DISTRIBUTIONAL CORPUS ANALYSIS OF KOREAN NEOLOGISMS USING ARTIFICIAL INTELLIGENCE

by

WONBIN KIM

(Under the Direction of William A. Kretzschmar, Jr.)

ABSTRACT

Distributional corpus analysis (DCA) is an approach which reveals lexical relations using large-scale corpora and computational techniques in natural language processing. It has an advantage of processing and analyzing lexical relations in a quantitative, consistent, and objective way. Although the DCA approach allows analysts to process large-scale linguistic data efficiently, there are few studies using the DCA approach to investigate language phenomena within corpus linguistics. Therefore, this study aims to bridge the gap between the DCA approach and corpus linguistics by designing and describing DCA from the perspective of corpus linguistics. Specifically, this study uses the DCA approach to analyze the distributional behaviors of three Korean neologisms *leyal*, *lwuce*, and *kay-* and track semantic change of the three neologisms. For the analysis of distributional behaviors, Korean Twitter data spanning about ten years is collected and three state-of-the-art techniques are employed. For *leyal*, word2vec and cosine similarity are used and for *lwuce*, Latent Dirichlet Allocation is employed. For *kay-*, long short-term memory is utilized. Regarding *kay-*, its connotational and attitudinal meaning is investigated. The results from DCA show that (i) between the two meanings of *leyal*, 'really' has always been more dominant than 'Real Madrid', (ii) between the new and existing meanings of *lwuce*, the existing meaning

has always been more dominant and the use of the new meaning most significantly decreased in 2015, and (iii) the semantic prosody of *kay-* has shifted from negative toward positive. This study has made several "first attempts". First, this work is the first study using artificial intelligence and Korean social media data to analyze the distributional behaviors of Korean neologisms and track their semantic change over time. Secondly, this work is the first study showing DCA from the perspective of corpus linguistics. Thirdly, this work has established specific methods to validate the DCA approach using a collocation analysis in corpus linguistics for the first time. This study making several "first attempts" will be able to encourage interdisciplinary research between corpus linguistics and artificial intelligence as well as function as a foundational study upon which further DCA studies can build in corpus linguistics.

INDEX WORDS: distributional corpus analysis, artificial intelligence, Korean neologisms, semantic change, collocation analysis, distributional frequency profiles

DISTRIBUTIONAL CORPUS ANALYSIS OF KOREAN NEOLOGISMS USING ARTIFICIAL INTELLIGENCE

by

WONBIN KIM

BA, Pusan National University, South Korea, February 20, 2009

MA, Pusan National University, South Korea, February 18, 2011

A Dissertation Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

DISTRIBUTIONAL CORPUS ANALYSIS OF KOREAN NEOLOGISMS USING

ARTIFICIAL INTELLIGENCE

by

WONBIN KIM

Major Professor:	William A. Kretzschmar, Jr.
Committee:	Linda Harklau
	Paula J. Mellom

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2022

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and respect to my advisor, Prof. Kretzschmar for his support and encouragement. His comments and advice on my doctoral research stimulated me to study my major in more depth and helped me to gain an insight into my research and grow as a professional linguist. His guidance during the period of my doctoral studies was really helpful for me to be equipped with the qualifications of an independent researcher. Also, I am grateful to the other two committee members, Prof. Harklau and Prof. Mellom for their encouragement and precious comments on my research.

Moreover, I cannot thank Prof. Yi at the department of comparative literature enough for hiring me as a Korean teaching assistant. Thanks to the teaching assistantships, I was able to continue studying without financial difficulties and build a teaching career. While teaching Korean, I could learn a lot including teaching skills. The teaching experience at UGA is an invaluable experience to me and I believe that it will be very useful when I get a teaching job.

Last but not least, I am deeply grateful to my family for their encouragement, support, and love. Especially, their faith in me gave me confidence and helped me to continue my studies. They provided another motivation to keep going. I could not have successfully completed my PhD and dissertation without them. I dedicate my doctoral dissertation to my grandmother, parents, and twin sister.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

In order to study word meaning, it is essential to examine the syntagmatic environment where the target word occurs. This is because words co-occurring with the word under scrutiny can provide much significant information on the meaning and properties of the target word. For example, when we distinguish two homonyms, their different contexts help us to differentiate one word from the other. The famous quotation from Firth, "you shall know a word by the company it keeps" (Firth 1957a: 11), supports the idea that it is necessary to analyze co-occurring words for the study of word meaning.

This study analyzes the context words of three Korean neologisms to investigate semantic change of those Korean neologisms. Specifically, this study tracks how the meanings of those Korean neologisms have developed over ten years by analyzing the distributional behaviors of the three neologisms. For a more efficient analysis of the distributional behaviors, I will use the approach of distributional corpus analysis.

Distributional corpus analysis (DCA) analyzes actual words in the contexts where the word under consideration occurs by means of computational techniques in natural language processing, in particular, word space models in order to reveal lexical relations centered around the target word. Because word space models structure the meanings of words in a vector space by converting words into context vectors based on the contexts of each word, it is possible to compute the relations of the target word to other words and process lexical relations in a quantitative, consistent, and objective way (Navigli 2009, Geeraerts 2010). DCA enables us to study meaning coming from the

combinations of words in the context systematically with sophisticated techniques and large-scale data.

Applying the DCA approach to Korean Twitter data, I analyze the context words of the three Korean neologisms *leyal* (*레알*), *lwuce* (*루저*), and *kay-* (*개-*)[1] and track their semantic change over ten years. As demonstrated in the statement from Firth (1935: 37), "the complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously", the analysis of the distributional behaviors of the three neologisms will be helpful for us to have a better understanding of the meanings of the neologisms.

1.1 Korean Neologisms

When new words are created, their frequency is low and their use is limited in the beginning, but as their frequency gets higher with time they are used in different contexts, resulting in change in their meanings. Considering this semantic characteristic of neologisms, this study targets neologisms. To be specific, I track the semantic change of three Korean neologisms *leyal*, *lwuce*, and *kay-* over ten years. Concerning the neologism *kay-*, its connotational meaning created by context is examined unlike the other two neologisms. The following three subsections introduce the three neologisms and describe what and how to study in brief.

1.1.1 Neologism *Leyal*

The neologism *leyal* appeared in late 2009. It derived from the English word *real* but its pronunciation is different from English. The neologism *leyal* means 'really' but depending on the contexts, it is used to refer to Real Madrid, which is a Spanish professional football club. That is,

[1] The three Korean neologisms *leyal*, *lwuce*, and *kay-* are romanized Korean. All the romanized Korean words in this dissertation follow the Yale system of romanization.

it is used as a contracted form of Real Madrid. In terms of these two meanings, it is examined which meaning is more dominant and how each of the two meanings has developed across time.

Specifically, the context words of *leyal* are compared with those of two alternative Korean words representing the two meanings, i.e., *cincca* and *leyalmatulitu* representing 'really' and 'Real Madrid' respectively. If the context words of *leyal* overlap with those of *cincca* more than *leyalmatulitu*, it means that the meaning of 'really' is more dominant. It is natural that similarity between the context words of two words is associated with semantic similarity of the two words because the contexts of semantically similar words are similar to each other (Harris 1954). This study looks into how much the context words of *leyal* overlap with those of each alternative word from year to year.

1.1.2 Neologism *Lwuce*

The word *lwuce* came from the English word *loser* and its meaning is similar to English. It means a person who is defeated or unsuccessful. However, in a Korean TV show aired in November 2009, a woman said that those men whose stature is below 180 centimeters (about 5 feet 11 inches) are losers. Back then, her remark became a hot issue and she was denounced for what she said. Since her remark, *lwuce* has been used to refer to a man whose stature is below 180 centimeters. The neologism *lwuce* is a case where a new meaning is added to the existing meaning.

To find out how the new and existing meanings of the neologism *lwuce* have developed across time and which meaning holds a predominant position, change in the use of each meaning is investigated. If the use of the new meaning gets more frequent with time and it becomes more than that of the existing meaning, it means that the new meaning has taken over from the existing meaning as the dominant meaning of *lwuce*. For the analysis of the use of each meaning, clustering is employed. Clustering used in this study classifies the context words of *lwuce* by topics. This

study examines whether the use of each meaning increases or decreases by tracking the number of topics associated with each meaning over time.

1.1.3 Neologism *Kay-*

The neologism *kay-* stemmed from the prefix *kay-* in the Korean language. The existing prefix is attached to nouns and has three senses. One of the three senses is ‘severe’ or ‘heavy’. When the prefix is attached to nouns with negative meaning, it intensifies the degree of negative meaning from nouns. In contrast, the neologism *kay-* is attached to adjectives or verbs with negative meaning. It means ‘really’ or ‘very’. It functions as an adverb which modifies adjectives or verbs but it cannot stand alone as the existing prefix. Also, it emphasizes the negative meaning of adjectives or verbs where it is attached as the existing prefix.

The neologism *kay-* has been mainly used among many teenagers as slang. Interestingly, combinations of the neologism and a positive word have frequently occurred recently. This suggests that there is change in its connotational and attitudinal meaning, i.e., semantic prosody of the neologism *kay-*. Therefore, based on the growing number of combinations of the neologism and a positive word, this study investigates how the semantic prosody of the neologism *kay-* has changed over time. For the exploration of semantic prosody, adjectives and verbs where the neologism is attached are sorted into positive and negative and the token frequencies of those positive and negative words are counted from year to year. The analysis of the token frequencies will show how the semantic prosody of the neologism *kay-* has changed.

1.2 Objective of this Study

Language keeps changing over time. Language is like an organism which gets old in that it is not immutable and unchangeable. As language is affected by its environment and incessantly interacts

with its surroundings, language continues to change. Therefore, linguists need to study how the language under consideration changes to get a better understanding of that language. Considering the importance of research on language change, in this study I focus on semantic change and examine how the meanings of words change over time. According to Sinclair's statement that word meaning is heavily dependent on context and is to be analyzed through the lexical and grammatical elements co-occurring with the target word (Sinclair 1991: 108), I will investigate the semantic change of words by analyzing the distributional behaviors of the words. To be specific, this study tracks how the meanings of three Korean neologisms *leyal*, *lwuce*, and *kay-* have changed over the past ten years from 2010 to 2019[2]. To analyze the distributional behaviors of the three neologisms, this study uses the DCA approach. The DCA approach has advantages of handling large-scale corpora efficiently and analyzing lexical relations in a quantitative way.

However, despite the advantages of DCA, there have been few studies employing the DCA approach to examine linguistic phenomena within the field of corpus linguistics. Studies using the DCA approach have been mainly carried out in computational linguistics and computer science because of the nature of methods. Therefore, this study aims to construct a bridge which can connect the DCA approach to corpus linguistics. In order to achieve this goal, I will design and describe this study using the DCA approach from the perspective of corpus linguistics. I will present how to apply the DCA approach to actual language data to analyze the distributional behaviors of the target words in detail and use methodology in corpus linguistics for method validation of the DCA approach. The detailed descriptions on the entire process from data

[2] This study constructs corpora based on tweets from 2010 to 2019 to investigate semantic change of the three neologisms but this does not mean that the three neologisms appeared in 2010. Please note that they may well have existed before 2010 and this study simply tracks their semantic change during a specific period.

collection to data preprocessing to data analysis will be helpful for corpus linguists who want to apply the DCA approach to their own corpora but have no background in that approach.

1.3 Organization of Chapters

In Chapter 2, I cover five theoretical frameworks functioning as the backbone of this study. The five theoretical frameworks have had a strong influence on designing this study. The first framework is corpus linguistics. I explain what a corpus is, what conditions are needed for the construction of a corpus, what types of corpora there are, and when and how the first modern electronic corpus was constructed. The second framework is distributional corpus analysis. I will put a main focus on distributional semantics, which is closely related to the methodology of this study. I introduce previous studies which have been performed in computer science and computational linguistics. The third framework is lexical semantics. As this study investigates how word meanings develop over time by analyzing the distributional behaviors of the words, I provide theoretical descriptions of lexical semantics and lexical relations. The lexical relations include collocation, colligation, semantic prosody, and semantic preference. The fourth framework is usage-based linguistics. According to an aphorism of usage-based linguistics, meaning comes out of language use. This aphorism is connected to the analysis of semantic change of the three neologisms in that the ultimate goal of the analysis is to demonstrate that the meaning of a word is determined by how people use the word. One of the important notions in usage-based linguistics is frequency. I will give explanations of type and token frequency. Token frequency is to be used for the analysis of the neologism *kay-*. The fifth framework is language as a complex system. I describe what a complex system is and why language (i.e., speech) is considered to be a complex system. One of the characteristics of language as a complex system is associated with the frequency profile of linguistic features, i.e., A-curve. I give a detailed description of A-curve because it is

used for method validation of computational techniques used for the analysis of each neologism (the concept of A-curve is applied to the selection of some collocates from a collocate list in a collocation analysis).

