

THESIS

**ACADEMIC VOCABULARY LIST FOR
UNDERGRADUATE STUDENTS IN WRITING
UNDEGRADUATE THESES: A CORPUS-BASED
ANALYSIS**

Written and Submitted by:

FISMA

F022182014



**ENGLISH LANGUAGE STUDIES
POSTGRADUATE PROGRAM
HASANUDDIN UNIVERSITY
MAKASSAR
2021**

**ACADEMIC VOCABULARY LIST FOR UNDERGRADUATE
STUDENTS IN WRITING UNDERGRADUATE THESES:
A CORPUS-BASED ANALYSIS**

Thesis

As a partial fulfillment of the requirements of Magister Degree

**English Language Studies
Faculty of Cultural Sciences**

Written and Proposed by

FISMA

**Post Graduate Program
Hasanuddin University
Makassar**

2021

APPROVAL SHEET (THESIS)

**ACADEMIC VOCABULARY LIST FOR UNDERGRADUATE STUDENTS IN
WRITING UNDERGRADUATE THESES: A CORPUS BASED ANALYSIS**

Written and Submitted by

**FISMA
F022182014**

Has been defended in front of the thesis examination committee which was formed in order to complete the study of the Master Program in English Language Studies Faculty of Cultural Sciences Hasanuddin University on February, 3rd 2021 and is declared to have met the graduation requirements.

Approved by:

Head of
The Supervisory Committee

Member of
The Supervisory Committee

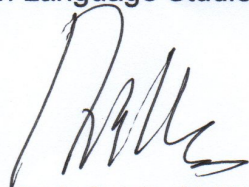


Dr. Abidin Pammu, M.A., Dip. TESOL
NIP. 19601231 198601 1 071

Dra. Ria Rosdiana Jubhari, M.A., Ph.D
NIP. 19660207 199103 2 001

The Head of
English Language Studies Program

The Dean of
Faculty of Cultural Sciences



Dr. Harlinah Sahib, M.Hum.
NIP. 19621128 198703 2 001



Prof. Dr. Akin Duli, M.A.
NIP. 19640716 199103 1

STATEMENT OF AUTHENTICITY

The undersigned:

Name : Fisma
Register number : F022182014
Study Program : English Language Studies
Level of Education : Postgraduate (S2)


States truthfully that this thesis entitled:

“Academic Vocabulary List for Undergraduate Students in Writing Undergraduate Theses: A Corpus-Based Analysis” was the result of my own work.

If it is proven later that some part of this thesis is the work of others, I am willing to accept any sanctions for my dishonesty.

Makassar, 3rd February 2021

The Researcher


PT TERAI TEMPEL
TGL
7998AHF865605988
6.000
ENAM RIBU RUPIAH
Fisma

ACKNOWLEDGMENTS

Alhamdulillah Rabbil 'Alamin. First and foremost, let me praise and give thanks to Allah *Subahanahu Wa ta'ala* for all of entrusted goods that was given to me especially the good health, knowledge, guidance and the opportunity to undertake this research and complete it satisfactorily. Peace and salutation be upon to the prophet Muhammad *Sallallahu Alaihi Wasalam*, his family, his companions and his followers.

I would like to express my gratitude and appreciation to my supervisors Dr. Abidin Pammu, Dip. TESOL., M.A and Dra. Ria Rosdiana Jubhari, M.A., Ph.D. for the guidance, support, encouragement, and countless hours of mentorship throughout my graduate studies. Many thanks also go to my examiners; Prof. Dr. Noer Jihad Saleh, M.A, Dra. Nasmilah, M.Hum., Ph.D., and Dr. Sukmawaty, M.Hum for the comments and suggestions. Their contribution leads me to accomplish this research much better. Furthermore, I thank all the lectures of English Language Study Program for wonderful academic support and Civitas Academica in Faculty of Cultural Science for a good service, the help and participation during my study in Hasanuddin University. For my best friend Purnama Cahya, thanks for sharing a great idea about corpus-based research. Hopefully, this research can give a good contribution for education in Indonesia, particularly for the lecturers, the students and the scholars in English Language Study Program.

I also thank to *Lembaga Pengelola Dana Pendidikan (LPDP)* as my sponsorship during my study. For Ibu Rosniati, thank you for giving me a place to stay while living in Makassar. For Ibu Hafirah and Ibu Magfirah thank you for your support and the help. For all my friends at ELS, thanks for your support, participation, the great experience and the amazing journey during our study.

Special thanks to my parents and my big family for the support, the endless prayers therefore this thesis can be completed properly. Last and most

special, I would like to thank my beloved husband for the support and encourage me to this journey.

Makassar, 3rd February 2021

A handwritten signature in black ink, appearing to be 'Fisma', enclosed within a large, horizontal, hand-drawn oval shape.

Fisma

ABSTRACT

FISMA (F022182014) *Academic Vocabulary List for Undergraduate Students in Writing Undergraduate Theses: A Corpus-based Research* (supervised by Abidin Pammu and Ria Rosdiana Jubhari)

This is corpus-based research aiming to find out the word frequency which appears in the undergraduate theses based on Coxhead's word selection criteria; to find out the distribution of Coxhead's (2000) Academic Word List (AWL) and Gardner and Davies' (2014) AVL in the undergraduate theses in Indonesia particularly in Education field and to establish new academic vocabularies for undergraduate students in writing undergraduate theses.

A 2.2-million word corpus was compiled comprising 200 undergraduate theses deriving from 10 universities in Indonesia. The LancsBox 5.0 software was exerted to calculate the word frequency and percentage of the text by applying the quantitative method.

