence of sentences and paragraphs that form a complete series. Text mining is an automated system for text exploration as a substitute or supporting device for reading text (JISC, 2006). Data from text mining is a collection of text. Text collections are grouped into four types: archives, electronic text libraries, corpus, and subcorpus (Atkins, 1991).

From these opinions, we concluded that all texts could be compiled, including undergraduate theses, theses, dissertations, and articles. The texts are collected and put together, and then made into a corpus. Based on our observations, the Faculty of Letters has not created a special database for these data.

(2) Technology Analysis

Websites or sites are internet applications and services designed to convey information accurately and widely, including the corpus.

Web scraping is the process of retrieving a semi-structured document from the internet, generally in the form of web pages in a markup language such as HTML or XHTML, and analyzing the document to retrieve certain data from the page for other purposes. Web scraping is often known as screen scraping. Web scraping cannot be included in data mining because data mining implies an effort to understand the semantic patterns or trends of the large amount of data obtained. Web scraping applications (also called intelligent, automated, or autonomous agents) only focus on obtaining data through data retrieval and extraction of various data sizes. Web scrapping is related to web indexing as a universal technique used by almost all search engines. The difference is that web scraping focuses more

on transforming an unstructured web, generally in HTML format, into a structured data format that can be stored and analyzed in a database or worksheet.

b. Design

The web design for this corpus of language, literature, and art uses the PHP program. The developed menu is a concordance, word frequency, and collocation. The access rights for the corpus web consist of corpus administrators and corpus category managers. The corpus administrator has the right to search for data on the corpus, view corpus reports, appoint managers of corpus categories, input corpus primary data, and update personal data.

The corpus flow based on access rights is shown in the following figure:

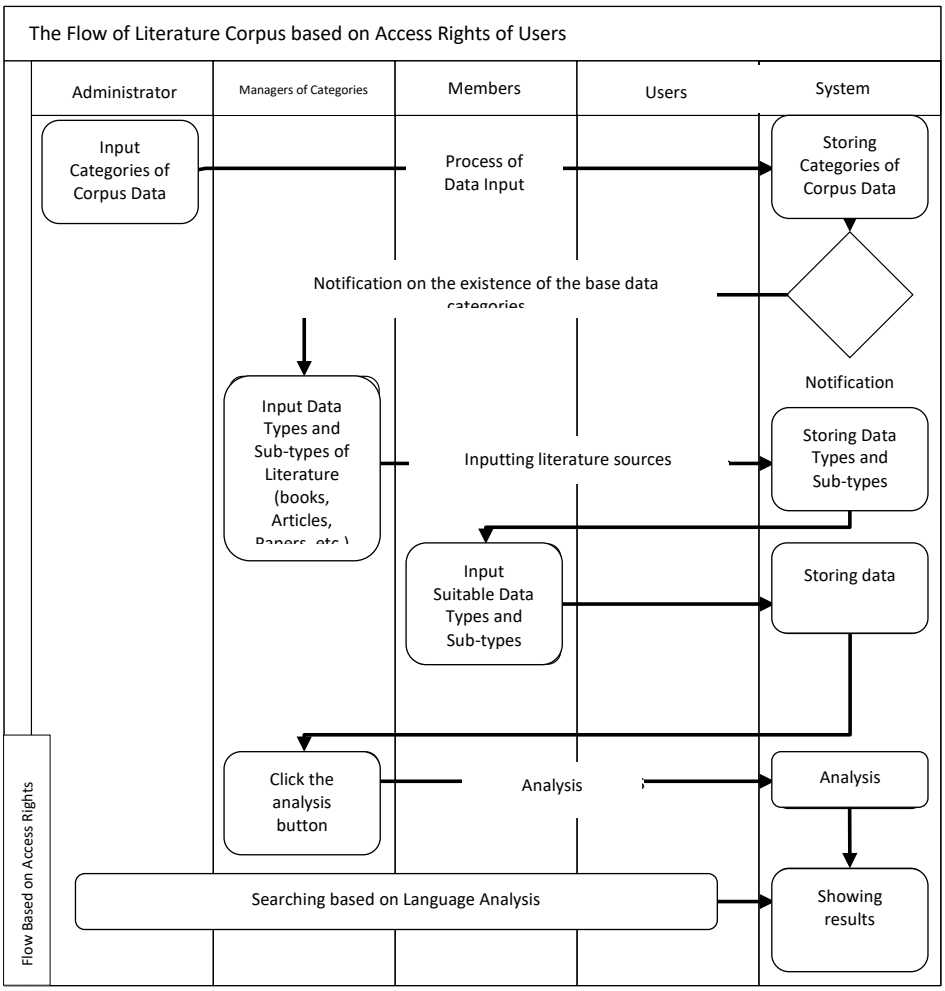


Figure 5.1 The Corpus Flow

(1) Concordance

Concordance is a collection of occurrences of word forms in their respective textual environments. The simplest form of concordance is an index. Each word formation is indexed, and the reference refers to a place in a text.