Chapter 3 provides background knowledge on artificial intelligence and natural language processing so that those who do not have any knowledge in those fields can better understand this study. Before I give explanations of them, I will introduce Korean corpora with focus on two main Korean corpora, i.e., Yonsei corpora and Sejong corpora. After going over the existing Korean corpora constructed in South Korea, I will show how to build Korean Twitter corpora using Python. Because the existing Korean corpora do not contain up-to-date data, I could not obtain enough data about the three neologisms from them. For research on the three neologisms, I collected Twitter data and constructed Twitter corpora myself. The construction of a corpus requires data collection and preprocessing. Data preprocessing includes tokenization, tagging, normalization, stemming, cleaning, and removal of stop words. The process of data preprocessing is carried out by means of KoNLPy in this study. After explaining how to build Korean Twitter corpora, I give brief descriptions of artificial intelligence and natural language processing. I describe what they are and introduce computational techniques to be employed to analyze the distributional behaviors of the three Korean neologisms: word2vec and cosine similarity, Latent Dirichlet Allocation, and long short-term memory. In the last section, I cover two model evaluation metrics, i.e., confusion matrix and precision-recall curve. They are used to evaluate the performance of long short-term memory.

Chapter 4, Chapter 5, and Chapter 6 show how to apply DCA to Korean Twitter data to track the semantic change of the three neologisms. I give details on every process from data collection to preprocessing to analysis for each neologism. The details include how and how much amount of Korean Twitter data was collected, how the scraped Twitter data was preprocessed, and

how models were trained with the preprocessed data. Based on the results from those models, I examine how the meanings of the neologisms have developed across time. With regard to long short-term memory trained in supervised learning unlike the other models, the section of evaluation is added. The section covers metrics to assess the performance of long short-term memory (i.e., validation accuracy, confusion matrix, and precision-recall curve). The most notable sections in the three chapters are the ones about method validation. The sections evaluate the suitability of each computational technique for semantic change research and the reliability of results from those techniques through the comparison of results between the employed techniques and method validation. Similar results between them will be able to demonstrate that the DCA approach has great potential as a systematic methodology for lexical semantic research. For method validation, not other computational techniques but more transparent and intuitive linguistic methods are used. Specifically, collocation analyses in corpus linguistics are used. In each discussion section, I explain what the results from the DCA approach mean, discuss the advantages and limitations of analyses using the DCA approach, and cover what further studies are needed in order to overcome such limitations. Lastly, Chapter 7 provides an overview, significance of this work, and conclusion. The overview sums up what this study has done. For significance, I mention in what aspects this study is significant and what contributions it can make. For the conclusion, I cover the advantages and limitations of the DCA approach and discuss what we should note concerning the use of the DCA approach.

CHAPTER 2

THEORETICAL FRAMEWORK

This chapter covers five theoretical frameworks which have had a strong influence on designing this study: (i) corpus linguistics, (ii) distributional corpus analysis, (iii) lexical semantics, (iv) usage-based linguistics, and (v) language as a complex system. Corpus linguistics, distributional corpus analysis, and language as a complex system have inspired the methodology used here for the analysis of distributional behaviors of the three Korean neologisms. Especially, language as a complex system has been applied to method validation of the DCA approach. In contrast, lexical semantics has given motivation to study word meaning. With that motivation, this study aims to track semantic changes of the three new words over time. Among lexical relations, collocation was used for method validation of the computational techniques employed to analyze the distributional behavior of the two neologisms *leyal* and *lwuce*. Semantic prosody was employed to reveal the semantic change of the neologism *kay-* more clearly. Lastly, usage-based linguistics is associated with the objective of this study. The ultimate purpose of this study is to support the aphorism of usage-based linguistics that meaning comes out of language use. I will show that the meaning of a word is determined by how people use the word. It will be able to function as empirical evidence proving that aphorism and make a contribution to supporting usage-based linguistics. The following sections provide specific descriptions of each theoretical framework.

2.1 Corpus Linguistics

A corpus is a collection of texts sampled from produced speech and writings. With the development of computer technology, its definition has been changed into a large set of authentic written or spoken texts saved in a computer readable format which can be representative of a particular language or situation of language use (e.g., Francis 1992, Atkins, Clear & Ostler 1992 for the definition of a corpus). A corpus allows linguists to observe how people actually use language and analyze such real language data. In particular, linguists can keep track of language change across time through monitor and diachronic corpora. The first modern computer readable corpus, the Brown Corpus of Standard American English was constructed in the early 1960s and the term "corpus linguistics" first appeared in the early 1980s.

Corpus linguistics is the study of language based on corpora. It involves constructing corpora of authentic language data and using such corpora to explore language. It provides empirical methods to research many aspects of real language in everyday life. Although corpora provide linguists with good resources of language data, Noam Chomsky argued that a corpus is not proper for language study because it is full of performance-related errors (Chomsky 1965). He introduced a binary approach to language: linguistic competence and linguistic performance. He suggested that linguists seek to delve into linguistic competence, i.e., internalized knowledge of a language. However, as computer-based corpus data became available, studies using corpora started to boom from 1980, leading corpus linguistics to mature methodologically and making schools of corpus linguistics like the neo-Firthians grow (McEnery & Wilson 2001).

Corpus linguistics is closely related to usage-based linguistics in that it can present empirical evidence proving the argument of usage-based linguistics that meaning is use and language structure emerges from language use (e.g., Caldwell-Harris, Berant & Edelman 2012,

Caines 2012). Moreover, corpus linguistics is linked to cognitive linguistics. Frequency is known to have an impact on language processing and learning (Bybee & Hopper 2001, Ellis 2002). The higher the usage frequency of linguistic structures, the stronger their degree of cognitive routinization (i.e., entrenchment) gets. The correlation between frequency and entrenchment is well supported by corpus-derived findings on frequency, which implies that corpus linguistics can be used for cognitive research as well.

Corpus linguistics is divided into two major approaches: corpus-based and corpus-driven approaches. This distinction was introduced by Tognini-Bonelli (2001). The corpus-based approach regards corpus linguistics as a methodology to research language. It uses corpora to analyze pre-defined linguistic features and test existing theories or hypotheses to validate, contradict, or refine the established theories or hypotheses. As linguists backing up the corpus-based approach consider the use of corpora as a methodology, they argue that the corpus-based approach can be applied to almost every branch in linguistics (e.g., Sinclair 1991, Biber 1993, Kjellmer 1994). Also, they are in favor of corpus annotation using grammatical categories, which is a main difference from the corpus-driven approach.

The corpus-driven approach rejects the definition of corpus linguistics as a method. It argues that corpus linguistics should be recognized as a completely new independent paradigm. That is, it aims to build its own theory based on corpus data alone without using existing linguistic theories. Linguists supporting the corpus-driven approach object against corpus annotation because the grammatical categories for tagging came from pre-set theories. They try to apply a holistic approach to language description entirely relying on corpus data, with no distinction between lexis, syntax, semantics, pragmatics, and discourse. Also, they claim that the large-scale size of a corpus necessarily accompanies corpus balance and representativeness because such a

huge corpus can balance itself when it grows to be large enough to achieve representativeness. However, this argument is controversial and it needs clear verification.

In fact, it is hard to strictly distinguish between the corpus-based and corpus-driven approaches because the boundary between the two approaches is fuzzy (McEnery, Xiao & Tono 2006). Kübler & Zinsmeister (2015) say that corpus linguistics can be both a theory and a tool and which one corpus linguistics becomes depends on how it is applied. Therefore, it would be desirable to use the two approaches flexibly rather than sticking to either one.

2.1.1 Conditions for Corpus

Representativeness, balance, and sampling are essential issues that must be considered in the construction of a corpus. As it is impossible to investigate every utterance or written sentence of a particular language, collecting a sample of data is inevitable. For sample collection, it is important to check whether the sample can represent the given language. Unless the characteristics of samples hold for the given language, the samples cannot be said to be representative. In order to secure representativeness, balance (i.e., the range of genres) and sampling (i.e., how texts are selected for each genre) are critical.

With regard to balance, the range of text categories determines how balanced a corpus is. To make a balanced corpus, as many text categories as possible should be proportionally contained in a corpus. In particular, the representativeness of a general corpus featuring an overall description of a particular language or language variety heavily depends on a broad range of text types. The British National Corpus, which represents spoken and written British English of the late twentieth century as a 100-million-word text corpus, is an example of a general corpus.

Concerning sampling, sampling units and the boundaries of a population have to be defined. To take an example of written texts, a sampling unit means a book, a newspaper, a periodical, and

so on. A population refers to the assembly of all sampling units. Decision on the boundaries of a population relates to deciding what texts to include and exclude from the population.

Another decision to be made about sampling is on sample size. For a corpus of written language, this relates to whether to collect full texts or text segments. In general, a large-sized corpus contains full text samples, which are used for discourse analysis or study on textual organization. However, the collection of full text samples might be able to bring about copyright issues. Moreover, it might prevent a corpus from being a balanced corpus when corpus size is finite since it makes a collector collect samples from relatively fewer text categories. Hence, it is usually recommended to collect text segments when corpus size is finite.

The last issue that has to be considered in sampling is the proportion and number of samples for each text category. The weight of samples for each genre should be determined by the population under consideration. That is, samples should be collected in proportion to their frequencies in the target population. Researchers should consider these conditions about balance and sampling and try to make their own corpora representative of the language or language variety under scrutiny.

To take an example of the British National Corpus (BNC), the BNC consists of written texts and spoken data, which account for ninety percent and ten percent of the BNC respectively. For written texts, they were sampled by the following three criteria: domain, time, and medium. Domain is the subject field of text, time refers to the period during which text was produced, and medium means the type of text publication, i.e., sampling units. Table 2.1 shows the written text composition of BNC depending on the three criteria.

Table 2.1 Written text composition of BNC depending on the three criteria

Domain	%	Date	%	Medium	%
Imaginative	21.91	1960-74	2.26	Book	58.58
Arts	8.08	1975-93	89.23	Periodical	31.08
Belief and thought	3.40	Unclassified	8.49	Misc. published	4.38
Commerce/Finance	7.93			Misc. unpublished	4.00
Leisure	11.13			To-be-spoken	1.52
Natural/pure science	4.18			Unclassified	0.40
Applied science	8.21				
Social science	14.80				
World affairs	18.39				
Unclassified	1.93				
Total	99.96		99.98		99.96

The specific sampling units include regional and national newspapers, published research journals, published and unpublished periodicals from various academic fields, fiction and non-fiction books, other published and unpublished materials such as brochures and letters, essays written by students, speeches, scripts, and many other types of texts. With regard to sample size, the target sample size for books was 40,000 words and no extract in the corpus exceeded 45,000 words. For texts shorter than the target size, they were included in the corpus after being reduced by ten per cent for copyright reasons.

The spoken data in the BNC was collected using the following two criteria: demographic and context-governed components. The demographic component is composed of transcriptions of spontaneous natural conversations in different contexts produced from volunteers who were selected by age group, sex, social class, and geographical region. The context-governed component consists of transcriptions of recordings made in formal settings such as meetings and lectures. Table 2.2 shows the spoken data composition of BNC depending on the two criteria.

Table 2.2 Spoken data composition of BNC depending on the two criteria

Region	%	Interaction type	%	Context-governed	%
South	45.61	Monologue	18.64	Educational/informative	20.56
Midlands	23.33	Dialogue	74.87	Business	21.47
North	25.43	Unclassified	6.48	Institutional	21.86
Unclassified	5.61			Leisure	23.71
				Unclassified	12.38
Total	99.98		99.99		99.98

The two columns of region and interaction type concern both the demographic and context-governed components while the column of context-governed only applies to the context-governed component. For spoken language, it can be particularly difficult to define a population because there are no sampling frames (i.e., the lists of sampling units) available for reference (McEnery, Xiao & Tono 2006).

2.1.2 Types of Corpora

There are many different types of corpora and it is possible that one corpus has the characteristics of different types of corpora. In this section, a total of ten types of corpora are covered, which are commonly classified corpora in corpus linguistics. Table 2.3 shows the characteristics and example corpora for each type (the example corpora focus on English corpora).