By implementing Coxhead's word selection criteria, only 1-million words which appear more than 63 times in the undergraduate theses and at least appearing 5 times from 5 universities. The research indicates that Coxhead's AWL is as much as 8.95% (201,286), Gardner and Davies' AVL is as much as 24.79% (557,658) and a new academic vocabulary is 4.46%. There are 268-word forms of AWL, 573-word forms of AVL, and 276 new academic vocabularies found in undergraduate theses. Moreover, AWL words, AVL words, and new academic vocabularies found in the undergraduate theses are accumulated into one list. In sum, 753 words become an academic vocabulary list for the students in writing their undergraduate theses. The total of those word appearances is 642,846 with a percentage of 28.75% of the total of word appearances in the corpus. Based on the findings of the results, it can be concluded that Coxhead's academic word distribution in undergraduate theses of English education is still low compared with the distribution of AWL in other fields, while Gardner and Davies' AVL still contains many general high-frequency words. The new academic vocabulary in this study still needs to be developed. Therefore, it is recommended to improve the development of academic vocabulary lists particularly in the more specific department such as English Education by involving more relevant academic texts. It is also important to consider the words that unfamiliar utilized by the students but high frequency used by native speakers or academicians in writing academic papers.

Keywords: Academic Vocabulary List; New Academic Vocabularies; Corpus-based; Undergraduate Theses

ABSTRAK

FISMA (F022182014) *Daftar Kosakata Akademik untuk Mahasiswa S-1 dalam Menulis Skripsi: Penelitian Berbasis Korpus* (dibimbing oleh Abidin Pammu and Ria Rosdiana Jubhari)

Penelitian ini adalah penelitian berbasis korpus yang bertujuan untuk mengetahui frekuensi kata yang muncul di skripsi berdasarkan pada kriteria penyeleksian kata-kata yang digunakan oleh Coxhead, mengetahui pendistribusian AWL oleh Coxhead (2000) dan AVL oleh Gardner and Davies (2014) dalam skripsi di Indonesia (khususnya bidang pendidikan) dan mengembangkan kata-kata akademik yang baru untuk mahasiswa S-1 dalam menulis skripsi.

Sebanyak 2,2 juta kata (korpus) yang terkompilasi dari 200 skripsi yang berasal dari 10 kampus yang ada di Indonesia. Penelitian ini menggunakan *software Lancsbox 5.0* untuk mengakulasi frequency dan persentase kata dari teks dengan menerapkan metode kuantitatif.

Dengan menggunakan seleksi pemilihan kata yang akademik dari Coxhead, hanya 1 juta kata yang muncul lebih dari 63 kali di skripsi dan muncul setidaknya 5 kali dari 5 universitas. Hasil penelitian ini menunjukkan bahwa AWL oleh Coxhead sebanyak 8,95% (201.288 dari total kata secara keseluruhan), AVL dari Gardner and Davies sebanyak 24.79% (557.658 dari total kata secara keseluruhan). Terdapat 268 bentuk kata di skripsi yang ditemukan di AWL, 573 di AVL dan 276 kata akademik baru yang ditemukan. Selanjutnya, kata-kata akademik dari AWL, AVL dan kata akademik baru yang ditemukan di skripsi digabungkan menjadi 1 daftar kosakata. Secara keseluruhan, 753 kata yang dijadikan sebagai daftar kosakata akademik untuk mahasiswa S-1 dalam menulis Skripsi. Artinya, pendistribusian kata akademik dari Coxhead dalam skripsi pendidikan bahasa Inggris masih rendah dibandingkan AWL di bidang yang lain, sedangkan kosakata akademik dari Gardner dan Davies masih mengandung banyak kosakata umum. Sementara kosakata akademik baru yang ditemukan di dalam penelitian ini juga masih perlu dikembangkan. Oleh karena itu, penelitian ini mendukung pentingnya pengembangan kosakata akademik baru yang melibatkan sumber akademik yang berkaitan dengan pendidikan bahasa Inggris. Dan juga mempertimbangkan kata-kata yang jarang digunakan oleh siswa, namun sangat sering digunakan oleh *native speaker* dalam menulis penelitian akademik.

Kata kunci: Kosakata Akademik; Korpus; Skripsi

TABLE OF CONTENTS

TITLE PAGE	i
APPROVAL SHEET	ii
STATEMENT OF AUTHENTICITY	iii
ACKNOWLEDGMENTS	iv
ABSTRACT	vi
ABSTRAK	vii
TABLE OF CONTENTS	viii
LIST OF TABLES	x
LIST OF APPENDICES	xi
LIST OF ABBREVIATIONS	xii
CHAPTER 1 INTRODUCTION	1
A. Background	1
B. Research Questions.....	7
C. Objectives of the Study	7
D. Significance of the Study	8
E. Scope and Limitation of the Study	8
CHAPTER II LITERATURE REVIEW	10
A. Previous Related Studies	10
B. Theoretical Background.....	14
1. Academic Vocabulary	14

2. Academic Word list	16
3. Academic Vocabulary List	19
3. Corpus-Based Analysis	21
4. Computer Software	25
C. Conceptual Framework	27
CHAPTER III RESEARCH METHODOLOGY	29
A. Research Design	29
B. Population and Sample	29
C. Research Instrument	30
D. Data Collection Procedure	31
E. Data Analysis	32
CHAPTER IV FINDING AND DISCUSSION	35
A. Findings	35
1. Frequency Analysis of AWL in Undergraduate Theses.....	36
2. Frequency Analysis of AVL in Undergraduate Theses.....	39
3. New Academic Vocabularies in undergraduate theses.....	41
B. Discussion	42
CHAPTER V CONCLUSIONS AND SUGGESTIONS	48
A. Conclusions	48
B. Suggestions.....	49
BIBLIOGRAPHY	51
APPENDICES.....	55

LIST OF TABLES

Table 1 The Proportion of words in undergraduate theses	36
Table 2 Top 10 most frequently used AWL in undergraduate theses	37
Table 3 Frequency per sublist of AWL in undergraduate theses.....	38
Table 4 Top 10 most frequently used AVL in undergraduate theses	40
Table 5 Top 10 most frequently used of new academic vocabularies in undergraduate theses	42

LIST OF APPENDICES

Appendix A : The Captures of LanscBox Software

Appendix B : The Captures of Data Collection Procedures

Appendix C : High Frequency Word Based on Word Selection Criteria

Appendix D : AWL in undergraduate theses

Appendix E : AVL in undergraduate theses

Appendix F : New Academic Vocabularies

Appendix G : Academic Vocabulary List in writing undergraduate theses

LIST OF ABBREVIATIONS

AWL	: Academic Word List (Coxhead)
AVL	: Academic Vocabulary List (Gardner & Davies)
GSL	: General Service List (West)
NGSL	: New General Service List (Browne & Phillips)
COCA	: Corpus of Contemporary American English
EFL	: English Foreign Language
SAJ	: Science Academic Journal
EAP	: English Academic Purposes
ESL	: English as Second Language

CHAPTER I

INTRODUCTION

This chapter presents background, research questions, objectives of the research, significances of the research, and scope of the research.