Sinclair's definition is important because it reminds us that, originally, a concordance was a manually prepared list of words found in a text or collection of texts along with their references and locations in the text. In computer-assisted linguistic analysis, concordance is still an index, but it can be generated for new purposes and across various texts and constantly evolving texts. Today's computers can be used to create concordance in the blink

of an eye and support a wider range of analytic purposes without suffering from the limited space of concordance in earlier times.

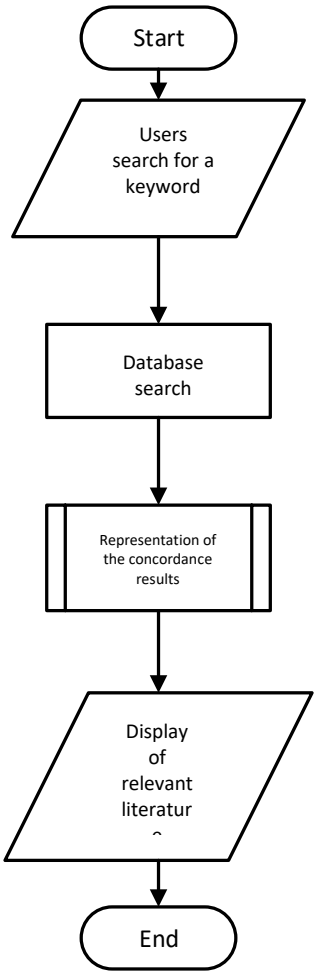


Figure 5.2 The Flow Chart Diagram to Count for Concordance

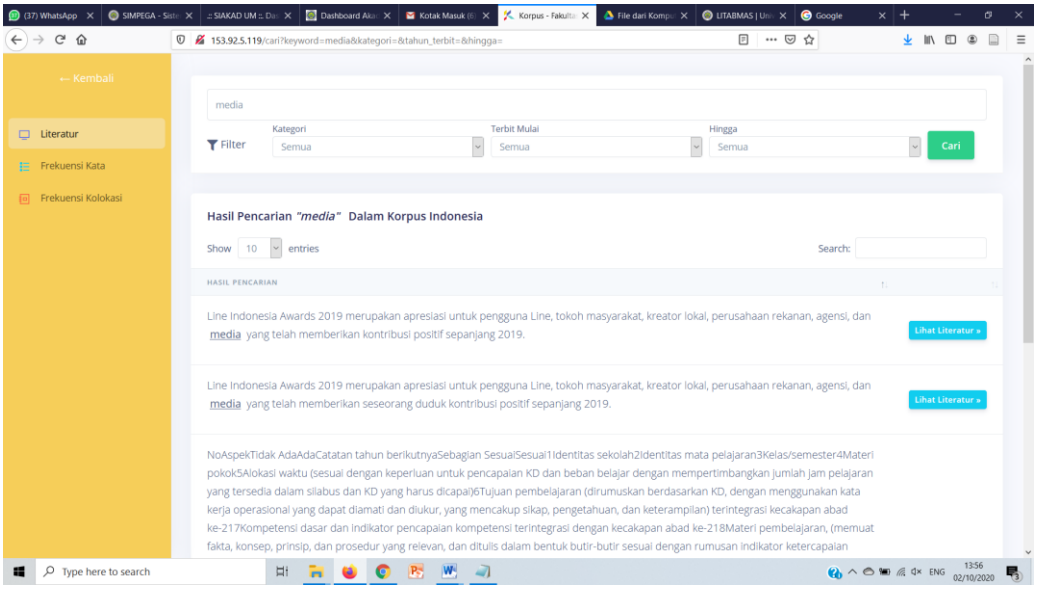


Figure 5.3 Concordance Display

(2) Word Frequency

Word Frequency is calculated based on the occurrence of the word in the corpus. The words in the corpus are then sorted by frequency.