Table 2.3 Characteristics and example corpora depending on corpus types

Type	Characteristic	Example
General	It contains texts from a number of different domains of spoken and written language. Its size is mostly so large that findings from it can be generalized. It provides general information on the language under scrutiny over a specific time span. A large general corpus can function as a reference corpus against which language varieties, particularly held in specialized corpora can be examined.	- American National Corpus - British National Corpus - Corpus of Contemporary American English
Specialized	It contains texts from a certain type/genre/register, or a specific time/context. In general, the texts are limited to one or a few domains, topics, or subject areas. It is useful for detailed research on a particular language or language variety. Its size can be large or small.	- Air Traffic Control Speech Corpus - Child Language Data Exchange System Corpus - International Corpus of Learner English - Lampeter Corpus of Early Modern English Tracts - Nottingham Health Communication Corpus - Michigan Corpus of Academic Spoken English - Michigan Corpus of Upper-level Student Papers - Uppsala Student English Corpus
Monitor	It is a corpus that keeps growing by including new texts on a regular basis, which aims to monitor language change over time. Due to the continuous addition of new texts, the relative proportions of different types of materials may vary. It covers texts from a relatively short span of time, compared to a diachronic corpus. It can be used to keep track of neologisms.	- Bank of English - Corpus of Contemporary American English
Balanced	It is a sample corpus which is representative of a particular language or language variety over a specific time period. It seeks to collect samples from a wide range of text categories for representativeness. The proportions of samples for each text category are determined according to a specific sampling frame which defines the population under consideration.	- Australian Corpus of English - British National Corpus - Brown University Standard Corpus of Present-Day American English - Freiburg-Brown Corpus of American English - Freiburg–LOB Corpus of British English - Lancaster-Oslo-Bergen Corpus

		- Kolhapur Corpus - Wellington Corpus of Spoken New Zealand English
Parallel	It contains the same texts translated into two or more languages. Because the translated texts are aligned, it is easy to compare languages and identify the translation equivalents in the other languages for a particular word in one language.	- Arabic English Parallel News Corpus - English-Norwegian Parallel Corpus - Open source Parallel Corpus
Comparable	It contains texts from the same domain in two or more languages. The texts are not translations of each other. The texts are collected along similar parameters. Corpora containing different varieties of the same language are also considered as comparable corpora.	- CorTec Corpus - International Corpus of English - International Corpus of Learner English
Diachronic	It contains texts from different/consecutive time periods, preferably comparable materials. It shows how language changes over time through texts from a relatively longer period of time than a monitor corpus.	- A Representative Corpus of Historical English Registers - Corpus of Contemporary American English - Helsinki Corpus of English Texts - Time Magazine Corpus
Multimedia	It contains multimedia materials such as audio/video recordings and transcriptions. It can be used for research on various aspects of language (e.g., prosody, speech, and non-linguistic gestures).	- System Aided Compilation Open and Distribution of European Youth Language Corpora
Learner	It contains written and/or spoken data from students who are learning a language, i.e., second or foreign language learners. It shows errors that learners frequently make. It can be usefully applied to the field of foreign language education.	- International Corpus of Learner English - Standard Speaking Test Corpus
Pedagogic	It contains language used in educational settings. It consists of academic textbooks, audio-visual materials, written texts/spoken transcripts in classroom settings, and so on. It can be used to examine teacher-student dynamics, or to develop self-reflective tools for teachers.	- BACKBONE Corpora - System Aided Compilation Open and Distribution of European Youth Language Corpora

In addition to the corpora described above, there are a number of different types of corpora depending on specific purposes. They include spoken and written corpora depending on the mode of data, monolingual and multilingual corpora depending on the number of covered languages,

reference and target corpora depending on what the objective for comparison is, raw and tagged corpora depending on whether annotation is present or not, paralinguistic and sign language corpora depending on targets under scrutiny, and so on.

2.1.3 The First Modern Electronic Corpus

The first modern computerized corpus is the Brown Corpus of Standard American English (abbr. the Brown Corpus), which is of great significance in corpus linguistics because it ignited the growth of that field. Before the early 1960s when the first computer-readable corpus was constructed, the corpus approach was severely criticized because corpora lacked representativeness due to their small size. It was impossible to construct massive corpora at that time because data collection was performed manually. However, with the development of computer technology, it has become feasible to make massive representative corpora. Starting with the Brown Corpus, many computer-readable corpora have been built.

The Brown Corpus was constructed by Henry Kučera and W. Nelson Francis at Brown University in the United States in the early 1960s. It consists of one million words from American English texts published in the United States in 1961. For tagging the words, eighty parts of speech and special indicators were used. The texts were collected from 500 samples over fifteen different text genres and the number of texts in each category did not have a balance. Table 2.4 shows the detailed composition of the Brown Corpus (see W. N. Francis & H. Kucera 1979 for the Brown Corpus manual).

Table 2.4 Detailed composition of the Brown Corpus

Category	Genre (Code)	Number of Texts	Total Tokens	Percentage of Tokens
INFORMATIVE	Press: Reportage (A)	44	88,000	8.8%
	Press: Editorial (B)	27	54,000	5.4%
	Press: Reviews (C)	17	34,000	3.4%
	Religion (D)	17	34,000	3.4%
	Skills and Hobbies (E)	36	72,000	7.2%
	Popular Lore (F)	48	96,000	9.6%
	Belles-Lettres, Biography, Memoirs, etc. (G)	75	150,000	15.0%
	Miscellaneous (H)	30	60,000	6.0%
	Learned (J)	80	160,000	16.0%
IMAGINATIVE	General Fiction (K)	29	58,000	5.8%
	Mystery and Detective Fiction (L)	24	48,000	4.8%
	Science Fiction (M)	6	12,000	1.2%
	Adventure and Western Fiction (N)	29	58,000	5.8%
	Romance and Love Story (P)	29	58,000	5.8%
	Humor (R)	9	18,000	1.8%
Total		500	1,000,000	100.0%

Although the Brown Corpus is considered a small corpus now, it is very significant in that it pioneered the field of corpus linguistics. It not only provided much empirical evidence such as frequency and the usage of English words but also laid a foundation for the construction of many later corpora: the Lancaster-Oslo-Bergen Corpus of British English, the Kolhapur Corpus of Indian English, the Wellington Corpus of New Zealand English, the Australian Corpus of English, the Freiburg-Brown Corpus of American English, and the Freiburg–LOB Corpus of British English.

With regard to spoken language, the first computer readable corpus of transcribed spoken language is the spoken Montreal French Corpus, which was directed by Gillian Sankoff and Henrietta Cedergren at the University of Montreal in 1971. The spoken Montreal French Corpus

is a corpus of one million words and consists of 120 recorded interviews from speakers selected in a balanced distribution in terms of variables such as gender and age. About 8000 words were collected per interviewee (see Sankoff, Lessard & Truong 1977 for more detailed information about the spoken Montreal French Corpus).

It is relatively harder to collect data of spoken language than written language because it requires more processes such as recording and transcribing, which are costly and time-consuming. However, corpora of spoken language are very important in fields where the analysis of utterances is necessary like phonetics, dialectology, and conversation analysis. Also, corpora of spoken language enable researchers to observe and analyze how language variation occurs depending on time and region. Therefore, the construction of corpora of spoken language should not be neglected.

2.2 Distributional Corpus Analysis

Geeraerts (2002) describes the development of lexical semantics with the following four approaches: (i) prestructuralist semantics, (ii) structuralist and neostructuralist semantics, (iii) generativist and neogenerativist semantics, and (iv) cognitive semantics. Distributional corpus analysis (DCA) belongs to the current version of neostructuralist semantics, which is a continuation of structuralist semantics. Neostructuralist semantics serves as one of the contemporary theoretical currents in lexical semantics with neogenerativist semantics and cognitive semantics.

Structuralist semantics puts stress on semantic structures. For the study of meaning, semantic structures should be analyzed and meaning should be explored in a linguistic way because linguistic structures determine the meaning of a language sign (Weisgerber 1927). Semantic structures have been defined as the following three lexical relations by structuralist semanticians (Geeraerts 2002): (i) the relationship of semantic similarity, (ii) lexical relations such

as synonymy, antonymy, and hyponymy, and (iii) syntagmatic lexical relations. The third type of lexical relations has developed into a collocational analysis in neostructuralist semantics, following Firth (1957b). DCA is a kind of collocational analysis, which analyzes co-occurrences[3] of the target lexical item in its context.

DCA employs computational techniques and large-scale corpora to investigate the distributional behaviors of lexical items. DCA analyzes actual words in the context in which the word under consideration appears to reveal the distributional patterns of the target word. It is a bottom-up, data-driven, and usage-based approach to lexical relations (Dobric 2013). It is closely linked to quantitative distributional models, that is, word space models in natural language processing and information technology. A word space model structures the meaning of a word in a vector space based on the context in which the target word appears, making it possible to compute the relations of the target word to other words through context vectors (Navigli 2009, Geeraerts 2010).

More specifically, the DCA approach makes use of distributional representations (i.e., vectorial representations) from word space models to analyze the distributional patterns of lexical items quantitatively. A distributional representation is a concept produced in the research area of distributional semantics in natural language processing. Distributional semantics aims to represent the meanings of words/sentences as distributional representations. The following subsection covers distributional semantics directly related to the methodology of this study in more detail.

[3] The term "co-occurrences" mentioned in this dissertation means lexical items co-occurring with the target word.

2.2.1 Distributional Semantics

Distributional semantics is a subfield of natural language processing, which is a research area that develops methods to quantify semantic similarities among linguistic items based on their distributional patterns, and studies related theories. Distributional semantics represents the meanings of words/sentences as distributional representations in a vector space according to the distributional hypothesis that words sharing contexts have similar meanings (Harris 1954). To produce such distributional representations, distributional semantic models need training, which allows them to learn word meaning from large-scale corpora. A number of studies in computational linguistics and computer sciences have employed distributional semantic models to investigate semantic change over time (e.g., Sagi, Kaufmann & Clark 2011, Gulordava & Baroni 2011, Kim et al. 2014, Jatowt & Duh 2014, Kulkarni et al. 2015, Hellrich & Hahn 2017, Bamler & Mandt 2017).

Distributional methods to embed words in vector spaces by means of their co-occurrence statistics started to be used in the 2010s (Boleda 2020). Initially, classic distributional methods were used. Sagi, Kaufmann & Clark (2009) employ a variation of latent semantic analysis to identify semantic change of some words from early to modern English. Latent semantic analysis is a technique that classifies documents into small groups by semantic similarity using a matrix containing word counts per document. Gulordava & Baroni (2011) use n-gram corpora and a distributional similarity model which computes context vectors based on the frequency counts of n-grams in order to detect cases of major diachronic context change and semantic change of words in the 1960s and 1990s. However, since the introduction of neural network representations by Kim et al. (2014), those classic methods have been predominantly replaced with neural network

representations in later studies (e.g., Hamilton, Leskovec & Jurafsky 2016, Szymanski 2017, Del Tredici, Fernández & Boleda 2019).

Kim et al. (2014) employ a neural language model to identify words the usages of which have changed over time or to detect specific periods when such changes occurred. They collect ten million 5-grams from the subcorpus of English fiction in the Google Books Ngram corpus for each year from 1850 to 2009 and use the yearly corpora to obtain word vectors. They have identified words which underwent change in their usages between 1900 and 2009 (e.g., *gay*, *actually*, *checked*, *check*, *supposed*, and so on) by comparing the values of cosine similarity among the word vectors in different years for the same word (the term cosine similarity will be explained further in Section 3.2.1). Also, they have detected the specific periods during which the usages of the target words changed by observing how the values of cosine similarity between the target words and their neighboring words have changed across time.

Kulkarni et al. (2015) detect and track linguistic shifts across time in terms of the meaning and usage of words through three approaches, i.e., frequency, syntactic, and distributional methods. They investigate different aspects of word evolution over time by applying the three approaches to Twitter data spanning from September 2011 to October 2013, movie reviews from Amazon spanning from August 1997 to October 2012, and the Google Books Ngram Corpus spanning from 1900 to 2005. They demonstrate that the distributional method outperforms the other two methods by means of multiple evaluation methods including a synthetic corpus, a reference dataset, and human evaluation. With the distributional method (i.e., neural language model), they track changes in the meaning and usage of words over time by putting context vectors from each time period in one unified coordinate system and tracking their displacements across time. Examples of the linguistic shifts from the Amazon reviews and Twitter datasets identified through their

distributional method include words like *ray*, *rays*, *streaming*, and *combo*, to which new meanings were added with the introduction of new technologies and products. These examples show that their distributional method can detect the introduction of new movies, books, products, games, events, and so on.

Del Tredici, Fernández & Boleda (2019) utilize a distributional model and data from the r/LiverpoolFC subreddit (an online forum of football fans) to explore meaning shift which arose during the period from 2011 to 2017. They have identified three main shift types, i.e., metonymy, metaphor, and meme using cosine distance among word embeddings across time. To assess the performance of their model, they use an evaluation dataset consisting of 97 words from the r/LiverpoolFC subreddit. The evaluation dataset was annotated with semantic shift indices which range from "0" (no shift) to "1" (shift) by twenty-six members of r/LiverpoolFC. Those semantic shift indices were determined according to the judgments of the members. They found that there is a positive correlation between the results from their model and the evaluation dataset, which demonstrates that their model can successfully detect most semantic shifts in their data.

In addition to cosine similarity and cosine distance that the above three studies used to detect and track semantic shifts, there is another method to track semantic development: topic models, such as latent semantic analysis. Topic models discover latent semantic structures in an extensive body of text. They classify documents into clusters of words sharing the same topic. Many studies using topic models explore the semantic shift of the target word through change in context words of the target word (e.g., Wijaya & Yeniterzi 2011, Lau et al. 2012, Frermann & Lapata 2016). As the frequency of a word gets higher with time, it tends to be used in different contexts, resulting in change in the word meaning. The way a word is used, i.e., the change of

words co-occurring with the target word reflects the semantic change of the target word. Studies employing topic models use this principle to catch semantic shift.

Wijaya & Yeniterzi (2011) employ a Topics-over-Time model and k-means clustering to look into 5-grams and detect the periods when the 5-grams move from one topic to another. They use the Google Books Ngram dataset, which is a corpus of over 500 billion words in English, French, German, Spanish, Russian, Chinese, and Hebrew. It consists of n-grams extracted from about 5 million digitized books published between the 1500s and 2008. To take a few examples mentioned by Wijaya and Yeniterzi, the word *gay* started to be used to represent 'a homosexual man' from the meaning 'cheerful and lively' around the 1970s, which is demonstrated by change in the topics of context words of the word *gay*. The word *Iran* shows a new cluster containing words related to republic and revolution after 1978. The new cluster identified through their k-means clustering is directly related to the change of regime in Iran from monarchy to an Islamic republic. This shows a correlation between the period identified by their model and a real historical event in the same period.