A. Background

The process of learning the words of a language is referred to as vocabulary acquisition. This acquisition of new vocabulary plays a strategic role in bridging the learners to acquire new knowledge especially foreign language learners. The ways in which English Foreign Language (EFL) students acquire the vocabulary of a native language differ from second language learners in acquiring the vocabulary of a second language. Since the acquisition of vocabulary knowledge affects the proficiency in language learning, understanding vocabulary is considered as the crucial component in learning English (Nation, 2001).

In an academic setting, vocabulary serves as a tool to ensure success in academic writing and publication. A substantial amount of researchers agreed that academic vocabulary plays an important role for learners and educators at University as a guide in writing academically and comprehending academic text. (It- Ngam & Phoocharoensil, 2019; Khani &Tazik, 2013; etc.)

The development of the academic vocabulary list can be traced back to the Academic Word List (AWL) from Coxhead (2000) as a widely used of word list. Coxhead's AWL focuses on word families and consists of

570-word families compiled from 3.5 million words of English Academic text by examining the range and frequency of words. It involves four disciplines: Arts, Science, Law, and Commerce. As a result, Coxhead (2000) ensures that the new academic word list coverage 10% of total tokens in the academic corpus. The list did not involve the first 2,000 words in General Service List (West, 1953). At the present time, Coxhead's AWL is widely used as a reference for building a new academic vocabulary list in English Academic Purposes or English Specific Purposes and as better sources for vocabulary learning in the English language. The effectiveness of Coxhead's (2000) Academic Word List has been discussed in some studies (e.g., Pathan et al., 2018; Vongpumivitch et al., 2009, etc.). Pathan et al. (2018), for instance, have proved that the use of Coxhead's AWL (2000) is effective in writing doctoral theses.

A number of studies in second language acquisition such as Martinez et al. (2009) and Mozaffari & Moini (2014) do not consider the AWL as a useful source in the terms of text coverage in the specific fields and disciplines. Sulaiman, Salehuddin, & Khairuddin (2018) found that Malaysian English undergraduates' knowledge of academic words based on Coxhead's Academic Word List (2000) still low because there are many fields and disciplines on there. Additionally, Mozaffari & Moini (2014) found the distribution of Academic Word List in Education Research Articles has low coverage of around 4.94%. In response, it is essential to develop a new academic vocabulary list specific in the education field.

Gardner and Davies (2014) also found a new Academic Vocabulary List (AVL) taken from a larger corpus, which contains 120 million academic texts of the 425-million-word Corpus of Contemporary American English (COCA; Davies 2012) which involve general discipline. Gardner and Davies' AVL focuses on lemmas than word families. They also provided part of speech in their list that can make it more accessible to determine the function of the word. Meanwhile, Durrant (2016) found that Academic Vocabulary List relatively has a small core in university students writing. Therefore, developing a list of academic vocabulary from university students' writing is needed.

Academic Vocabulary is one of the most challenging aspects of making decisions in which words are worth on teaching (Coxhead, 2000; Vongpumivitch et al., 2009). The teachers do not know the occurrences of the word and which words are genuinely representative in teaching new vocabulary (Chanasattru & Tangkiengsirisin, 2017). The effectiveness of teaching academic vocabulary is determined by the sources. However, providing academic vocabulary from the general field seems ineffective because some words which are frequent in one field may be absent in another field (Xue and Nation, 1984), and the students in different areas have different needs.

Some scholars have built a new academic vocabulary in several fields (e.g., It-ngam & Phoocharoensil, 2019; Lei & Liu, 2016; Coxhead, 2000; etc.). The results of all their research have an excellent contribution

to the learning process. However, all of these studies have addressed the academic vocabulary from journal articles, meaning that they provided the academic vocabulary that the students required to know receptively. Beside, Durrant (2016) has argued that "productive use of vocabulary requires more knowledge than receptive; therefore, a pedagogical focus on productive vocabulary is at least as important as one of receptive vocabulary." In response to these issues, the researcher formulates that it is crucial also to know the productive use of vocabulary gathered from the undergraduate students' theses. Furthermore, the list of words from the undergraduate thesis can be used as a reference to find out whether the words are meaningful for the students' writing and need to be developed to improve vocabulary learning.

In the University, the students should read and write academic text. However, insufficient vocabulary knowledge is the most problem of English as a Foreign Language (EFL) learners faced in learning academic discourses because it is rarely used in daily activities (Malmström et al., 2018; Mozaffari & Moini, 2014). Sometimes the students find the difficulty to distinguish which words are academic words and which are general words. The preliminary study reported that choosing the appropriate term is one of the challenges in writing theses. Sometimes students also found it difficult in paraphrasing the idea and looking for the appropriate synonym of the word to avoid plagiarism. The students also rarely used theses as references for academic Writing (Prihantoro, 2016). In Indonesia, some

scholars also reported that the vocabulary and academic vocabulary knowledge of university students is relatively low (Nasir & Chinokul, 2018; Novianti, 2016). A recent account of tertiary learners' acquisition of vocabulary, for example, has informed that EFL learners despite years of their learning English as a foreign language have problems comprehending academic texts because they have limited possession of vocabulary.