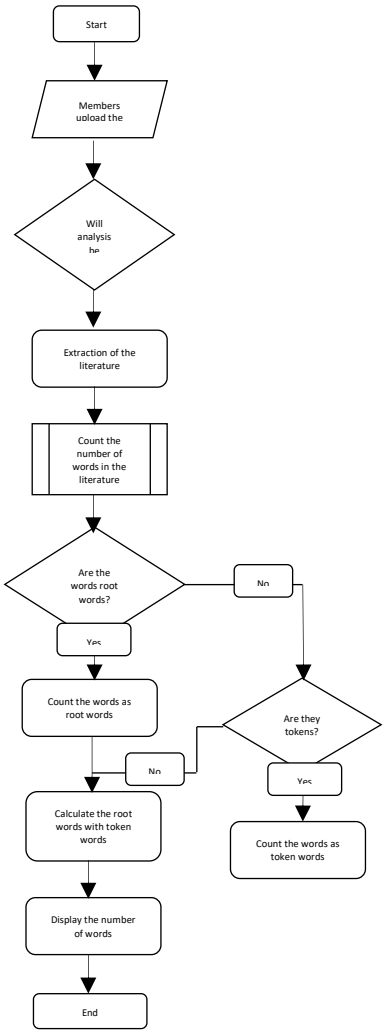


Figure 5.4 The Flow Chart Diagram to Count for Word Frequency

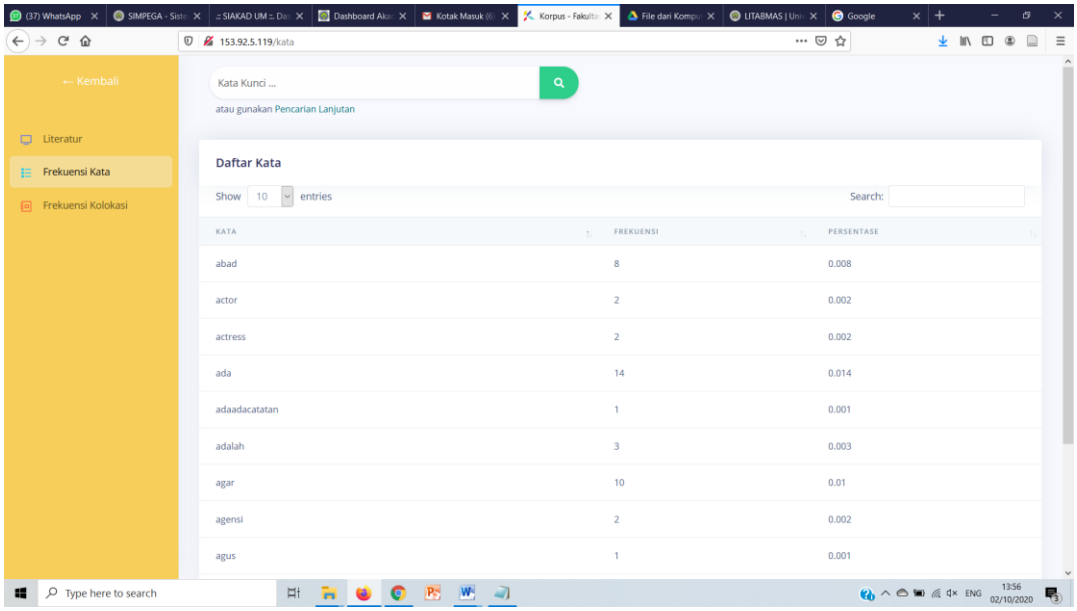


Figure 5.5 Word Frequency Display

(3) Collocations

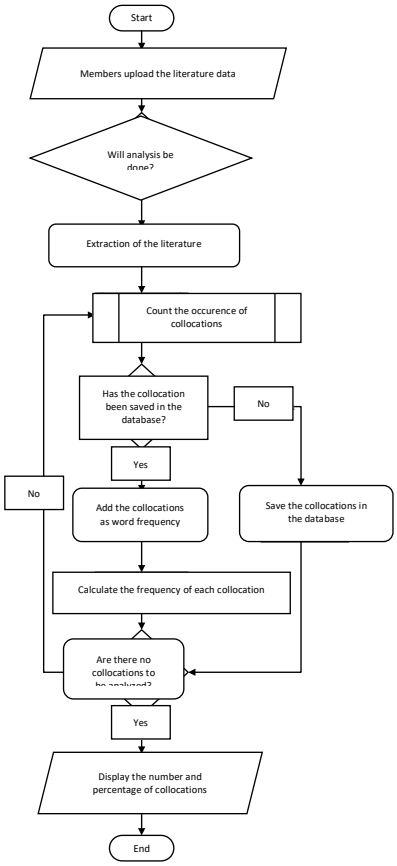


Figure 5.6 The Flow Chart Diagram to Count for Collocations

c. Development

After the script was created, it was displayed in the form of a website. Here is the explanation:

(1) Main page

To go to the main page, click korpus.sastra.um.ac.id, and it will look like this:

Figure 5.7 Corpus Main Page

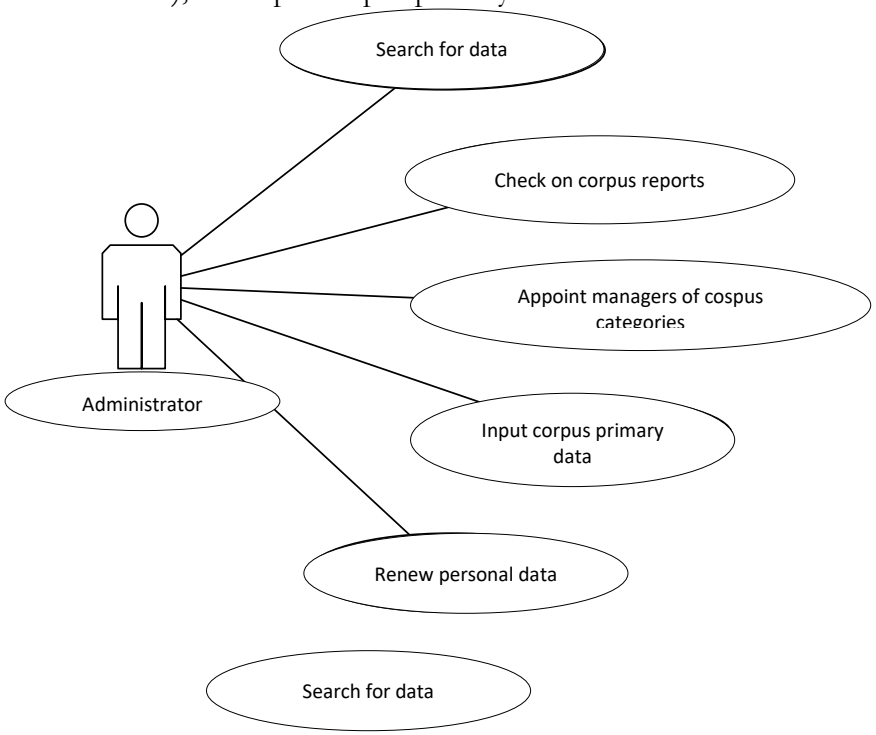
The main page is the initial display menu for the corpus and the menu used by website visitors to enter the corpus application. If users want to visit **Indonesian Literature, click Indonesian. If they want to visit German Literature, click German, and so on**.