Frermann & Lapata (2016) utilize a dynamic Bayesian model for study of diachronic meaning change. They built a diachronic text corpus spanning the period from 1700 to 2010 using documents from three sources. After preprocessing the diachronic text corpus, they extracted 5-grams from the corpus (a 5-gram refers to a contiguous sequence of five lexical items) and trained their model with them. They tracked how the senses of a word gradually and smoothly evolve across a sequence of time intervals using their model. The result showed that the word *mouse* had one sense (i.e., 'a small rodent') until the mid-twentieth century and subsequently gained a second sense (i.e., 'a device connected to a computer'). Also, they showed that their model can infer subtle changes within a single sense as well as capture the emergence of new senses, the dominance of

some senses, including semantic broadening and narrowing. They demonstrated that their model is suitable for research on meaning change through four experiments.

Distributional semantics has been increasingly getting attention in the recent decade. For the rise of distributional semantics, the construction of large-scale diachronic corpora of various languages is needed. Tahmasebi, Borin & Jatowt (2018)'s survey shows that 19 out of 23 datasets employed for diachronic semantics are English datasets. Also, a majority of studies on distributional semantics use the Google Book Ngrams corpus covering only the period from 1850 to 2009 (Boleda 2020). Therefore, the construction of a number of large-scale corpora covering various languages and periods will be able to promote active research in distributional semantics, leading to the development of distributional semantics.

2.3 Lexical Semantics

Lexical semantics is the study of word meaning. Research into word meaning was established as a subdiscipline of linguistics by the middle of the nineteenth century. To mention the flow of the history of lexical semantics briefly, lexical semantics begins with prestructuralist semantics, which dominated the scene between 1870 and 1930. Prestructuralist semantics emphasizes a diachronic change of meaning and psychological aspects of lexical meaning. In the 1930s, structuralist semantics appeared in reaction to prestructuralist semantics. Structuralist semantics takes a synchronic perspective of meaning and views meaning and semantic structures as an autonomous system.

Generativist semantics succeeded structuralist semantics in terms of methodology. It is considered as the combination of structuralist analysis and generative grammar. The late 1960s to the 1970s is characterized by the generativist model introduced by Katz & Fodor (1963). Generativist semantics was followed by cognitive semantics. Cognitive semantics emerged in the

1980s. It links meaning to mind. It regards lexical meaning as a concept in the mind on the basis of experiences with the entity or relation.

Neostructuralist semantics and neogenerativist semantics came from structuralist semantics and generativist semantics, respectively. Neostructuralist semantics is based on structuralist ideas but pays attention to issues raised by generativist semantics (i.e., formalization and the borderline between linguistic meaning and cognition). Neogenerativist semantics agrees with cognitive semantics in terms of the flexibility and dynamism of meaning but adheres to the distinction between linguistic meaning and world knowledge unlike cognitive semantics. The three approaches, neostructuralist semantics, neogenerativist semantics, and cognitive semantics serve as current approaches of lexical semantics (Geeraerts 2002).

As this study is closely related to structuralist semantics, I focus on structuralist semantics here. Weisgerber (1927) presented the first theoretical basis and methodology of structuralist semantics. Weisgerber criticizes the features of prestructuralist semantics, which focuses on change of meaning across time. It regarded lexical meanings as psychological entities so meaning changes were considered to result from psychological processes. However, Weisgerber argues that the study of meaning should be synchronic and meaning should be identified and analyzed in a linguistic way, i.e., through linguistic structures. That is, semantic structures should be examined for the study of meaning.

The development of synchronic, non-psychological, and structural semantic theories is contingent on how semantic structures are defined. Semantic structures were defined as the following three lexical relations by structuralist semanticians, which constitute the methodological basis of lexical semantics: (i) the relationship of semantic similarity, which forms the basis of semantic field analysis; (ii) lexical relations such as synonymy, hyponymy, and antonymy, which

were for the first time chosen as the methodological basis of structuralist semantics by Lyons (1963); (iii) syntagmatic lexical relations, which were identified by Porzig (1934). The third category, syntagmatic lexical relations, later appeared again in the form of selectional restrictions in the approach of neostructuralist semantics and they were incorporated into generative grammar through Katz & Fodor (1963).

These three distinct definitions were passed to neostructuralist semantics. The third semantic structure evolved into a sort of collocational analysis by Firth (1957b). A collocational analysis was later incorporated into the framework developed by Halliday (Halliday & Hasan 1976) and into lexicography (Sinclair 1987a, Moon 1998). Halliday & Hasan (1976) cover cohesion in English, which stems from semantic relations between sentences. They describe different types of cohesion such as reference, substitution, ellipsis, conjunction, and lexical cohesion and present a method for the analysis of cohesion. This study is significant in that it is regarded as the root of collocational analysis. The incorporation of collocational analysis into lexicography is based on the idea that a collocational analysis should use actual language data to describe meaning.

Given that a collocational analysis involves co-occurrences in actual contexts, it naturally links up with corpus linguistics, which studies language using corpora of actual language data. An introduction to corpus-based collocational analysis is Partington (1998), which emphasizes that the exploration of corpus data allows us to have access to information that language textbooks, dictionaries, and other resources cannot provide. Moreover, it shows how corpora can be employed to solve problems about language phenomena through case studies using corpora and concordance technology. A more detailed and systematic explanation of collocational analysis is found in Stubbs (2001b). It demonstrates the usefulness and importance of corpora in lexical semantics as well as provides theoretical background in meaning, the definitions of critical notions in lexical

semantics and lexical relations with examples, and the methodology for corpus studies of lexical semantics (for additional studies on a collocational analysis, see Hoey 1991, 2005 on lexical priming; Louw 1993 on semantic prosody; Moon 1998 on idioms).

Lexical relations treated in a collocational analysis include collocation, colligation, semantic prosody, and semantic preference. Those lexical relations are the ones commonly investigated in corpus linguistics. Among the four lexical relations, semantic prosody is applied to the distributional corpus analysis of the neologism *kay-* and collocation is used for method validation of techniques employed to analyze the context words of the neologisms *leyal* and *lwuce*.

2.3.1 Lexical Relations

This section covers four lexical relations in detail: collocation, colligation, semantic prosody, and semantic preference. Among the four lexical relations, collocation is to be used for method validation of distributional corpus analysis and semantic prosody is to be used for the analysis of semantic change of the neologism *kay-*. Specifically, the subsection on collocation covers how collocation is employed for method validation of distributional corpus analysis of the neologisms *leyal* and *lwuce*. The subsection on semantic prosody covers what the limitations of previous studies on semantic prosody are and how long short-term memory can help to overcome such limitations.

2.3.1.1 Collocation

According to Sinclair (1991, 1996), a collocation is defined as a word combination that recurs more often than by chance. In a collocation, the target word is called the node word and words surrounding it are called collocates. The term collocation was introduced by Firth (1957b). Firth argues that the most typical collocates of the node word characterize the meaning and usage of the

node word to some extent. This led to his famous statement, “you shall know a word by the company it keeps.” Traditionally, collocations have been regarded as a relationship between only two lexical elements (Jones & Sinclair 1974, Sinclair 1987b).

It is impossible to find collocates relying on absolute frequencies alone. Collocates can be found through statistical association measures. Those association measures compute the syntagmatic attraction between the node word and its collocate, i.e., collocation strength. The most commonly used association measures include the z-score, the t-score, the log likelihood ratio, MI (Mutual Information), and Fisher’s exact test. Because the result of collocation analysis can be different by which association measure is used, the choice of association measures requires a careful determination (Lehecka 2015).

Collocation analysis has been widely performed in corpus linguistics. In particular, it has been extensively used to compare near-synonyms (Lehecka 2015). To take an example of *strong* versus *powerful*, although their meanings are similar to each other, their usages differ: *a powerful car* and *strong tea* make sense but *a strong car* and *powerful tea* sound awkward (Halliday 1966: 150, Church et al. 1991, Church et al. 1994). Paying attention to this characteristic of near-synonyms, a number of studies have demonstrated that near-synonymous words have similar meanings but their lexical contexts are different.

Kennedy (1991) performed an empirical analysis of differences between collocations and semantic functions of *between* and *through* using the one-million-word Lancaster-Oslo/Bergen corpus of adult written British English. The comparison between their collocations shows that the immediately preceding words which occur most frequently before each one are clearly different. Specifically, while *between* is typically preceded by nouns, *through* is preceded by verbs. Another notable difference is that 72 percent of the *-ed* forms before *between* are, grammatically, past

participles in passive voice constructions whereas 61 percent of the *-ed* forms before *through* are finite past tense forms. The patterns of collocations following *between* and *through* markedly differ as well. For example, plural pronouns are more dominantly used after *between*. Concerning their semantic functions, a majority of the tokens of *between* are composed of non-locative uses and *through* more commonly occurs with a dynamic sense.

Kjellmer (2003) compared the usages of *almost* and *nearly*. Kjellmer used the CobuildDirect Corpus to examine their differences in terms of their frequency, style and text type preference, and collocations. The findings show that (i) as *almost* is much more frequently used than *nearly*, it tends to be less specialized than *nearly*; (ii) while *almost* occurs more often in literary writings rather than in popular text types, *nearly* occurs more frequently in the news media (neither *almost* nor *nearly* is used much in the spoken language); (iii) the collocates of *almost* and *nearly* are clearly different (*almost* co-occurs with adjectives, adverbs, pronouns, and prepositions, whereas *nearly* co-occurs with numerals).

Le & Kim (2018) used the 560-million-word online Corpus of Contemporary American English to explore the collocational behaviors of *wide* and *broad*. They investigated the similarities and differences of the two near-synonymous adjectives with focus on their overall usage patterns, nominal collocates, semantic preferences, and semantic prosodies. They limited the nominal collocates to nouns on the right side of the two near-synonymous adjectives. To obtain the collocates, they employed MI scores and frequency measure. They obtained and analyzed forty-five most frequent collocates, whose MI scores are higher than "3". They found out that (i) *wide* is more often used across all the genres and in the entire corpus, (ii) while *wide* is most often used in the genre of Magazine, *broad* is most frequently used in the Academic genre, and (iii) *broad* has more various semantic preferences than *wide* and both of them have neutral semantic prosodies.

Phoocharoensil (2021) used the Corpus of Contemporary American English to look into the differences of the two synonymous verbs *persist* and *persevere* in terms of their distribution across genres, collocations, semantic preferences, and semantic prosodies. Collocates were limited to noun collocates and the collocates were obtained using MI scores in conjunction with frequency measure (i.e., twenty most frequent noun collocates whose MI scores are greater or equal to "3"). The results show that *persist* is most frequently used in academic texts whereas *persevere* is most frequently used in webpages. Also, they do not share any noun collocates although their meanings are similar to each other; *persist* is combined with a wider range of collocates than *persevere*. Concerning semantic preferences and semantic prosodies, *persist* tends to co-occur with lexical items expressing continual unpleasant situations, leading to a negative semantic prosody. In contrast, *persevere* co-occurs with lexical items expressing hardship/difficulties or determination/effort, leading to a neutral semantic prosody.

In this dissertation, collocation is used for method validation of word2vec and LDA. Word2vec and LDA are computational techniques employed to analyze the distributional behaviors of the neologisms *leyal* and *lwuce*. Method validation is needed to check whether the word2vec and LDA models have been well trained and results from those models are reliable. Similar results between those models and method validation will be able to verify the results from the models. Based on the statement from Firth (1957b) that the most typical collocates of the node word characterize the meaning and usage of the node word to some extent, I will look into how top collocates of *leyal* and *lwuce* change over time to track the semantic change of each neologism.

To be specific, for *leyal*, I will obtain top collocates of *leyal* from the 2010 corpus and examine whether there is change in their frequency and order across time. Also, I will compare the top collocates with collocates of its alternative words *cincca* and *leyalmatulitu* representing the

two meanings of *leyal* respectively. For *lwuce*, I will obtain collocates of *lwuce* related to the new meaning 'a man's stature' and investigate whether there is change in their frequency and order over time. Moreover, I will look into how the number of collocates related to the new and existing meanings changes over time within the range of top eleven collocates.

2.3.1.2 Colligation

The definition of colligation varies depending on scholars. The notion of colligation originates from Firth (1968), in which colligation is defined as the syntagmatic interrelation of grammatical categories. Sinclair (1996) defines colligation as the attraction between a grammatical feature and a lexical item, in other words, the co-occurrence of grammatical choices. According to Tognini-Bonelli (2001) and Stubbs (2001a), colligation is the correlation between a grammatical category and a lexical item. Hoey (2005) provides the most extensive definition of colligation. He argues that colligation embraces the following three distributional attractions: (i) the attraction between a lexical item and the grammatical company preferred or avoided by the lexical item in its context, (ii) the attraction between a lexical item and the grammatical functions preferred or avoided by the context where the lexical item occurs, and (iii) the attraction between a lexical item and the position in text and discourse preferred or avoided by the lexical item.

Colligation analyses have been mainly carried out in comparative studies of near-synonyms (Lehecka 2015). Some colligation studies have demonstrated that the grammatical contexts where near-synonyms occur can be different from each other, although their meanings are similar (e.g., Biber, Conrad & Reppen 1998, Gilquin 2003). Other colligation studies have questioned the syntactic characteristics of some near-synonyms described in traditional reference materials such as the *Oxford Advanced Learner's Dictionary* and the *Longman Dictionary of Contemporary English* (e.g., Atkins & Levin 1995, Liu 2010). They analyze corpora to show that the distributional

behaviors of those near-synonyms are actually different from descriptions of them from the traditional reference materials.