Studies also have documented that the majority of these learners have performed poorly in the test because they have failed to infer meaning from the reading due to vocabulary problems. Written evidence from a number of Language Centre around the Eastern part of Indonesia also indicates that many students from different disciplines have been prevented from the opportunity of going abroad because they did not fulfill the required TOEFL/IELTS Band score. Such reading problems and academic writing have recently been one of the major concerns of researchers and teaching practitioners at the University of Hasanuddin and they have been aware of the need to provide pedagogical remedies. Since underperformance in reading has been considered a serious problem, researchers and academicians have attempted to offer various pedagogical approaches suitable to EFL social and cultural context in Indonesia, such as extensive reading intervention and reading strategy training.

The development of the Academic Vocabulary List from the undergraduate thesis is very helpful not only for the students but also for

the teachers. Knowing academic vocabulary since writing undergraduate theses will bridge the students to publish their research. It also will help them in their future studies. Direct attention to the specific word from the vocabulary list can lead the teacher and the students' development in writing academically. It can help the teacher in designing the syllabus and construct relevant teaching material.

Nowadays, the researcher did not find academic vocabulary research compiled from undergraduate theses. It-ngam & Phoocharoensil (2019) suggested exploring other text types of academic text, such as theses, for the future research in investigating academic vocabulary. Therefore, the current study focuses on investigating the words taken from the undergraduate theses, by using corpus-based analysis, the frequency of the occurring words that the students use can be tracked. It is also supported by Mackey (1965) cited in Durrant, P. (2016) that "since items occurring the most frequently are those which the learner is more likely to meet, they are the ones which are selected for teaching." Corpus-based research is research that involves the collection of the word identified by using the software program. The software program shows how many times the words occur on the corpus. This study developed an academic vocabulary list obtained from students' undergraduate theses by considering several criteria in determining academic vocabularies such as those used in Coxhead's AWL (2000) and still comparing the academic vocabulary list produced by Gardner & Davies (2014) and Academic Word

List from Coxhead (2000). Therefore, the teacher can choose which words should be taught in teaching vocabulary. Based on the previous description, the researcher conducted the research entitled: Academic Vocabulary List for Undergraduate Students in Writing Undergraduate theses. Academic vocabulary list refers to a collective of academic words / academic vocabulary that found in this study

B. Research Questions

The following are formulated Research Questions:

1. Which are the words that frequently occur in undergraduate theses based on Coxhead's word selection criteria?
2. What are the academic vocabularies in the Coxhead's AWL that occur in undergraduate theses?
3. What are the academic vocabularies in Gardner & Davies's AVL that occur on undergraduate theses?
4. What are the new academic vocabularies found in the students' undergraduate theses based on Coxhead's word selection criteria?

C. Objectives of the Study

The present study is guided by the objectives of:

1. To find out the most frequent occurrence of the words in undergraduate theses.
2. To find out the distribution of the Coxhead's AWL that appear on undergraduate theses.

3. To find out the distribution of Gardner & Davies's AVL that appear on undergraduate theses.
4. To establish the new academic vocabularies for undergraduate students in writing undergraduate theses.

D. Significance of the Study

Understanding academic vocabulary and the words that frequently appear on undergraduate theses, particularly in English Education Study Program can help the students to acquire academic vocabularies and to enrich their competence in writing an undergraduate thesis. The result of this research can help the teacher to teach the academic vocabularies effectively and analyze which words their students used in their academic text. This study reports vocabulary that the students know, so the teachers can analyze what items of vocabulary that the learners need to know as well as can use in Academic Writing subject as material and designing syllabus. Academic Vocabulary List produced from this study can help the students in writing academic text, mainly writing an undergraduate thesis.

E. Scope and Limitation of the Study

This research focused on the distribution of academic vocabulary from Coxhead (2000) and Gardner and Davies (2014) and the development of a new academic vocabulary list for undergraduate students in writing undergraduate theses. The English Education Study Program was addressed all of them. To analyze the vocabulary, the researcher used a corpus-based analysis and Academic Vocabulary List from Gardner and

Davies (2014) and Academic Word List from Coxhead (2000) and compare it with New General Service List (Brownie and Philips, 2014).

CHAPTER II

LITERATURE REVIEW

This chapter presents previous related studies, theoretical review, and conceptual framework. This chapter is essential as it acts as the basis of knowledge for the researcher.

A. Previous Related Studies

A review of the literature indicates that there is a good number of studies conducted on academic vocabulary in specific fields, specific subjects, and also across different fields. Some of researchers have developed new academic word lists in the field of Medical (Chen & Ge, 2007; Lei & Liu, 2016), Science (It-ngam & Phoocharoensil, 2019), Linguistics (Khani & Tazik, 2014), Chemistry (Valipouri & Nassaji, 2013) and Nursing (Yang, 2015). Additionally, some researchers only identify the distribution of academic words in academic texts (Vongpumivitch et al., 2009; Martinez et al., 2009; *etc.*).

Pathan *et al.* (2018) have investigated Academic Vocabulary Use in Doctoral thesis entitled: Academic Vocabulary Use in Doctoral Theses: "A Corpus-Based Lexical Analysis of Academic Word List (AWL) in Major Scientific Disciplinary Groups". This study involves 200 doctoral theses from physical sciences, biological and health sciences, and covering 17 disciplines areas to investigate the frequency and coverage of academic vocabulary in scientific doctoral theses texts using AntConc version 3.4.4 software. The results revealed that 550 world families (96.50%) among the

total 570-word families of Coxhead's (2000) AWL are found to be frequently used in the doctoral theses of scientific disciplinary groups including Biological & Health science and Physical sciences. The coverage of AWL in Pakistani Doctoral Theses text is 8.76%. It is accurate to conclude that Coxhead's (2000) AWL is sufficient for writing doctoral theses. However, the corpus in this study is not balanced because the researcher did not involve the same number of doctoral thesis in each discipline. Focusing on disciplinary groups is not effective because words have different functions and meanings across different disciplines (Martinez et al., 2009).