To find out the purpose of this corpus, users can click on the "About" menu, and a description will appear: "the corpus serves as a database for linguistic analysis." The inputted data are undergraduate theses, theses, dissertations, articles, papers, and books.

The main page also presents information on the Development Team in the "Member" menu, so users know the corpus developers.

Figure 5.8 The Corpus Development Team

The "Login" menu is used for inputting data. The right of the corpus category manager is to search for data on the corpus, view the corpus report, input general corpus data (categories and members), and input corpus primary data.



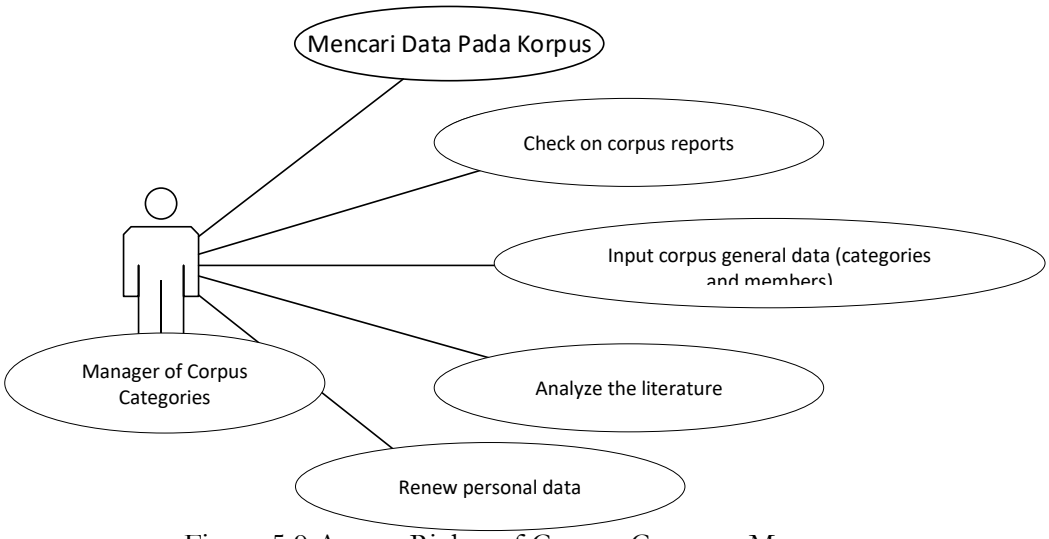


Figure 5.9 Access Rights of Corpus Category Manager

(2) Concordance

Concordance is a collection of occurrences of word forms in their respective textual environments. The simplest form of concordance is an index. Each word formation is indexed, and the reference refers to a certain place in a text.

Sinclair's definition is important because it reminds us that, originally, a concordance was a manually prepared list of words found in a text or collection of texts along with their references and locations in the text. In computer-assisted linguistic analysis, concordance is still an index, but it can be generated for new purposes and across various texts and constantly evolving texts. Today's computers can be used to create concordance in the blink of an eye and support a wider range of analytic purposes without suffering from the limited space of concordance in earlier times.

Before the era of digitizing text with modern computers as it is today, concordance was made by dedicated individuals or teams over a long time. Team members read the text, identified important words for analysis, and built a painstaking table that allowed one to record where each word example was found.