Atkins & Levin (1995) investigate the colligational patterns of the two English verbs *quake* and *quiver*. The Collins COBUILD Dictionary and the Longman Dictionary of Contemporary English state that both of the verbs are intransitive, whereas the Oxford Advanced Learner's Dictionary describes that *quake* is an intransitive verb and *quiver* is a transitive verb. However, Atkins and Levin demonstrate that *quake* and *quiver* can function as transitive verbs as well as intransitive verbs through their examples in transitive constructions in a 50-million-word corpus (e.g., *the insect quivered its wings* and *it quaked his bowels with fear*).

Biber, Conrad & Reppen (1998) look into how the English adjectives *little* and *small* differ in terms of predicative versus attributive positions. Using two corpora of academic prose and conversation, they show that *small* is more frequently used in a predicative position in both registers, i.e., academic prose and conversation and this preference is stronger in the latter register. In conversation, *small* in the predicative position and *little* in the attributive position are similar in that they both characterize physical size. However, they have different functions. Predicative *small* is used to depict the attribute of smallness from the person being described, whereas attributive *little* is used on the one hand to represent an identifying characteristic that makes the noun being described recognizable, but on the other to express other purposes (e.g., *he's like any little kid I think*).

Gilquin (2003) uses the British component of the International Corpus of English (ICE-GB) to inspect the distributional behaviors of the causative verbs *get* and *have*. With regard to the syntactic characteristics of them, she showed the following four findings: (i) *get* is mainly used with infinitives and past participles (42.6 percent and 36.6 percent respectively) and rarely used

with present participles (20.8 percent); (ii) *have* is largely construed with past participles (71.4 percent) and is not used with present participles and infinitives very often (15.6 percent and 13 percent respectively); (iii) the demotion of the CAUSEE appears more frequently with *have* than *get*; (iv) the promotion of PATIENT appears more often with *have* than *get*.

Liu (2010) probes the five English synonymous adjectives *primary*, *principal*, *chief*, *main*, and *major* using the approach of corpus-based behavioral profile analysis. He showed the following four main things: (i) *chief* and *principal* rarely modify abstract/dual nouns ("dual" means belonging to abstract and concrete) but when they do, the phrases from such modification are largely used in formal registers like academic writings; (ii) when *principal* is used in non-position titles, it means the highest degree of share or contribution instead of power or authority; (iii) *major* and *primary* can serve as predicative adjectives; (iv) the meanings of the adjectives under consideration, particularly *primary*, change depending on the context. These findings dispute some of the existing descriptions on the distributional characteristics of those adjectives from traditional reference materials like dictionaries.

Liu & Espino (2012) examine differences in the distributional patterns of the four English synonymous adverbs *actually*, *really*, *truly*, and *genuinely*. According to them, when *actually* functions as a disjunct, it is typically used in contexts implicating surprise/contradiction. When *actually* is used as an intensifier/emphasizer of verbs and adjectives, it seldom occurs with verbs and adjectives related to emotion/attitude/desire/cognition. In contrast, *really* serves as both a versatile disjunct and a versatile intensifier/emphasizer for verbs/adjectives. It is most frequently used with evaluative adjectives. The adverbs *truly* and *genuinely* are largely used as intensifiers/emphasizers of verbs/adjectives related to emotion/attitude/desire/cognition. They

more often function as adjective intensifiers/emphasizers rather than being used as verb intensifiers/emphasizers.

These studies show that corpus analysis is required for colligation analysis. Because the usage of a word can be changed over time, results from the analysis of up-to-date corpora and descriptions from traditional reference materials can be different. In order to find out the most recent usage of the target word, it is needed to analyze corpora containing the latest information of the target word.

2.3.1.3 Semantic Prosody

Research on semantic prosody dates back to Sinclair (1987a), which found that the word *happen* and the phrase *set in* mainly co-occur with unpleasant events. Although the study of semantic prosody started off with Sinclair (1987a), it was Louw (1993) that introduced the term semantic prosody. According to Louw (1993), semantic prosody refers to "a consistent aura of meaning with which a form is imbued by its collocates". Sinclair (1996) defines semantic prosody as an impression of a pragmatic and/or attitudinal meaning which is created by the usage of a word. In Partington (1998), semantic prosody is defined as "the spreading of connotational coloring beyond single word boundaries". Hunston & Thompson (1999) describe semantic prosody as follows; the frequent occurrence of a given lexical item in the context composed of predominantly positive or negative words/phrases leads to association between the lexical item and the positive or negative meaning coming from the context.

According to Louw (2000), semantic prosody is "a form of meaning which is established through the proximity of a consistent series of collocates". Xiao & McEnery (2006) argue that the semantic prosody of a lexical item comes from the interaction between the lexical item and its collocates. Through the definitions of semantic prosody from all of these studies, semantic prosody

can be summarized as attitudinal and connotational meaning established by the evaluative or emotive attitude from the collocates of the word under scrutiny. Semantic prosody is generally categorized as good/bad, pleasant/unpleasant, or favorable/unfavorable.

Semantic prosody research has been conducted in the domain of corpus linguistics because corpora allow linguists to observe semantic prosody empirically (Whitsitt 2005). Specifically, semantic prosody has been explored by means of corpus analysis using computational methods. Many scholars have emphasized the importance of the use of corpora and computational methods for semantic prosody research. Louw (1993, 1997) claims that the study of prosodic profiles needs computational methods because semantic prosody is hard to access through human intuition. Stubbs (1995) argues that the investigation of semantic prosody should be performed through attested data, not a native speaker's intuition. Bublitz (1996) supports their importance by mentioning that the application of computational methods to large-scale corpora can best and only reveal the relationship between a lexical item and its context.

Widdowson (2000) states that language behavior not accessible to intuition can be disclosed through a quantitative analysis of texts using a computer, which buttresses the employment of computational methods for semantic prosody research. Also, Adolphs & Carter (2002: 7) argue that large-scale corpora and suitable software only enable researchers to delve into semantic prosody. Sinclair (2003) agrees with such methodology by arguing that semantic prosody research should be conducted using large-scale corpora and corpus analysis techniques. But for the development of technology and computers in the past few decades, semantic prosody studies could not have been actively carried out on the basis of corpora (Begagić 2013).

The semantic prosody of a given lexical item is determined by whether its co-occurrences are pleasant or unpleasant and specifically, it is based on the frequencies of pleasant and unpleasant

co-occurrences (Stewart 2010). In corpus linguistics, co-occurrences have been inspected through concordances and collocational profiles. A concordance refers to an alphabetical index of primary words in a text or a group of texts with the instances of immediate contexts in which each of those words occurs. A collocational profile concerns the types of collocations and their frequencies. Co-occurrences surrounding the node word which have high frequency are called collocates. They are identified through statistical association measures like the Pointwise Mutual Information index (Church & Hanks 1990). Corpus analysis software tools such as WordSmith Tools, Sketch Engine, and AntConc have functions to apply statistical association measures to corpora and display concordances and collocational profiles.

Several previous studies employed those corpus analysis software tools to investigate semantic prosody. Tognini-Bonelli (2004) found that the semantic prosody of the English verb *face* is negative based on its negative collocates. The collocates include *challenges*, *competition*, *dilemma*, *difficulties*, *obstacles*, *pressures*, *problem*, *shame*, *threat*, and so on. The co-occurrence with these collocates demonstrates that the verb *face* is used with unfavorable things. She emphasizes that (i) collocates are obtained through computer software tools such as MonoConc Pro and WordSmith Tools and (ii) it is a quantitative analysis on the basis of corpora that has made semantic prosody tangible. This is, the corpus-based quantitative analysis enables researchers to observe the connotational and attitudinal meaning pervading the instances of the target word.

Xiao & McEnery (2006) delve into the semantic prosodies and collocations of some near synonyms in the English and Chinese languages from the viewpoint of a cross-linguistic approach. The purpose of their study is to find out whether Chinese also shows semantic prosody and semantic preference as English does and how similar or different the semantic prosodies and collocational behaviors of English near synonyms and their close Chinese equivalents are. They

investigated three groups of near synonyms: (i) the consequence group, (ii) the CAUSE group, and (iii) the price/cost group. They used the Freiburg-Brown Corpus of American English, the Freiburg-LOB Corpus of British English, and the Lancaster Corpus of Mandarin Chinese as their principal corpora.

They employed WordSmith for the English data and Xaira for the Chinese data in order to obtain the collocates of the target words. They regarded four words to the left and right of the target words with a minimum Mutual Information score of "3" and a minimum co-occurrence frequency of "3" as collocates. They sorted all the instances of near synonyms and their collocates into positive, neutral, and negative. They found out that the semantic prosodies and collocational behaviors of some English near synonyms are fairly similar to those of their Chinese equivalents but other English near synonyms have different semantic prosodies and collocational behaviors from their Chinese equivalents. Xiao and McEnery suggest that the differences between the two languages may derive from morphological variations which apply in English but do not in Chinese.

Zhang (2013) traces the diachronic change of semantic prosody. She examines the development of semantic prosodies of the four English adverbial intensifiers *awfully*, *terribly*, *dreadfully*, and *horribly* by counting the frequencies of their non-negative (i.e., neutral and positive) and negative collocates across three periods: (i) from 1710 to 1850, (ii) from 1850 to 1920, and (iii) since 1980. The collocates were extracted from the Bank of English corpus, a fiction corpus that she complied herself, and the Corpus of Late Modern English Texts by means of WordSmith. The collocates were classified into negative, neutral, or positive by the author and a trained rater. The results show that the frequencies of neutral and positive collocates have increased, while those of negative collocates have decreased across time for all the four adverbs. This implies that their negative semantic prosodies have become weak.

Alcaraz-Mármol & Almela (2016) explore the semantic prosodies of the two Spanish nouns *inmigrante* (immigrant) and *inmigración* (immigration) in written media. They collected articles about immigration between 2003 and 2013 from the two national newspapers El País and El Mundo, which have different political ideologies. They used WordSmith to analyze the corpus of articles that they compiled. They regarded three words to the right and left of the target words as their co-occurrences, with the minimum co-occurrence frequency set as "9". The results show that most of the co-occurrences of the two target words have unfavorable meanings, which implies that the two nouns *inmigrante* and *inmigración* have negative semantic prosodies.

Although it is true that a systematic study on semantic prosody has become possible with the help of corpora and corpus analysis software tools, the existing software tools cannot translate results from concordances and collocational profiles into semantic prosody. That is, the analysis of semantic prosody requires analysts' interpretations of concordances and collocational profiles. However, interpretations carry a risk of involving analysts' subjectivity to a certain degree. This risk can cause different results depending on analysts, although they investigate the semantic prosody of the same lexical item. In particular, because the process of sorting co-occurrences or collocates into favorable and unfavorable is not straightforward, the classification results depend on analysts. Concerning this issue, some studies suggest methods to quantify prosodies (e.g., Osgood, Suci & Tannenbaum 1957, Dilts & Newman 2006) but the methods are limited to words of a particular part of speech. Other studies make reference to previous studies showing what lexical items are positive or negative (e.g., Ahmadian, Yazdani & Darabi 2011). However, neither of these methods provides a clear resolution to the problem.

In addition, previous studies have mainly focused on the synchronic analysis of semantic prosody, although its diachronic analysis is of great importance for semantic prosody research. To

borrow terms from the definition of semantic prosody by Louw (1993) (i.e., “a consistent aura of meaning with which a form is imbued by its collocates”), because the process of a form being infused with the meaning from its collocates takes a long time, semantic prosody research presupposes tracking the development of semantic prosody with a huge amount of data over a long period of time. However, there are few studies investigating semantic prosody over a very lengthy period of time. Given that state-of-the-art artificial intelligence methodology, i.e., long short-term memory suggested in this study is able to classify lexical items as positive or negative with no analyst’s subjective judgment and process large datasets more efficiently, long short-term memory is expected to address the issues from previous studies over subjectivity and processing a huge amount of data.

2.3.1.4 Semantic Preference

Semantic preference concerns looking into the semantic features that collocates of the target word share (Sinclair 1991, 1996). That is, it approaches co-occurring words from a semantic perspective. Stubbs (2001b: 65) provides a more specific definition of semantic preference. According to Stubbs, semantic preference refers to “the relation, not between individual words, but between a lemma or word form and a set of semantically related words”. To have a better understanding of semantic preference, it is needed to know the relation of semantic preference to semantic prosody because the two things are interdependent collocational meanings (McEnery, Xiao & Tono 2006).

Partington (2004) argues that they interact: whereas semantic preference makes a contribution to developing semantic prosody, semantic prosody forms the general environment which restricts the preference of the node word to co-occur with particular items. In addition, Partington (2004) describes the difference between semantic preference and semantic prosody using their different operating scopes. Semantic preference is regarded as a feature of collocates.

It connects the node word to another word from a particular semantic set. In contrast, semantic prosody is viewed as a feature of the node word. It can affect wider stretches of text.

To mention some previous studies on semantic preference, Partington (1998) investigated the semantic preference of the intensifying adjective *sheer* using newspaper and academic corpora. He found out that *sheer* collocates with particular semantic sets, i.e., items expressing "magnitude", "weight", or "volume", items expressing "force", "strength", or "energy", items expressing "persistence", nouns expressing "strong emotion", and physical quality. Also, Partington compared the collocational behavior of *sheer* with the collocational behaviors of its synonyms such as *complete*, *pure*, and *absolute* and found out that none of them have the same semantic preferences as *sheer*.