The distribution of the Academic Word List has also been addressed in several research articles on various subjects. Chanasattru & Tangkiengsirisin (2017) investigate the use of Academic Word List (AWL) and New General Service List (Browne & Philips, 2014) entitle: "The Word List Distribution in Social Science Research Articles". Sixty-four English were selected and loaded in AntWordProfiler 1.4.0. The result reported that the distribution of Academic Word in Social Science Research Articles is a good percentage covering 13.86% and the coverage of NGSL accounted for around 70%. The finding of this research revealed that AWL has an excellent contribution to vocabulary learning, particularly in preparing students for reading and writing social science research articles.

In contrast, Mozaffari and Moini (2014) also try to investigate the use of Academic Word List in the field of education entitled: " Academic Words in Education Research Articles: A Corpus Study". 239 Education

research Articles were involved and loaded into WordSmith Tools software. The finding reported that word lists from the Education Research Articles corpus could be considered as one of the efficient and best methods in language learning. However, the use of AWL in Education research articles is still low. The coverage of AWL word forms in the Education Research Articles Corpus was only 4.94%. Based on the findings, they recommended an attempt to create a discipline-specific word list that leads the students to learn which one of the words is necessary for their field of study. Concerning this issue, it means that AWL in Education specific fields is very crucial to be investigated.

Khani and Tazik (2013) tried to establish an academic wordlist specific to applied linguistics entitled: "Towards the Development of an Academic Word List (AWL) for Applied Linguistics Research Articles". This research contains 1,553,450 running words taken from 240 Applied Linguistic Research Articles from prestigious journals; there was an attempt to develop an academic word list for this field of study. The data were compiled by using the Range Program. By the predetermined word selection criteria plus exclusion of GSL words, 773-word types were finally selected as the academic words which appeared 193,989 times in the corpus. Similarly, the finding of this research showed that academic vocabulary is hugely influenced by pedagogical implications. They also argued that academic word is very crucial to be acquired by the students.

Another study concerning academic word is the development of the academic science word list provided by It-ngam & Phoocharoensil (2019) entitled: "The development of science academic word list". This study compiled specialized academic words across 11 sub-disciplines of natural science. The words were identified by using a corpus-based approach and an expert-judged approach. Two concordance programs were utilized: AntWordProfiler and AntConc. Their study involved a 5.5-million-word corpus conducted from 1,062 journal articles in science disciplines called the Science Academic Journal (SAJ) Corpus. By considering the word selection criteria, this study found that 513-word families met the word selection criteria, and the experts agreed to remove 81 words from the list. Therefore, the Science Academic Word List was created with 432-word families and provided 5.82% coverage of the running words in the SAJ corpus. Similarly, word lists from this study will be useful for learning and teaching vocabulary in natural science.

At the present time, some scholars had developed a new academic vocabulary in general disciplines (Gardner and Davies, 2014; Coxhead Averil, 2000). However, academic vocabulary compiled from multiple disciplines are seemed not effective, therefore some scholars also try to develop a new academic vocabulary in a specific discipline, for example, Lei & Liu (2016) develop a new academic word list in the field of medical, It-ngam & Phoocharoensil (2019) in the field of Science, Khani &

Tazik (2014) in the field of Linguistics, Valipouri & Nassaji (2013) in the field of Chemistry and Yang (2015) in the field of Nursing and et cetera.

The previous researches showed that all of the scholars were interested to investigate the distribution of Academic Word List from Coxhead (2000), but only a few of the scholars tried to investigate the distribution of Academic Vocabulary List from Gardner and Davies (2014) that more update. Nowadays, the researcher did not found the academic vocabulary list compiled from undergraduate theses in a specific subject. Most of the previous study only involved journal articles in their study to know how many times of the word occur. In fact, knowing the distribution of Academic Word List and Academic Vocabulary List compiled from the undergraduate theses are also important as references for teachers in teaching vocabulary and for students in writing academically particularly in a specific subject. To fill this gap, the current study focuses on the distribution of AWL and AVL from the undergraduate theses focus on Education Study Program and provides a new academic thesis academic wordslist. This study also used a new computer software program that was never used in previous research particularly in AWL research.

B. Theoretical Background

1. Academic Vocabulary

The phenomenon of Academic Vocabulary is one of the crucial topics in English language learning. English vocabulary was categorized into four items (Nation, 2001); high-frequency words, academic words,

technical words, and low-frequency words. High-frequency words have a large capacity and as the essential foundation for all language use. It covers about 80% of most English text, such as found 2,000-word families in West's (1953) General Service List (GSL). Academic words frequently occur in the academic text that encompasses 8%-10% of running words, such as Coxhead's Academic Word List (AWL) (2000), which found around 10% of total words in academic text. Technical vocabulary which distinct by subject area covers up to 5% of texts. Low-frequency words contain the narrow word and infrequently occurring words within the text. Commonly, they appear once or twice and will not appear again for a long time.

Academic vocabulary is an essential element in University students because it is used to write academic text, mainly writing theses. Academic vocabulary "refers to a set of lexical items that do not core words but are relatively frequent in academic texts" (Paquot, 2010). It commonly appears in a large number of academic texts, such as journal articles, theses/dissertations, research papers, conference papers, academic books, etc. However, the students tend to be unfamiliar with academic vocabulary because it has low frequency than general vocabulary. Consequently, knowing academic vocabulary is a demanding task for students to be acquired.

Scholars around the world have attempted to identify academic vocabulary and compiled them in corpora. A variety of vocabulary list has been categorized into two types: general academic word list and specific

academic word list (Liu & Han, 2015). General academic word list compiled from frequent academic words across disciplines. Thus, the specific academic word list is a more specific area, also known as a technical word list, field-specific academic vocabulary list, discipline-specific academic word list, and discipline-based.

2. Academic Word List

Academic word list (AWL) is developed by Averil Coxhead in 2000. Averil Coxhead teaches English for academic purposes in the School of Linguistics and Applied Language Studies at Victoria University of Wellington.