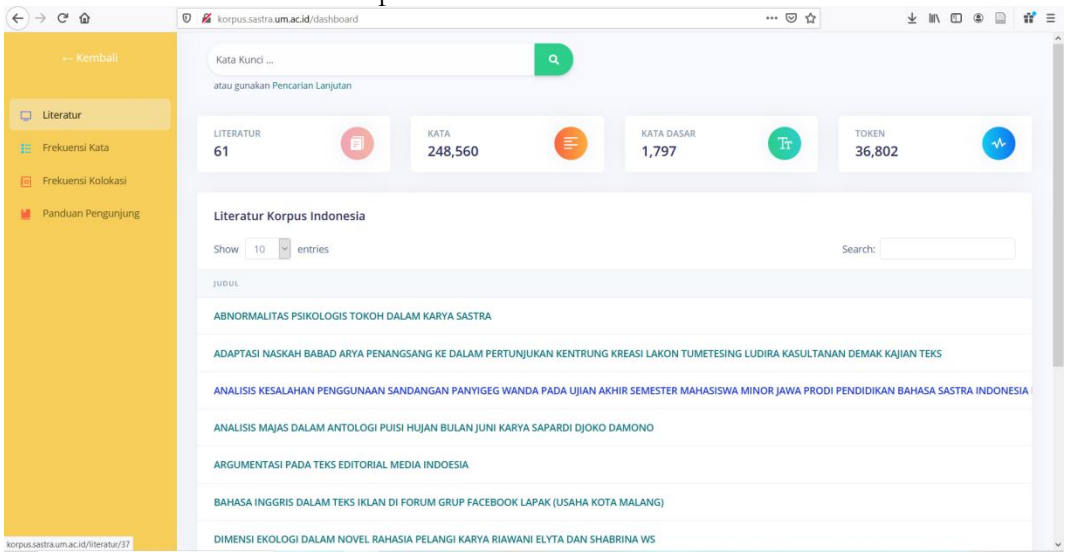


Figure 5.10 The Display of Corpus of the Indonesian Department

An example of concordance in the corpus web is "*Bahasa*", so the display will be as follows.

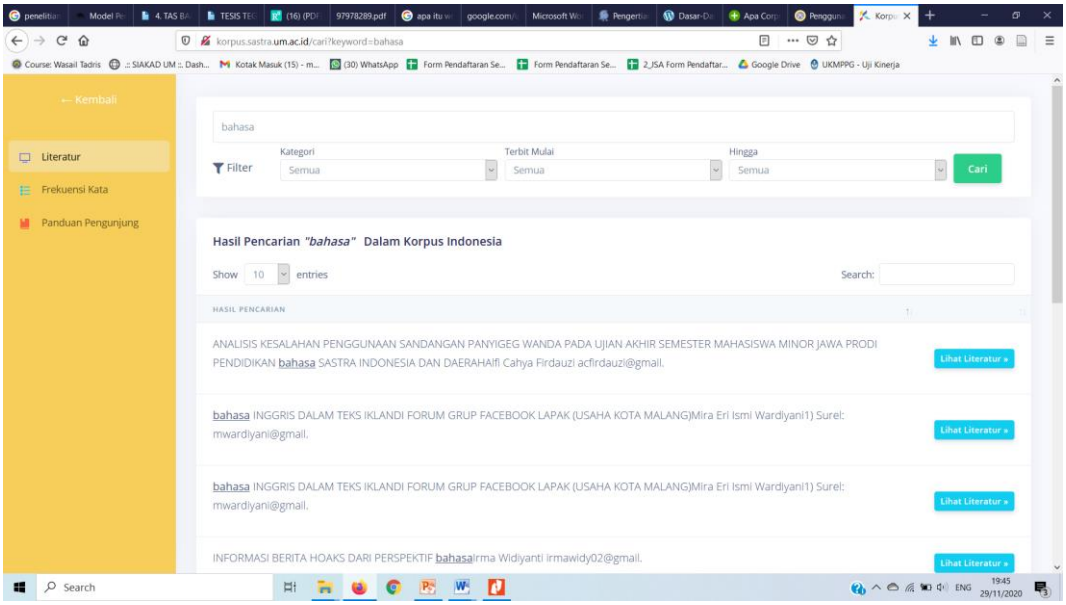


Figure 5.11 Concordance Results

(3) Word Frequency

Word frequency is calculated based on the occurrence of the word in the corpus. The words in the corpus are then sorted by frequency.