Stubbs (2001b) used a 200-million-word corpus to analyze the semantic preference of *large*. Stubbs found out that it collocates with words for "quantities and sizes" such as *number*(*s*), *amounts*, *quantities*, *scale*, and *part* in at least a quarter of its 56,000 occurrences. Furthermore, Stubbs (2001b: 89-95) looked into the semantic preference of the verb *undergo*. The collocates following the verb indicate that *undergo* collocates with several semantic sets, i.e., "medicine" (e.g., *brain surgery*, *treatment*, *hysterectomy*), "tests" (e.g., *training*, *examination*), and "change" (e.g., *a historic transformation among others*, *dramatic changes*). These preferences result in an unfavorable semantic prosody because those collocates are things that people are forced to undergo although they do not want to.

As a further study of Partington (1991), Partington (2004) investigated the collocational behaviors of four maximizers *completely*, *entirely*, *totally*, and *utterly*, which share many collocates. He discovered that (i) the four maximizers collocate with items expressing "absence" or "change of state", (ii) *entirely* collocates with items expressing "(in)dependency" as well as

"absence" or "change of state", and (iii) *utterly* has an unfavorable semantic prosody whereas the other three maximizers evenly occur with favorable and unfavorable items. Partington (2004) adds the semantic preferences of *absolutely* and *thoroughly* and the semantic prosody of *perfectly* based on findings from Partington (1991): (i) *absolutely* collocates with "hyperbole" and "superlatives"; (ii) *thoroughly* collocates with items related to "emotions" or "liquid penetration"; (iii) *perfectly* tends to co-occur with "good things", leading to a favorable semantic prosody.

Stewart (2010) used the British National Corpus to show the semantic preference of the verb *break out*. A total of 1,126 occurrences including all the inflected forms of *break out* were analyzed. The findings show that *break out* collocates with words such as *war, conflict, infection,* and *crisis,* leading to an unfavorable semantic prosody. Those co-occurrences belong to semantic sets of "diseases", "situations of conflict", or more broadly "problematic circumstances".

Begagić (2013) examined the semantic preference and semantic prosody of the collocation *make sense* using the Corpus of Contemporary American English (COCA). She hypothesized the following two things in her study: (i) the semantic preference and semantic prosody of the collocation *make sense* in the newspaper register are different from those in the academic register; (ii) it is possible to infer the semantic preference and semantic prosody of the collocation *make sense* (i.e., the collocation has its semantic preference and semantic prosody). She investigated four word forms of *make sense,* i.e., *make sense, makes sense, made sense,* and *making sense* in the two registers. She selected 50 instances for each word form, totaling 400 examples (200 examples per register) and manually analyzed approximately ten words to the left and right of the collocation.

All the four word forms of the collocation occur more frequently in negative environments in the newspaper register as opposed to the academic one, which demonstrates that the first

hypothesis is true. She explains that this phenomenon is because of the nature of news that news items are more likely to cover negative contents such as problematic and tragic events. More specifically, while *make sense* and *making sense* occur more frequently in a negative environment, *makes sense* and *made sense* occur more frequently in a positive environment. This means that the collocation *make sense* cannot be said to have a negative semantic prosody overall.

The second hypothesis was also found to be true. All the four word forms of *make sense* collocate with words related to "difficulty" such as *try*, *attempt*, *struggle*, and *help*, leading to unfavorable semantic prosody. In addition, it can be said that the semantic preference of the word form *make sense* is "uncertainty" in that it commonly occurs in hypothetical phrases as well as it collocates with various modals. In contrast, the word forms *makes sense* and *made sense* occur in more factual and definite environments, which means that they have a rather favorable semantic prosody.

Alrajhi (2019) explored the semantic preference and semantic prosody of the four maximizers *completely*, *entirely*, *totally*, and *utterly* in EFL (English as a foreign language) Saudi students' writings. The results were compared to the findings of Partington (2004). Partington (2004) investigated those four maximizers using data obtained from the COBUILD corpus. With regard to semantic prosody, all the four maximizers used in the students' writings have favorable semantic prosodies. However, in Partington's study, *utterly* has an unfavorable semantic prosody whereas the other three maximizers evenly occur with favorable and unfavorable lexical items. Concerning semantic preference, the collocates of the four maximizers in the students' writings belong to semantic sets associated with "emotion" and "state of mind". In contrast, in Partington's study, all the four maximizers occur with words related to "absence" or "change of state". The collocational behavior shared between the two studies is that the four maximizers occur with words

concerned with “change” and *entirely* collocates with words related to “(in)dependency”. Alrajhi (2019) shows how different the way EFL Saudi students use the four maximizers is from real English through this study.

As the other three lexical relations are used to distinguish synonyms, semantic preference can also be used. Semantic features that collocates share can be different between synonyms. Semantic preference approaches differences between synonyms semantically while semantic prosody approaches them pragmatically and colligation approaches them syntactically. Collocation shows them in terms of what words occur next to each synonym. Given that the four lexical relations enable linguists to investigate differences between synonyms in various aspects, they are very useful for research on synonyms.

2.4 Usage-based Linguistics

Usage-based linguistics is based on the following two aphorisms: (i) meaning is use; (ii) structure emerges from use (Tomasello 2009). The first aphorism is associated with the symbolic or functional dimension of linguistic communication. It focuses on the use of linguistic conventions for achieving social ends. Usage-based linguistics emphasizes the symbolic or functional dimension, rather than grammar, which is derivative (Tomasello 2003). The second aphorism is related to the grammatical or structural dimension of linguistic communication. It focuses on how grammatical constructions emerge from actual language use and generalizations made over individual acts of language use. As people string symbols together into sequences and use the sequences for communication, patterns of use emerge and get consolidated into grammatical constructions. The grammatical or structural dimension is a product of grammaticalization. Research on grammaticalization was carried out by a number of scholars (e.g., Bybee 1985, 1995,

2002, Langacker 1987a, 1991, 2000, Hopper 1987, Croft 1991, 2001, Hopper & Traugott 1993, Goldberg 1995, Givón 1995).

In order to have a better understanding of usage-based linguistics, we need to know Chomskian generative grammar because usage-based linguistics appeared in opposition to generative grammar. Generative grammar was introduced by Noam Chomsky in the 1950s and had a profound effect on the study of language acquisition in the 1960s and 1970s. Chomsky (1968, 1980a, 1986) argued that human beings are born with abstract principles that guide the process of language acquisition, which constitute what he calls "universal grammar". According to Chomsky, universal grammar is situated at the center of human linguistic competence and on the periphery, there are such things as the lexicon, irregular constructions and idioms, the conceptual system, and pragmatics. Those things on the linguistic periphery can and need to be learned but universal grammar, the core of linguistic competence, is an innate property so it does not require learning. Universal grammar is characterized as a unified set of abstract algebraic rules irrelevant to meaning. It is the lexicon that contains meaningful linguistic elements. The meaningful linguistic elements function as variables in the rules.

In contrast to generative grammar, usage-based linguistics provides a different view of language acquisition. The usage-based theory of language acquisition suggests that language acquisition is input-driven and depends on the learner's experience of language. It argues that two sets of skills in particular are important for language acquisition: intention-reading and pattern-finding skills. The former skills are the ones that involve the attempt of one person to manipulate the intentional or mental states of other persons, which are related to the symbolic or functional dimension of linguistic communication. The latter skills are the ones that enable children to find patterns in the way adults use linguistic symbols across different utterances and to construct the

grammatical dimension of human linguistic competence (Tomasello 2003). Both of these skills are domain-general, which implies that human linguistic competence is integrated with other cognitive skills, not separate and unique mental capacity like the language acquisition device (LAD) proposed by Chomsky (universal grammar is the knowledge contained in the LAD).

Because usage-based linguistics evolved from cognitive and functional linguistics in which pragmatic and conceptual factors are considered to play an important role in the emergence of language structure and meaning (e.g., Talmy 1983, Langacker 1987b, Lakoff 1987), it has emphasized the symbolic or functional dimension of linguistic communication. However, as the focus of usage-based analysis has shifted to frequency, many studies have investigated the effects of frequency on the development of linguistic knowledge (Hay 2001, Bybee & Hopper 2001, Krug 2003, Goldberg 2006, Bybee 2006, 2007, 2010, Arnon & Snider 2010). As we can see from those studies, frequency is an important concept in usage-based linguistics and has become essential to the usage-based analysis of linguistic structure and meaning. In the following subsection, I focus on frequency. To be specific, I distinguish between type and token frequency, explain what they refer to and how they interact with language learning, and mention how token frequency is applied to this study.

2.4.1 Type and Token Frequency

A type is a category. In corpus analysis, type frequency is the number of distinct lexical items which can be substituted in a given position in a word-level construction for word formation (i.e., inflection) or a syntactic construction specifying lexical relations. That is, it means how many different lexical items are used in a given slot in a construction. To take a word-level construction example from Ellis, the English past tense *-ed* has a high type frequency because it is used with a number of different verbs, which is a regular method of making past tense verbs. As *-ed* applies to

many distinct verbs, it has a very high type frequency, compared to the vowel change exemplified in *rang* and *swam*.

Type frequency determines the productivity of morphological, syntactic, and phonological patterns (Bybee 1995; Bybee & Hopper 2001). When a number of different lexical items are used in a certain slot in a construction, the construction is less likely to be associated with a particular lexical item, leading to a more general category over the lexical items that occur in that slot. The more items the category covers, the more general the criterial features for that category get and the broader the category boundary gets. The use of many distinct lexical items in the given slot ensures that the representational schema of that construction gets more strengthened as well as the construction is more frequently used (Bybee & Thompson 2000).

In contrast, a token is an occurrence. In corpus analysis, token frequency refers to the number of occurrences of a lexical item, i.e., how often a word or a phrase occurs in a given sample. Token frequency shows how people actually use lexical items. A high token frequency of a lexical item promotes entrenchment of the lexical item in the mind (Langacker 1987a). A particular construction with high token frequency is produced more quickly and easily, recognized faster, processed more readily, and remembered better, compared to a construction with low token frequency (Ellis 2002). Multiple repetitions are known to be required for entrenchment of representations, automatization of processing, readiness of accessibility, autonomy of idiomatic expressions, and fluent, fast, and phonetically reduced production (Ellis 2009). Also, a high token frequency of a lexical item has an impact on the survival of irregular forms and idioms. The frequent use of irregular forms and idioms guarantees their conservation.

The notion of token frequency is applied to the analysis of semantic prosody of the neologism *kay-*. Because the investigation of semantic prosody is related to the actual use of the

neologism *kay-*, not the productivity of the neologism *kay-*, it is based on token frequency. The change in the token frequencies of positive and negative words to which the neologism *kay-* is attached will demonstrate whether the semantic prosody of the neologism *kay-* is positive or negative and how the semantic prosody has changed over time.

2.5 Language as a Complex System

A complex system is defined as "a system in which large networks of components with no central control and simple rules of operation give rise to complex collective behavior, sophisticated information processing, and adaptation via learning or evolution" (Mitchell 2009: 13). Examples of complex systems include economic and social organizations, transportation or communication systems, infrastructures, an ecosystem, Earth's global climate, organisms, a living cell, the human brain, and so on. The term complex systems is often used to refer to the study of complex systems that examines how interactions between a system's components produce its collective behaviors and how the system builds relations with its environment through interactions between itself and the environment.

Collective behaviors are hard to model because they are more than the sum of a system's constituent parts. They stem from the interactions and relationships between a given system's parts or between the system and its environment. Because the fundamental object of the study of complex systems is collective or system-wide behaviors, the study of complex systems is regarded as an alternative paradigm to reductionism, which is an approach to describe systems in terms of their components and individual interactions between them. The study of complex systems is drawn from a number of different fields including physics, mathematics, social sciences, biology, and many others so it is often used as a broad encompassing research approach to problems in diverse disciplines like information theory, nonlinear dynamics, statistical physics, computer

science, anthropology, sociology, economics, psychology, biology, and meteorology (Bar-Yam 2009).

According to Kretzschmar (2015), principles underlying the basic operation of complex systems are summarized as follows: (i) the activity in complex systems continues to be dynamic and open; (ii) their components randomly interact; (iii) their components exchange information with feedback; (iv) behaviors from feedback are reinforced; (v) stable patterns emerge without central control. For example, flocks of birds show these principles. When individual birds fly together, we can see flocks of birds. Although there is not any lead bird directing how individual birds act, individual birds do not bump into each other and maintain similar speeds as well as a minimum distance from other birds in the flock. They influence and are influenced by their neighboring birds in close proximity. Each bird in the flock makes moment-to-moment decisions through interaction with neighbors in its own immediate context. This example shows how individual elements act at a collective level. The interaction between individual elements leads to unplanned behaviors or properties at a collective level.

To mention some important characteristics of complex systems, they include spontaneous order or self-organization, emergence, adaptation, nonlinearity, and networks. Spontaneous order or self-organization refers to unplanned organized behavior which arises from interactions between an initially disordered system's components or a process where that unplanned order takes place. Self-organization is seen, for example, in snowflakes. Snowflakes show a radial symmetrical beauty. The symmetry results from interactions involving attractive and repulsive forces between the smaller entities that make up snowflakes (i.e., water molecules) and their surrounding environment. Emergence is concerned with the appearance of emergent behaviors or properties. Those behaviors or properties cannot be explained with a system's components in isolation

because they come from the interactions, dependencies, competitions, or relationships that the components build when placed together.