Up to the present time, the most widely used Academic Vocabulary List is the compilation Academic Word List (AWL) from Coxhead (2000). The AWL focuses on general English language vocabulary that the students need in language learning, particularly in academic writing. The AWL was compiled around 10% of the total words in academic texts but only 1.4% of the total words in a fiction collection of the same size. It consists of at least 15 of the 28 disciplines within four subject areas: arts, commerce, law, and science. The representative of the text, organization of corpora, size, and word selection criteria was considered to develop academic corpora and word lists. Academic words were selected based on three criteria. Firstly, the words family did not come from the first 2,000 most frequently occurring English words from West's (1953) General Services List (*specialized occurrence*). Secondly, a word family member had to

appear ten times in each of the four main sections of the corpus and 15 or more of the 28 subject areas (*range*). Lastly, word family includes in AWL had occurred 100 times in the academic corpus (*frequency*). The corpus analysis program Range was utilized to count the word's frequency and sort the words in the academic corpus. As a result, the AWL contains a 570-word family from 3.5 million running words. Furthermore, the word families were divided into ten sub-lists according to their frequency ranking in the academic corpus. The first sublist consists of the 60 most frequent word families in the AWL, the second sublist contains the next 60 most frequent word families, and so on.

Coxhead's (2000) AWL is widely used in various academic corpora until this day. The coverage of AWL in a specific field and specific disciplines are not the same. Variety of discipline-specific AWL provides high and low coverage: 4.94% AWL in Education Research Articles (Mozaffari & Moini, 2014); 13.86% AWL in Social Science Research Articles (Chanasattru & Tangkiengsirisin, 2017); 10.07% AWL in English Medical Research Articles (Chen & Ge, 2007).

Meanwhile, AWL in the specific field provides high coverage: 11.96% AWL in Applied Linguistics Research Articles (Khani & Tazik, 2014); 11.17% AWL in Applied Linguistics Research papers (Vongpumivitch et al., 2009); 12.82% AWL in Environmental Science (Liu & Han, 2015); 9.06% AWL in Agricultural Research Articles (Martinez et al., 2009); 10.46% AWL in the financial corpus (Li & Qian, 2010); 9.96% AWL in Chemistry Research

Articles (Valipouri & Nassaji, 2013). The coverage of AWL in various disciplines continuously reached around 10%, which relevant to investigating AWL distribution in texts from Coxhead (2011). Whereas, Mozaffari & Moini (2014) reported the low lexical coverage of the distribution AWL in Education Research Articles. It means that pay attention to the education field is very crucial.

Although Coxhead's AWL has a decisive role in academic learning and research and is widely used on English Academic Purposes (EAP) and English Specific Purposes, it has limitations and has been criticized for several issues. Coxhead's AWL was built on West's GSL (1953), which contained a more general old list, as pointed out by Gardner and Davies (2014). All the academic texts are from the early 1960s to the late 1990s. The use of word families also is claimed as a problematic issue in Coxhead's AWL by them. Providing word families in the list of Academic Word is not efficient due to the word that might not share the same core meaning, and the list does not consider grammatical parts of speech to make it easier in using the word. Also, Coxhead's AWL involved general disciplines, which means that the AWL focus is too general but did not cover all the fields. Therefore it is better if the researcher builds Academic words based on specific fields because words from different disciplines and fields have different functions and meanings. Moreover, the distribution of the words from the corpora is not the same (Martinez et al., 2009).

3. Academic Vocabulary List

Academic Vocabulary List was developed by Dee Gardner and Mark Davies in 2013 and published in 2014. They are from the Department of Linguistics and English Language, Brigham Young University. Gardner & Davies (2014) introduced a new Academic Vocabulary List derived from a 120-million-word academic subcorpus of the 425-million-word Corpus of Contemporary American English (COCA; Davies 2012). The corpus is composed of nine disciplines: Education, Humanities, History, Social Science, Philosophy, Religion, Psychology, Law and Political Science, Science and Technology, Medicine and Health, Business, and Finance. All the text in the academic corpus was compiled from the USA. Academic Vocabulary List (AVL) was created by considering four criteria. 1) *Ratio*: The frequency of the word must occur at least 50% higher in the academic corpus than in the non-academic reference corpus of COCA (per million words). 2) *Range*: The word must achieve at least 20% of the expected frequency in at least seven out of nine academic disciplines represented. 3) *Dispersion*: the words must have a dispersion of at least 0.80 to measures that the word occurred evenly in the corpora or superior to the range measures. 4) *Discipline Measures*: a word could not appear more than three times the expected frequency in any of nine disciplines.

The studies showed that the AVL discriminates between academic and other materials and covers 14% of academic materials in both COCA (120 million+ words) and the British National Corpus (33 million+ words).

The corpus can be used in settings where academic English is the focus of instruction. AVL found a new web-based interface that can be used to learn AVL words and to identify and interact with AVL words in any text entered in the search window. The entire list is available at www.academicwords.info in two formats (a lemma version and a word-family version) to meet various academic needs.

Nowadays, only a low substantial of researchers have attempted to involve Gardner and Davies' (2014) AVL in their research. Csomay & Prades (2018) have attempted to investigate academic vocabulary in ESL student paper. They reported that the overall AVL use in ESL student paper was 12.03%. The data have shown that the contribution of AVL in students' papers is highly skewed and useful.

While the previous argument has supported that AVL is effective in students' papers, it must be recognized that AVL also has a limitation. Durrant (2016) stated that AVL is not built on any pre-existing list. The word list directly composed of COCA and considered word selection criteria to make a new academic word list.

Up to the present time, the use of Coxhead's AWL (2000) has been widely used in previous research papers than Gardner and Davies' AVL (2014). Nevertheless, the finding from Qi (2016) supported that AVL was more useful for learners than AWL because AVL contains the lexical sizes of the most frequent BNC/COCA 1,000-8,000 word families. Coxhead's AWL was to develop academic word by considering GSL (West, 1953). The

words that occur in GSL are removed from the list. However, GSL is an out-of-date corpus, therefore Browne and Philips' created a New General Service List (NGSL) by compiling 273 million-word corpus, which larger than GSL. Nowadays, only a few numbers of research try to involve NGSL in the corpus. To fill this gap, this research also compared the words in undergraduate theses corpus by using NGSL.