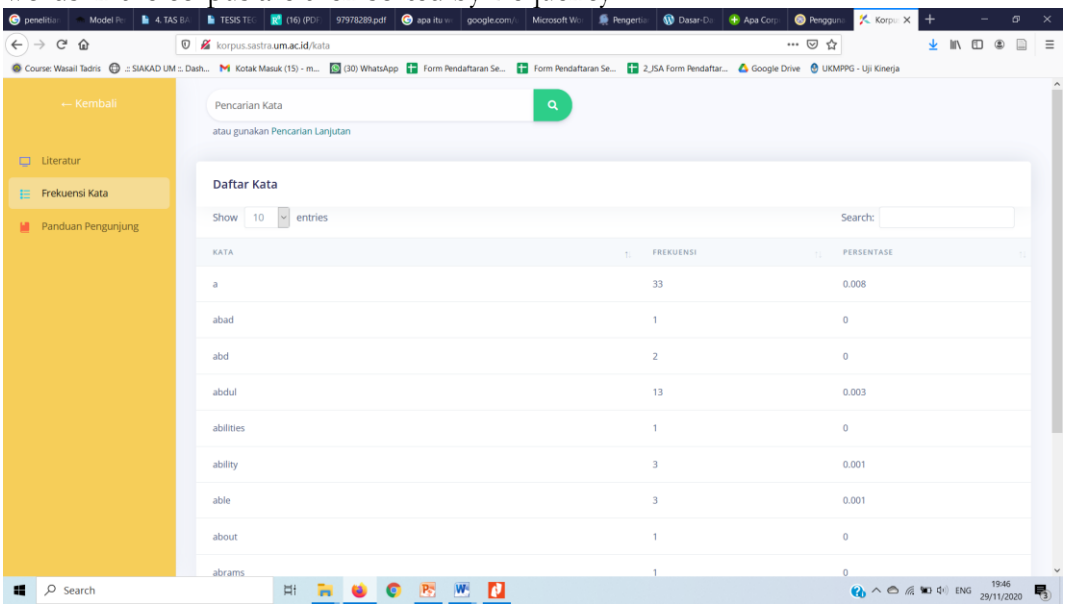


Figure 5.12 Word Frequency Results

d. Implementation

At this stage, we conducted a test of the corpus web to members of the eight IMLA Indonesia's Arabic corpus and students. The trial was carried out for users to view the web corpus and fill out a prepared questionnaire. The questionnaire was in the online form on a google form.

e. Evaluation

At this last stage, we evaluated the validation results.

2. Describing the results of the feasibility test of the language, literature, and art corpus of the Faculty of Letters, Universitas Negeri Malang

After the product trial (implementation), we received the assessment results related to the developed product. The assessment results were used as a reference and to determine the level of feasibility of the product. The assessment was obtained from Media Expert validators and users.

Table 5.1 Validation Results

	Items	Score 5	4	3	2	1	X	Xi	%
	Software Engineering								
1	The development of the application is quite effective.		√				5	5	100%
2	The development of the application is quite efficient.		√				4	5	80%
3	The analysis output displayed has a very good level of consistency.	√					5	5	100%
4	The application maintenance is easy.	√					5	5	100%
5	The application is relatively easy to install.	√					5	5	100%
6	The application is relatively simple in its operation.		√				4	5	80%
7	The selection of the website platform on the application is very appropriate.	√					5	5	100%
8	The application can run well on desktop and mobile devices.		√				4	5	80%
9	The installation package for the application is complete and easy to follow.		√				4	5	80%
10	The installation instructions on the application are clear, concise, and complete.		√				4	5	80%
11	The message box display (warning, notification, alert, etc.) is suitable.		√				4	5	80%
12	There is input validation if there are fields that have not been filled in.	√					5	5	100%
13	The program design in the user manual is clear and easy to understand.	√					5	5	100%

	Items	Score 5	4	3	2	1	X	Xi	%
14	The program code that has been installed can be easily developed into other or similar products.	√					5	5	100%
	Visual Communication Design								
15	The menu layout in the application is easy to access and use.		√				4	5	80%
16	Easy-to-read font selection	√					5	5	100%
17	Easy-to-read font size selection	√					5	5	100%
18	The colors in the application are attractive.	√					5	5	100%
19	Clear navigation icons		√				4	5	80%
20	Complete navigation icons		√				4	5	80%

Qualitative data came from suggestions and comments of validators, as follows.

Table 5.2 Website Quality Evaluation Criteria

No	Aspects	Good	Enough	Lack
1.	Accessibility	√		
2.	Design	√		
3.	Content		√	
4.	Technological aspects and interactivity	√		
5.	Creativity/originality	√		

- Accessibility and Design
  - Easy access
  - Good interoperability
  - Simple yet attractive design
- Content

Menu:
1. Literature: display the selected literature data.
2. Word Frequency: display word list/word count(frequency)
3. Collocation Frequency: display collocation and frequency/number
4. Visitor Guide: display a guide on the content and use of the corpus

- Technological aspects, creativity, and originality
  - The technological aspect is good and open for further development.
  - In terms of creativity, the web design of the corpus is quite original.
  - However, there are a few questions:
  - How about roles?

- Is the web a CQT (Corpus Query Tool) or a CMS (Corpus Management System)?

Quantitative data collected were then determined using the following calculations:

$$\sum = \frac{x}{xi} \times 100\%$$

$$\sum = \frac{91}{100} \times 100\%$$

$$= 91\%$$

The questionnaires from users are as illustrated in the following table:

Table 5.2 QuestionnaireResults from Users

	Items	Score				
		5	**4**	**3**	**2**	**1**
	Learning Design					
1	The analysis results of the number of word frequencies can be understood easily.	5 (62.5%)	3 (37.5%)			
2	The analysis results of the number of word collocations can be understood easily.	3 (37.5%)	5 (62.5%)			
3	Concordance analysis results can be understood easily.	4 (50%)	4 (50%)			
4	Applications can be used easily to find information from the literature.	4 (50%)	4 (50%)			
5	The application is systematic in its analysis process.	3 (37.5%)	5 (62.5%)			
6	The application applies a coherent process in its analysis process.	3 (37.5%)	5 (62.5%)			
7	The logical flow (the process of entering data, performing analysis, and displaying results) of the application is quite clear.	2 (25%)	6 (75%)			
8	The user manual is clear and easy to understand.	3 (37.5%)	4 (50%)	1 (12.5%)		
	Software Engineering					
9	The application runs normally on Windows, Linux, MAC OS, and other operating systems.	3 (37.5%)	4 (50%)		1 (12.5%)	