Adaptation means the ability to learn from experience and adjust to the changing environment. Complex systems that are adaptive are called complex adaptive systems, which are special cases of complex systems. For instance, when the human immune system is attacked by pathogens, it produces antibodies to destroy them. If the immune system encounters the same pathogens again, it responds to them much more quickly and efficiently thanks to memory cells. This demonstrates the learning ability and evolvability of the immune system. As pathogens develop, the immune system evolves as well.

Nonlinearity is a term used in mathematics and physics to describe a system in which the change in the size of output is not proportional to the change in the size of input. Complex systems often show nonlinear behavior, which means that they may produce not changes in output proportional to given changes in input but greater or less than such proportional changes, or even no output at all. Depending on their current state or context, they may differently respond to the same input. Nonlinear behavior is unpredictable, counterintuitive, or chaotic.

Lastly, complex systems can be regarded as networks. A network is a collection of distinct objects and relationships between them. In complex systems, the distinct objects are constituent parts of a complex system and the relations are formed by the interacting constituent parts. The view of complex systems as networks allows the relationships between components to be analyzed by means of many useful approaches such as graph theory and network science.

2.5.1 Language and Complex Systems

Complexity science originates from the disciplines of physics, mathematics, and biology but it has a wide range of applications; it has been applied to other disciplines including economics,

meteorology, and ecology (Larsen-Freeman & Cameron 2008). The attempts to apply complexity science to linguistics were made in 1980s (e.g., Lindblom, MacNeilage & Studdert-Kennedy 1984; Hopper 1987). More papers trying to couple complex systems with linguistics appeared in the 1990s (e.g., Van Geert 1991, Schneider 1997, Larsen-Freeman 1997). In recent years, complexity science has been applied to the study of second language learning (e.g., de Bot, Verspoor & Lowie 2005, Larsen-Freeman & Cameron 2008, Dörnyei 2009).

However, there had been no studies showing how each of the properties of complex systems corresponds to the aspects of speech until Kretzschmar (2009). To demonstrate speech as a complex system, he associated experimental evidence from speech with the fundamental principles of complex systems. Specifically, he showed that language in use satisfies the requirements of a complex system: (i) speech is open and dynamic; (ii) a large number of interacting components constitute speech; (iii) emergent order is observed in speech; (iv) the distribution of units in speech shows nonlinearity; (v) speech has the property of scaling or nesting.

The first requirement is fulfilled by the fact that new conversations and writings continue to occur and new feature variants continuously emerge among interactions between members of the speaking population. Also, the continuous exchange of speakers, i.e., new speakers' entering and leaving any speaking population because of birth and death or movement across social and geographical space adds to change in speech. Speech variation can be captured through change in possible realizations of feature variants and a continuum in the frequency of such realizations. Possible realizations and the frequency of feature variants are continuously variable, which clearly demonstrates that speech is dynamic. The second requirement is satisfied by that speech is composed of a number of various speech units exchanged in conversations. The components include speech sounds, words, and other entities extracted from the speech stream.

Regarding the third requirement about emergent order, he mentions the correlation of feature variants with geographical and social factors. They are correlated with each other in complex ways, leading feature variants to have statistically significant associations with independent variables related to different geographical and social factors. The grouping of variants by location or social conditions plays an important role in the occurrence of emergent order. The grouped variants show patterned behavior in their distribution (nonlinear distribution), which is regarded as emergent order. The fourth and fifth requirements are covered in the following subsection in detail. I will focus on giving a more detailed explanation about the nonlinear distribution of speech because it is the most striking evidence proving that speech is among complex systems (Kretzschmar 2015).

2.5.2 A-curve

According to Kretzschmar (2009, 2015), the frequency distribution of variants for any given linguistic feature always takes a nonlinear pattern, what he calls an "A-curve". Figure 2.1 shows the shape of A-curve, which consists of a few variants with high frequencies, some variants with moderate frequencies, and most variants with very low frequencies. The variants with very low frequencies account for the long tail of the curve. The A-curve, i.e., the nonlinear distribution was introduced to linguistics by the American linguist George Kingsley Zipf. Zipf's law named after that linguist says that the frequency of any word in texts is inversely proportional to its rank. That is, the frequency of the first-ranked word (the most frequent word) is approximately twice that of the second-ranked word (the second most frequent word), three times that of the third-ranked word (the third most frequent word), and so on. A frequency profile reflecting Zipf's law shows the same curve shape as the A-curve: a few very frequent words, some moderately frequent words, and most low frequent words.

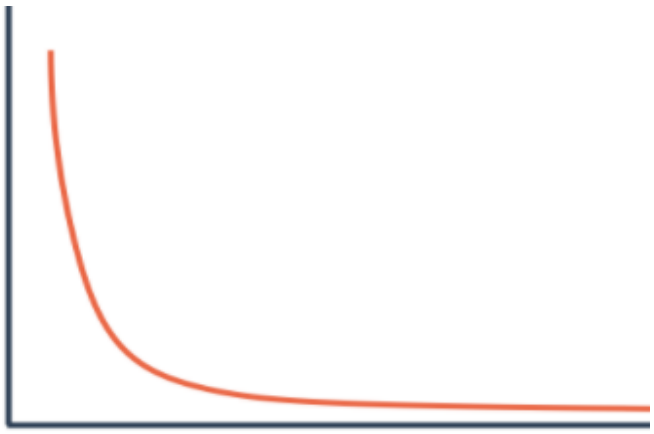


Figure 2.1 A-curve

The frequency distribution of variants for linguistic features provides clear evidence supporting that speech is a complex system. It demonstrates that speech has the characteristics of complex systems, i.e., emergence, nonlinearity, and self-organization. The A-curves show "nonlinearity". The nonlinear pattern of A-curve emerges from the frequencies of many interactive variants. That is, the pattern is created from not any single variant but the relationships among variants. The occurrence of the pattern is connected to "emergence". The important thing is that control by any external agent is not involved in the emergence of such pattern. The frequencies of variants in isolation seem to be disordered but the relationships among variants create the organized behavior of A-curve pattern, which is robust so it is always present for any linguistic feature. This is associated with "self-organization".

Kretzschmar (2015) presents some experimental evidence supporting the A-curves of variants, which was drawn from his American Linguistic Atlas Project (Kretzschmar et al. 1993 for more information). The first evidence is *dry spell* meaning 'weeks without rain'. There are 39 different lexical variants including *drouth(s)*, *drought*, and so on. Their distribution profile based

on their counts shows the A-curve. The second evidence is *parlor* meaning 'place to meet guests', which has 99 lexical variants including *living room(s)*, *sitting room(s)*, and so on. The third evidence is the vowel in *fog*, which has 242 phonetic variants and the fourth evidence is the vowel in *six*, which has 29 phonetic variants. All of the three evidence shows that the frequency distribution of their variants has the A-curve. The A-curves from these four examples are a little different from each other but they are all alike in that they have a few common variants, some moderately common ones, and many uncommon ones. Kretzschmar (2015) emphasizes that this nonlinear distribution occurs everywhere in the Atlas survey research on language in use and other surveys (e.g., Johnson 1996, Hoover 2001, Burkette 2001, 2009, Hatfield 2005, Mello 2013).

In the A-curve, we can find the scaling property of complex systems as well as emergence, nonlinearity, and self-organization. The scaling property means that the shape of a particular structure appears at any level of scale. To explain the scaling property using the above third example *fog*, the distribution profile of variants from the subsample of women as well as that of variants from the whole survey shows the A-curve, although the particular realizations at the top of the count list from the subsample are not the same as those from the whole survey. This is true of the subsample of speakers with a high-school education, too. All the A-curves from the subsamples and whole survey clearly show the scaling property of complex systems.

The most remarkable thing about the A-curve is that it can neatly explain how users of speech perceive language variation (Kretzschmar 2015). Depending on regions, social groups, or many kinds of interest groups, the way that speakers use their language changes, leading to language variation. Accordingly, a large number of variants come into existence for linguistic features. Interestingly, the distributional frequency profile of variants always follows the A-curve as mentioned above. However, because the most common variants vary depending on regions,

social groups, or many kinds of interest groups, those common variants can be considered as characteristic language behavior of each group. That is, the varying common variants can be used for users to distinguish one from the other group. The most common variants on the A-curve (a few top-ranked ones) are perceived as "normal" by users and infrequent variants "different". The top-ranked variants regarded as "normal" are assembled into linguistic systems, which are called "observational artifacts" in Kretzschmar (2015). The linguistic systems come from not reality but users' perceptions of reality.

More interestingly, people have their own linguistic systems based on language that they experience. They rely on their linguistic systems to decide what variant(s) to use. Since we speakers belong to a number of different groups at the same time, we choose variants which are judged to be best (i.e., normal) depending on given contexts. However, in this process, we know the existence of many uncommon variants for each context and understand them. This perspective on the full range of variants is associated with speech as a complex system. Every single variant corresponding to a component within a complex system is acknowledged and such acknowledged variants interact to create the A-curve. The A-curves of variants function as a perceptual aid which helps us to make sense of speech interactions. Also, they have an advantage of consistently explaining characteristic language behavior of many different groups within the frame of an overall language.

In this study, an A-curve will be used to validate results from DCA. For method validation of DCA, I will use collocates of the neologisms to perform collocation analyses. This process includes examining the distributional frequency profiles of collocates and selecting top collocates with high frequencies from the A-curves. Because the top collocates represent and characterize the entire collocates of a given word, their change means a significant change to the entire set of

collocates. Through the comparison between results from the analysis of top collocates and those from DCA, I will evaluate whether the DCA approach is suitable for semantic change research and whether results from DCA are reliable.

CHAPTER 3

BACKGROUND ON METHODOLOGY

3.1 Korean Corpora

The history of Korean corpora is not that long; it has only been about 33 years since the first electric Korean corpus was constructed. The interest in a corpus began in the late 1980s, after which corpora started to be constructed by a number of institutes and universities revolving around Yonsei University, Korea University, and the Korea Advanced Institute of Science and Technology. In this section, I focus on two main Korean corpora, i.e., the Yonsei Corpora and the Sejong Corpora.

Korean corpora consist not of words but eojuls, unlike English corpora. An eojul[4] refers to a content word itself or the morphosyntactic combination of a content word and thing(s) in charge of grammatical function such as particles and endings. Eojuls compose a Korean sentence and they are separated by spacing. This is why eojuls are used to define the size of Korean corpora and to parse sentences in Korean corpora.

[4] e.g. 비가 온다. NM: Nominative case particle
pi-ka *o-n-ta.* IN: Indicative mood suffix
rain-NM come-IN-DC DC: Declarative sentence-type suffix
'It is raining.'
The example sentence consists of the following two eojuls: (i) *pi-ka* and (ii) *o-n-ta*. To be specific, the first eojul is the morphosyntactic combination of the noun *pi* and the nominative case *ka* and the second eojul is the morphosyntactic combination of the verb stem *o-*, the indicative mood suffix *-n,* and the declarative sentence-type suffix *-ta*.

3.1.1 Introduction of Korean Corpora

This section introduces two main Korean corpora, the Yonsei Corpora and the Sejong Corpora with other Korean corpora. As the Yonsei Corpora contain the first Korean corpus, and the Sejong Corpora are national Korean corpora, the two main corpora are very significant in South Korea.

3.1.1.1 Yonsei Corpora

The first electric Korean corpus was built by the Yonsei Institute of Language and Information Studies in 1988, called the Yonsei Corpus 1. The Yonsei Corpus 1 was constructed based on how people read, i.e., their actual reading habits, with the assumption that their vocabulary is formed and reinforced by what categories of reading materials to read, especially which categories to read more. In order to find out how people read, a survey was conducted among a thousand adult Koreans. The result from the survey of asking how much time to spend on average reading each kind of written text showed that the participants spend 36.6 percent of their time on daily newspapers, 22.8 percent on magazines, 20.4 percent on books of fiction, 11.2 percent on books on general culture and information, and 9.0 percent on biographies. On the basis of these proportions, the Yonsei Corpus 1 was built. It consists of 2.9 million eojuls.

After the Yonsei Corpus 1, the Yonsei Institute of Language and Information Studies has worked on expanding their corpora. Among them, I introduce twenty-six Yonsei Corpora that are considered as the main corpora of the institute. Table 3.1 shows the size and brief description for each of the twenty-six corpora (for the information in Korean, visit https://ilis.yonsei.ac.kr/). At first the Yonsei Corpora were not open to the public, but since 2016 anyone can have access to some of them online by visiting the website at https://ilis.yonsei.ac.kr/corpus.