Thus, the purpose of this study is to use Gardner and Davies' AVL (2014) combined with Coxhead's AWL (2000) to know the academic word list that the student used in writing theses and to develop a new academic word list for undergraduate students in writing undergraduate theses.

4. Corpus-Based Analysis

Corpus (pl. corpora) is a collection of spoken and written language archived on computer software (McKay, 2006). A similar tone is noted in the definition of a corpus, as formulated by McCarthy (2004). In his terms: "A corpus is a collection of texts, written or spoken, usually stored in a computer database." A corpus, then, is simply an extensive collection of texts that we can analyze using computer software, just as we can access the millions of texts on the Internet. It is not a theory of language learning or a teaching methodology. However, it does influence our way of thinking about language and the kinds of texts and examples we use in language teaching" (McCarthy 2004: 7). In essence, a corpus is a word bank that provides an extensive collection of the words in this world that are collected from spoken and written language stored in computer software. However,

corpus in software enables re-arranging the storage so that the investigations of various kinds in language can be made (Hunston 2002: 3).

In the term of corpus-based research, there are two fields especially valuable for developing and assessing L2 textbooks adopted from McKay (2006), they are word frequency lists and concordancing. Those are the common areas that the researchers did in corpus-based research. A frequency list is a list of all the types of words that appear in a corpus, along with the total number of word occurrences. Concordance programs involve the user to bring together all the instances of a particular word and the words surrounding it. The selected word is referred to as the node word/phrase like a noun phrase, verb phrase, etc.

The primary goals of corpus-based research in linguistics are to describe and explain systematic linguistic patterns that are extremely frequent or rare in discourse from a particular are extremely frequent or infrequent in discourse from a particular pointed out by Dauglas (2010). Dash (2018) also noted that corpus-based research purposes are to investigate, describe, apply, and analyze that appropriate with all branches of linguistics. It means that corpus-based research can guide the researcher to obtain the data in linguistics quickly.

Dash (2018) has proposed ten general corpus features to all kinds of language corpus and explained some of them in the explanation below:

1. Quantity

The quantity of the corpus itself determines the authenticity and reliability of the corpus. Nowadays, corpus computer software provides a more significant collection number of the word. The present technology has provided us the liberty to access and increase the quantity of corpus quickly. The quantity of corpus refers to the sum of total linguistic components included in it. Corpus is a lot of word collection representing a variety of languages from various written and spoken sources. Language corpora also have a significant increase in overtimes. At first, the corpus contained just 1 million words; however, nowadays, the total number of language corpora reached 400 million within the last few years. This increase will continue to grow over time from various sources and fields.

2. Quality

Quality of corpus related to authenticity. It implies that the text's accumulation should be from real communication activities in written and spoken language. In collecting the data, the corpus collector has no right to alter, modify, or distort the actual image of the language data collected. In other words, collecting data should follow the principles proposed for the task.

3. Representation

The corpus should involve text from various fields and disciplines to reach representativeness. It should be accurate and balanced from the corpus representing language and collecting it both from written and spoken

text. The representative of the field of the study depends on the purpose of the research itself. According to Leech (1991 cited in Dash 2018), A corpus can be representative only when finding based on an analysis of it can be generalized to the language as a whole or a specified part of the language. Hence, the variety of data represented proportionately from all possible domains of language use better to be emphasized rather than focusing on the quantity of data. Dash (2018) stated that those features are not absolute and changeable. Some features can be redesigned to address the unique form and the content of a specific corpus.

In Corpus-based research, four aspects affect the designing corpus noted by Hunston (2002: 25), some of them have similarity from features pointed out by Dash (2018):

1. Size

The advancing of computer technology increases the size of corpora and makes it easier to be accessed. The store of the computer provides millions of corpora. Dash (2018) revealed that "a corpus is always expected to contain a large number of words and sentences since the basic point of assembling a corpus is to gather data from a variety of sources in a larger quantity of words." The data in a corpus can be overwhelming. Therefore some arguments support that the use of a small corpus can be useful under specific research (Hunston, 2002). In essence, the size or quantity of a corpus should be relevant to the research's purpose. If one intends to examine the language used in a particular genre, such as

academic textbooks, a more limited corpus can be quite helpful (McKay: 2006).

2. Content

Proper research depends on the content's suitability with the purpose of the research, as Hunston (2002) stated. For example, if a researcher wants to determine the authenticity of informal spoken dialogues in an L2 textbook, two corpora are needed: one containing textbook dialogues and the other corpora of spoken, informal conversations (McKay: 2006). It relates to the feature of quality in the corpus pointed out by Dash (2018) that has been explained above. It implies that quality relies on the content of the study.

3. Balance and Representativeness

The specific terms of language should be represented in corpora. An object of study has to build a corpus that is representative of the study. The balance corpus has to consist of an equal number of words. The representative of McKay (2006) stated that "If the purpose of the research is to examine general spoken, informal conversational English, then the corpus should include conversations that demonstrate variation in the gender and age of the speakers, as well as diversity in the situations used." Representativeness also has been explained in the features of the corpus by Dash (2018) above.

4. Computer Software

There are some free computer tools for corpus-based research.