	Items	Score				
		5	**4**	**3**	**2**	**1**
10	The selection of the website platform on the application is very appropriate.	3 (37.5%)	4 (50%)	1 (12.5%)		
11	The application can run well on desktop and mobile devices.	6 (75%)	1 (12.5%)		1 (12.5%)	
12	Error handling messages are easy to understand.	3 (37.5%)	4 (50%)	1 (12.5%)		
13	The application uses communicative language.	5 (62.5%)	3 (37.5%)			
14	The application has attractive display (interface).	5 (65%)	2 (25%)	1 (12.5%)		
15	Easy-to-read font size selection	5 (62.5%)	3 (37.5%)			
16	*Clear navigation icons*	3 (37.5%)	5 (62.5%)			

The comments from users are as follows:

- I hope the quality can be improved.
- In my opinion, the design is quite good because it is simple. Yet, it would be better for the design to be given other features on each icon and item because the public will use it.
- There should be an index on National and International Corpus Clusters.
- It must be developed even more.
- Literature needs to be added
- It should be directed to developing multi-disciplinary studies and new issues on the initiation of foreign language learning.
- Sorry, I could not open my cellphone yesterday. But on my friend's cellphone, I could.
- Concordance frequency experienced an error (appears 404 not found).

Discussion

Corpus linguistics is defined as the study of the compilation and analysis of corpora (Cheng). McEnery and Hardie define corpus linguistics as a field that focuses on procedures or methods of studying or researching language (McEnery dan Hardie). McEnery and Hardie also mention the approach used in corpus linguistics, which Tognini-Bonelli also proposes. Tognini states two corpus linguistic approaches: corpus-based and corpus-driven (Tognini-Bonelli). Both have differences in viewing the corpus as evidence that supports the theory. The first uses a deductive approach. Meanwhile, the corpus-driven approach considers the corpus as evidence that must be a reference for theory so that it is inductive.

The development of the language, literature, and art corpus web was carried out through several stages. The first stage was requirements gathering. In the first stage, all requirements, information, and initial data needed for application development were explored and collected. In addition to those directly related to application development, at this first stage, the design of the corpus of the undergraduate theses, theses, and dissertations

was also determined in the form of abstracts and articles as the content in the web corpus. In terms of the internal structure of the corpus, the data collection consisted of written data. The corpus application had some features, namely concordance, word frequency, and collocation features.

Concordance is the mainstay of the corpus. Concordance allows strings and related words (McEnery). According to Setyawan, concordance is a collection of word forms in their respective textual environments. The simplest form of concordance is an index. Each word formation is indexed and the reference refers to a place in a text (Setyawan, "Pengertian Konkordansi dan Cara Penggunaannya dalam Korpus Linguistik"). The most common format used to bring up word formations is KWIC (Key Word in Context), which displays the searched word with several characters before and after the word appears. Like the word 'المدرسة' in KorSA. The first ten occurrences of the word formation are presented in text order in the middle of a seventy character context (spaces and punctuation marks count as characters).

Related to word frequency, one of the text processing steps in text information retrieval systems or text mining is text cleaning from irrelevant words as an index. In a text document, there may be many types of words such as prepositions, conjunctions, pronouns, adjectives, and so on. Some of these words may not have the potential to be indexed for documents because their occurrence is not unique or has never been used in search queries. For this reason, a filtering process for these words (Luhn and Flood) is carried out. Filtering is done by providing a list of words that are not indexed (stopword list). Zipf's law is sometimes used to form stopword lists, especially in word occurrence analysis (Zipf).

The word frequency, apart from getting data on how many times the word appears in the web corpus, can also be used to analyze the word. Setyawan has asserted that the benefit of word frequency is understanding each unit in the text (Setyawan). Each text has several levels of meaning, and these levels tend to be related to physical and structural units, ranging from single words, phrases, clauses, sentences to the whole text. One of the fundamental problems encountered when processing text is the question of what exactly a word is. We might naively argue that words are entities in text separated by spaces or punctuation marks. This definition of a word largely ignores the fact that, in practice, a word does not consist of only one entity delimited by spaces or punctuation marks.

Conclusion

Based on the analysis results and discussion, the following conclusions are presented:

1. In designing and developing the language, literature, and art corpus of the Faculty of Letters, Universitas Negeri Malang, several stages were carried out: analysis, design, development, implementation, and evaluation.
2. The feasibility test results for the corpus of language, literature, and art at the Faculty of Letters, Universitas Negeri Malang, are valid with a value of 91%. This means that the web corpus developed is feasible to use.

References

Ahsanuddin, Mohammad. (2018). *Tashmim Al-mudawwanah Al-Mutawaziyah Li mustakhlash Al-buhuts Al-Ilmiah Al-Indunisiya Al-Arabiyah 'Ala Dhaui Nadzariyah Mona Baker Li Al-Takafu' Al-Lughawi Fi Al-Tarjamah*. Disertasi. Malang: UIN Maulana Malik Ibrahim Malang.

Anthony, Laurence. (2013). A Critical Look at Software Tools in Corpus Linguistics. *Linguistic Research*, *30* (2), 141–61. DOI.org (Crossref), doi:10.17250/khisli.30.2.201308.001.

Baker, Paul. (2010). Sociolinguistics and Corpus Linguistics. Edinburgh University Press.
Baker, P. (2010). *Corpus Methods in Linguistics. In Litosseliti, Lia. Research Methods in Linguistics*. New York: Continuum International Publishing Group
Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.
Cheng, W. (2012). *Exploring Corpus Linguistics: Language in Action*. Routledge. https://books.google.co.id/books?id=yqgdkgAACAAJ.
Eddakrouri, Ayman. (2016). *Web-based (Searchable) corpora*. Infoguistics https://sites.google.com/a/aucegypt.edu/infoguistics/directory/Corpus-Linguistics/arabic-corpora.
Flood, B. J. (1999). Historical Note: The Start of a Stop List at Biological Abstracts. *JASIS*, *50*(12).
Halliday, M.A.K, Wolfgang Teubert, Collin Yallop, & Anna Cermakova. (2004). *Lexicology and Corpus Linguistics: An Introduction*. London: Continuum.
Hunston, Susan, Sara Laviosa, dan Nicholas Groom. n/a. *Corpus Linguistics*. Birmingham: The Centre for English Language Studies, University of Birmingham.
Kushartanti, Untung Yuwono, dan Multamia RMT Lauder. (2007). *Pesona Bahasa: Langkah Awal Memahami Linguistik*. Jakarta: Gramedia Pustaka Utama.
Kilgarriff, Adam, dan Gregory Grefenstette. (2003). Introduction to the Special Issue on the Web as Corpus. Computational Linguistics, *29*(3), 333–47. DOI.org (Crossref), doi:10.1162/089120103322711569.
Luhn, H. P. (1959). Key word-in-context index for technical literature (kwic index). *American Documentation*, *11*(4).
McEnery, Tony. (1997.). Multilingual Corpora - Current Practice and Future Trends. pp. 12.
McEnery, Tony, & Andrew Hardie. (2012). *Corpus Linguistics: Method, Theory, and Practice*. Cambridge University Press.
McEnery, Tony, dan Andrew Wilson. (1966). *Corpus Linguistics*. Edinburgh University Press, 1996.
Sasongko, Jati. (2010). *Aplikasi untuk Membangun Corpusdari Data Hasil Crawling dengan Berbagai Format Data Secara Otomatis*.
Setiawan, T. (2017). *Linguistik Korpus dalam Pengajaran Bahasa. Seminar Nasional Perspektif Baru Penelitian Linguistik Terapan*. Yogyakarta: Universitas Negeri Yogyakarta.
Setyawan, Aan. (2018). Manfaat daftar Frekuensi dalam Korpus Bahasa." belajarbahasa.id, https://belajarbahasa.id/artikel/dokumen/516-manfaat-daftar-frekuensi-dalam-korpus-bahasa-2018-03-25-23-57.
Setyawan, Aan. (2018). Pengertian Konkordansi dan Cara Penggunaannya dalam Korpus Linguistik." belajarbahasa.id, https://belajarbahasa.id/artikel/dokumen/512-pengertian-konkordansi-dan-cara-penggunaannya-dalam-korpus-linguistik-2018-03-25-22-00.
Tognini-Bonelli, E. Corpus Linguistics at Work. J. Benjamins. (2001) https://books.google.co.id/books?id=6YDRH45MpL8C.
Teubert, Wolfgang, and Ramesh Krishnamurthy. (2007). *Corpus Linguistics: Critical Concepts in Linguistics*. London: Routledge.
Tim Redaksi KBBI. (2008). *Kamus Besar Bahasa Indonesia Edisi Keempat*. Jakarta: Gramedia Pustaka Utama.
Wagner, Joachim. (2006). Nadja Nesselhauf, Collocations in a Learner Corpus: John Benjamins, Amsterdam. *Machine Translation*, *4*(6), 301–03. DOI.org (Crossref), doi:10.1007/s10590-007-9028-8.
Zipf, H. Human Behaviours and the Principle of Least Effort. Addison-Wesley, 1949.

http://www.americannationalcorpus.org/OANC/index.html. Diakses 3 Agustus 2014.
http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus Linguistics Practical Introduction.pdf. Diakses pada 13 Juli 2014.
http://corpus.byu.edu/coca/compare-boe.asp. Diakses 3 Agustus 2014.
http://corpus.byu.edu/coca/compare-bnc.asp. Diakses 3 Agustus 2014.
http://www.macmillandictionaries.com/features/from-corpus-to-dictionary/. Diakses 14 Juli 2014
http://www.oxforddictionaries.com/words/about-the-oxford-english-corpus. Diakses 14 Juli 2014
http://www.oxforddictionaries.com/words/the-corpus-and-the-dictionary-entry. Diakses 14 Juli 2014
https://the.sketchengine.co.uk/auth/corpora/. Diakses 14 Juli 2014.
http://www.ucl.ac.uk/english-usage/projects/ice.htm. Diakses 3 Agustus 2014.