1. AntConc

AntConc is a freeware concordancing software developed by Laurence Anthony (Anthony, 2019). This software is the most widely used today because simple to be used. It can be run on Windows, Macintosh, and Linux, but it only reads the txt file. This program contains several tools: concordance, concordance plot, file view, cluster/N-Grams, collocates, word list, and keyword list. The software available from <https://www.laurenceanthony.net/software>

2. Range Program

The range program is computer software for analyzing the word's frequency in a particular text or group of texts. Paul Nation developed it, and it is handy to be used. However, the tools of this program are very limited, only focus on range and frequency. This program available on <https://www.wgtn.ac.nz/lals/resources/vocabulary-analysis-programs>

3. Lextutor

The lexical tutor is an analysis corpus tool that can be used on the web site. This tool is appropriate for vocabulary learning because many tools can be used by teachers and learners to use a corpus in class for research-based learning (McKay: 2006). This tool is provided on <https://www.lex tutor.ca/>

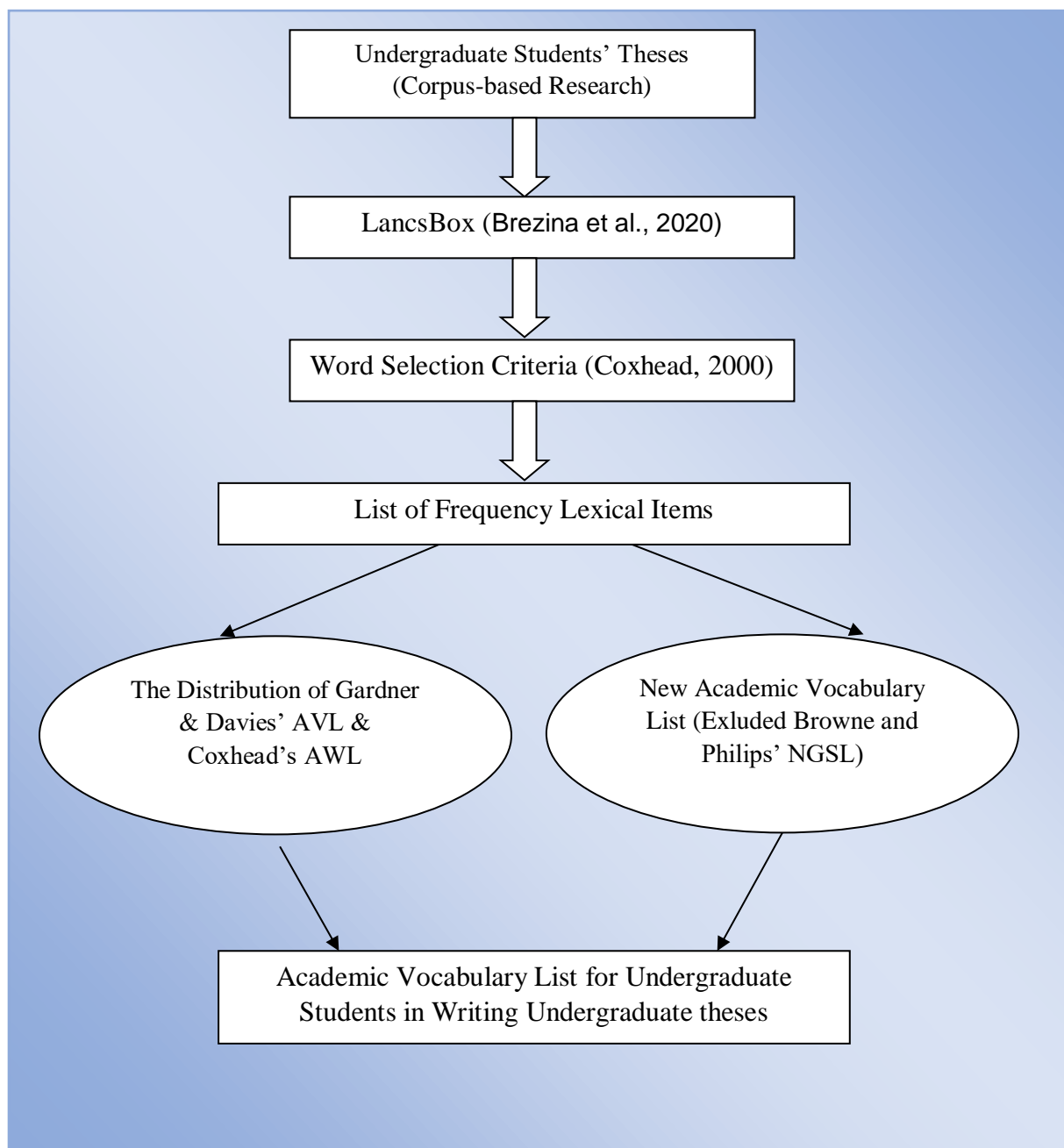
4. Lancsbox

Lancaster Corpus Tool Box is a new software language analysis developed by Lancaster University. This software is free and provides many

tools to analyze the language used. The main features of this tool are works with the data or existing corpora. It can be used by linguists, language teachers, historians, sociologists, educators, and anyone interested in language; visualizes language data; analyses data in any language; automatically annotates data for part-of-speech; works with any primary operating system (Windows, Mac, Linux) (Brezina et al., 2020). The result of this tool will report directly on document form. The software is available on <http://corpora.lancs.ac.uk/lancsbox>.

C. Conceptual Framework

The concept of this research is conducted by compiling undergraduate students' theses. All of the theses are imported to LancsBox software and analyzed based on Coxhead AWL's criteria. The conceptual framework is provided below:



CHAPTER III

RESEARCH METHODOLOGY

This chapter presents the research design, the sample of the study, research instrument, data collection procedures, and data analysis procedures of this research.

A. Research Design

The design of this study is quantitative research by using a corpus-based approach. A corpus-based analysis is an approach that involves an extensive collection of the text that can be analyzed by using computer software (McKay, 2006; McCarthy, 2004). This corpus-based approach focuses on the frequency, and distribution of academic vocabulary and develop new academic vocabulary in undergraduate students' theses. The frequency of the words on the text appears in a software program. It was identified based on word selection criteria to determine the distribution of AWL and AVL and the new academic vocabulary in undergraduate theses.

B. Population and Sample

The population of this study is all of the open-access undergraduate theses from 10 universities in Indonesia majoring in the English Education Study Program during the period of 2016 to 2020. They are *Universitas Cokroaminoto Palopo; Universitas Muhammadiyah Makassar; IAIN Ponorogo; UIN Sunan Ample; UIN WaSlisongo; UIN Syarif Hidayatullah; Universitas Negeri Yogyakarta; IAIN Tulungagung; UIN Makassar; and Universitas Kristen Satya Wacana*. 1784 undergraduate