Table 3.1 Size and brief description for each of the twenty-six Yonsei corpora

Corpus	Size	Description
#1. Yonsei Corpus 1	2,900,000	It was collected from materials published from 1980 to 1987. What types of materials and how much for each type to collect were based on the actual reading habits of a thousand adult Koreans.
#2. Yonsei Corpus 2	1,100,000	With the aim of making a balanced corpus, it was mainly collected from books across ten categories from 1987 to 1988: Generic (7.8%), Philosophy (9.9%), Religion (10.7%), Social science (12.8%), Language (5.7%), Pure science (11%), Applied science (11.7%), Art (8.1%), Literature (11.2%), and History (11.3%). The proportions were determined by how frequently related books are checked out for each category. The Dewey Decimal Classification was used for the classification of Korean literature into those ten categories.
#3. Yonsei Corpus 3	5,980,000	It was collected from materials selected as excellent publications in 1980s.
#4. Yonsei Corpus 4	770,000	It consists of real colloquial and quasi-colloquial languages. Specifically, it is composed of Dialogues (26%), Lectures (24%), Counsel (14%), Plays·Scripts (13%), DJ broadcasts (13%), Discussions (8%), Meetings (2%), and so on. It contains information about the age, gender, and occupations of speakers, the number of speakers, information about transcribers, the types of utterances, and recording time.
#5. Yonsei Corpus 5	8,600,000	It was collected from a range of literature in 1970s: Newspapers (10%), Fictions & Essays (50%), General books (35%), and Textbooks (5%).
#6. Yonsei Corpus 6	7,230,000	It was collected from literature in 1960s.
#7. Yonsei Corpus 7	13,670,000	It was collected from literature until the middle of 1990s with main focus on fiction and essays. It was constructed over the period from 1994 to 1995.
#8. Yonsei Corpus 8	870,000	It consists of teaching materials from every elementary school subject and the ones from the subjects of Korean and social studies in middle and high schools. Those materials are part of the 5th and 6th curricula.
#9. Yonsei Corpus 9	1500,000	It was collected from a sample of early childhood education books and was constructed in 1996.
#10. Yonsei Corpus 10	780,000	It is composed with separate volumes from the first period (1945-1965) corpus supplemented for the compilation of Yonsei contemporary Korean dictionary.

#11. Yonsei Corpus 11	730,000	It is composed with textbooks from the first period (1945-1965) corpus supplemented for the compilation of Yonsei contemporary Korean dictionary.
#12. Yonsei Corpus of Korean in the 20th Century	150,378,870	It is a raw corpus of written language which consists of 20th literature collected depending on publication dates and text types.
#13. Corpus of Korean Textbooks (Complete)	724,856	It was collected from Korean textbooks from Korean language education institutes in 1990s.
#14. Corpus of Korean Textbooks (Conversation)	119,598	It was collected from dialogues in the introductions of Korean textbooks from Korean language education institutes in 1990s.
#15. Yonsei Korean Learner Corpus	278,542	As a Korean learner corpus, it was collected from compositions by students at the Yonsei Institute of Language Research and Education.
#16. Korean Elementary Textbook Corpus after Independence	1,496,280	It was collected from elementary school Korean language textbooks published after the period from 1945 to 1954.
#17. The 6th and 7th Korean Elementary Textbook Corpus	1,681,769	It was collected from textbooks in the 6th and 7th curricula. It provides annotations on homonyms.
#18. Yonsei Balanced Corpus of Written Discourse	1,054,362	It is a corpus of written language composed of texts from a range of genres.
#19. Yonsei Balanced Corpus of Spoken Discourse	998,934	It is a corpus of spoken language collected from monologues and public & private dialogues.
#20. Yonsei Corpus of Polysemy	1,165,224	It is a corpus providing annotations on polysemy, which was constructed for a Korean meaning frequency dictionary.
#21. Yonsei Corpus of Hangul tripitaka	386,472	It was collected from the doctrines of Buddhism, Buddhist scriptures, Tripitaka, and so on.
#22. Corpus of <Tongnip Sinmun> Newspaper	144,309	It was constructed based on <Tongnip Sinmun>, an early Korean newspaper and the first privately managed modern daily newspaper in Korea.
#23. Corpus of Popular Songs in the Modern Era	29,339	It was collected from the lyrics of popular songs in 1930s and 1940s.
#24. Yonsei Corpus of Multimodal Data	18,986	It is composed of videos of utterances, transcription texts, and annotations on non-linguistic actions.
#25. Twitter Corpus	945,175,620	It was collected from Korean tweets which were written for one month, October 2011.

#26. Political Discourse Corpus	306,681	It was constructed for discourse analysis with the topic limited to politics.
Total	1,148,089,842	

3.1.1.2 Sejong Corpora

The Sejong Corpora are the most important landmark in Korean corpora. They are national Korean corpora, open to the public. They were built by Korea University and Yonsei University under a 12-million-dollar government-sponsored national project named the 21st Century Sejong Project, which was performed from 1998 to 2007. Constructing a large-scale national corpus comparable to the British National Corpus in the UK was one of the goals that the project pursued in order to promote the development of language research and technology in South Korea. When the project finished, the Sejong Corpora were distributed on DVD in the beginning, with each version updated four times, 2007, 2009, 2010, and 2011, respectively. They are no longer distributed on DVD.

The size of the Sejong Corpora is about 200 million eojuls. They consist of seven corpora: written modern Korean, spoken modern Korean, North Korean/Korean used abroad, old Korean, Korean-English parallel corpora, Korean-Japanese parallel corpora, and Korean terminology. Table 3.2 shows the format and size for each corpus in the Sejong Corpora (see Kang 2008 for more detailed information about the Sejong Corpora). In the table, a raw corpus refers to a corpus of original text. A tagged corpus means a corpus with morphological information added to the raw corpus. A word-disambiguated corpus is a semantically tagged corpus with disambiguated sense information added to the tagged corpus. A treebank is a corpus with syntactic structural information added to the word-disambiguated corpus.

Table 3.2 Format and size for each corpus in the Sejong Corpora

Category		Format	Size (million eojul)
Modern Korean	Written	Raw	62.0
		tagged	15.0
		word-disambiguated	12.5
		treebank	0.8
	Spoken-transcript	Raw	3.7
		tagged	1.0
North Korean/Korean abroad		Raw	9.5
		tagged	1.6
Old Korean		Raw	5.6
		tagged	0.9
Parallel	Korean-English	Raw	4.8
		tagged	1.0
	Korean-Japanese	Raw	1.1
		tagged	0.3
Korean terminology		Raw	75.0
Total			194.8

- Written and spoken modern Korean corpora: They consist of materials after 1910s. The corpus of written modern Korean was collected from various types of texts such as newspapers and magazines. The corpus of spoken modern Korean was collected from monologues and dialogues. The monologues and dialogues are divided into two subcategories respectively: public and private.
- North Korean/Korean abroad corpora: They were constructed from Korean texts used in North Korea, China, and Commonwealth of Independent States.
- Old Korean corpora: They were collected from Korean texts from the 15th century to the beginning of 20th century.
- Korean-English and Korean-Japanese parallel corpora: They contain both source language and their translated language texts.
- A Korean terminology corpus: It was built from professional texts in various fields.

When the Sejong Corpora were released, the size was large enough to match corpora in the United States, the UK, Japan, and so on (their sizes were 200 million to 500 million words back then). However, they have been left behind since 2007 when the project stopped. Since the 21st Century Sejong Project finished in 2007, the Sejong Corpora have not been updated with new data so they are considered out of date now. Although a number of new words have been created and commonly used for about fifteen years after 2007, the Sejong Corpora do not include them.

A corpus has to continue to grow to reflect the dynamics of language change over time because it cannot be considered as a representative corpus unless it is regularly updated (Hunston 2002). It is very important to keep track of new language data and include it in a corpus so that it can be a representative corpus, especially when the corpus is used for special tasks such as the establishment of language policies, the improvement of efficiency in language education, and information processing.

Fortunately, the National Institute of the Korean Language has recently constructed a Web Corpus consisting of language data on the Web such as social media platforms and blogs. They aimed to collect two million posts from social media platforms, ten thousand posts from blogs, ten thousand posts from bulletin boards, and a hundred thousand posts from reviews like comments on products, with the investment of sixty million dollars in the data collection. In addition to the Web Corpus, the National Institute of the Korean Language has built many different kinds of additional corpora. Those corpora will be able to supplement the Sejong Corpora and further be widely applied to the field of artificial intelligence in South Korea. They are available at https://corpus.korean.go.kr/#down.

3.1.1.3 Other Korean Corpora

The Trends 21 Corpus is also remarkable in terms of size. It was built by Research Institute of Korean Studies at Korea University. It consists of 600 million eojuls and was collected from texts from four main daily newspapers in South Korea (i.e., The Dong-a Ilbo, The Chosun Ilbo, JoongAng Ilbo, and The Hankyoreh), which are materials for fourteen years from 2000 to 2013. In addition to the Yonsei Corpora, the Sejong Corpora, and the Trends 21 Corpus, many different kinds of corpora have been constructed by a large number of research institutes and universities depending on various research objectives. For example, Newspaper Corpus, Chinese-English-Korean Multilingual Corpus, Korean Tree-Tagging Corpus, Automatically Analyzed Large Scale KAIST Corpus, Terminology Corpus, and so on were built by the Korea Advanced Institute of Science and Technology (http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus for access to those corpora). Korean Sentiment Analysis Corpus was constructed by Seoul National University, the Korean-German parallel Corpus by Hankuk University of Foreign Studies, the Corpus of Historical Materials by Kyung Hee University, and the Korean Resident in Japan Corpus by Jeju National University. However, the size of all these corpora is much smaller than that of the Yonsei corpora, the Sejong Corpora, and the Trends 21 Corpus.

3.1.2 Construction of a Korean Twitter Corpus using Python

The Yonsei Twitter Corpus in the Yonsei Corpora is a Korean Twitter corpus which boasts a large size. It was built for the analysis of political inclinations in collaboration with a social media analytics company. It consists of tweets during the period of the Seoul mayoral election campaign in 2011. The tweets were randomly collected with no specific keywords. The Yonsei Twitter Corpus (945,175,620 eojuls) accounts for about 82 percent of the entire Yonsei Corpora

(1,148,089,842 eojuls). In terms of size, the Yonsei Twitter Corpus has made a great contribution to Korean corpora.

However, the Yonsei Twitter Corpus is not open to the public and it is an old corpus now because the tweets in the corpus are tweets written for a single month, October 2011. With the old and short-term corpus, it is impossible to study on language change. Accordingly, this study builds a new Korean Twitter corpus on the basis of up-to-date data over a longer period. For the construction of the new Korean Twitter corpus, this study collects tweets containing each of the three neologisms as a keyword from 2010 to 2019 (in the case of *lwuce,* to July in 2021). The following subsections show how the new Korean Twitter corpus was constructed and preprocessed using Python.

3.1.2.1 Data Collection

Twitter is a microblogging social platform that allows users to share posts of up to 280 characters. Despite the character limit of posts, it is hard to collect a great number of tweets manually using computer commands of "control+C" and "control+V", which is time-consuming and energy-consuming. As a more efficient method to collect a large number of tweets from Twitter, I chose Python because it automatically scrapes tweets that I want to collect within a short time. Python is one of the most popular computer programming languages. It is widely used by big companies like Google, Instagram, and IBM. Because Python is relatively easy to learn, many non-programmers are learning it.

Python has a number of useful libraries which can be used to perform various tasks. A library offers a collection of packages, and each package is a collection of modules. A module is a Python file containing various Python functions and variables. Thus, a library is a set of code created to perform a certain task. Thanks to already-made libraries/packages, users do not need to

write Python code for themselves from scratch. However, users have to modify some of the Python code in already-made libraries/packages or write their own code depending on their specific research objectives.

In order to collect tweets from Korean Twitter, I used snscrape, which scrapes data from social networking services such as Facebook, Instagram, and Twitter. For Twitter, it has the advantage of allowing scraping tweets without personal API keys and returning thousands of tweets in seconds. Employing the Python library of snscrape, I collected tweets containing the three Korean neologisms as keywords (the specific period and number of scraped tweets for those neologisms are covered in Section 4.1, 5.1, and 6.1). Figure 3.1 shows Python code which collects data of *leyal* using snscrape.

```
import os
import pandas as pd
import time
from pandas import DataFrame

tweet_count = 15500
text_query = "레알"

for j in range (2010,2011):
    time_list = []
    length_list = []
    for i in range(1,13):
        print(j,i)
        start_time = time.time()
        since_date = "%d-%d-1" %(j,i)

        if i == 12:
            until_date = "%d-%d-1" %(j+1,1)

        else:
            until_date = "%d-%d-1" %(j,i+1)

        os.system('snscrape --jsonl --max-results {} --since {} twitter-search "{} until:{}"> ₩
        text-query-tweets.json'.format(tweet_count, since_date, text_query, until_date))
        tweets_df = pd.read_json('text-query-tweets.json', lines=True)
        length_list.append(len(tweets_df))
        tweets_df.to_csv('twitter_raw_data_%d.%d.csv' %(j, i), index=False, encoding='utf-8-sig')
        timevalue = time.time() - start_time
        time_list.append (timevalue)

    df_time_length = DataFrame({'Time': time_list, 'Length': length_list})
    df_time_length.to_csv('collection_time_length_%d.csv' %(j), index=False, encoding='utf-8-sig')
```

Figure 3.1 Python code which collects data of *leyal* using snscrape. This example code shows the process of scraping 15,500 tweets per month for the year 2010

For the construction of corpora, it is important to make them readily accessible. If a file size is too large it takes too much time to open the file, so I saved the scraped tweets in the form of month-to-month files (Figure 3.2). The files were saved in CSV format. Also, I grouped the monthly files by year for convenience (Figure 3.3). Each monthly file consists of twenty-one columns: url, date, content, renderedContent, id, user, outlinks, tcooutlinks, replyCount, retweetCount, likeCount, quoteCount, conversationId, lang, source, sourceUrl, sourceLabel, media, retweetedTweet, quotedTweet, mentionedUsers. The content column contains messages in tweets, which become the target of data preprocessing.

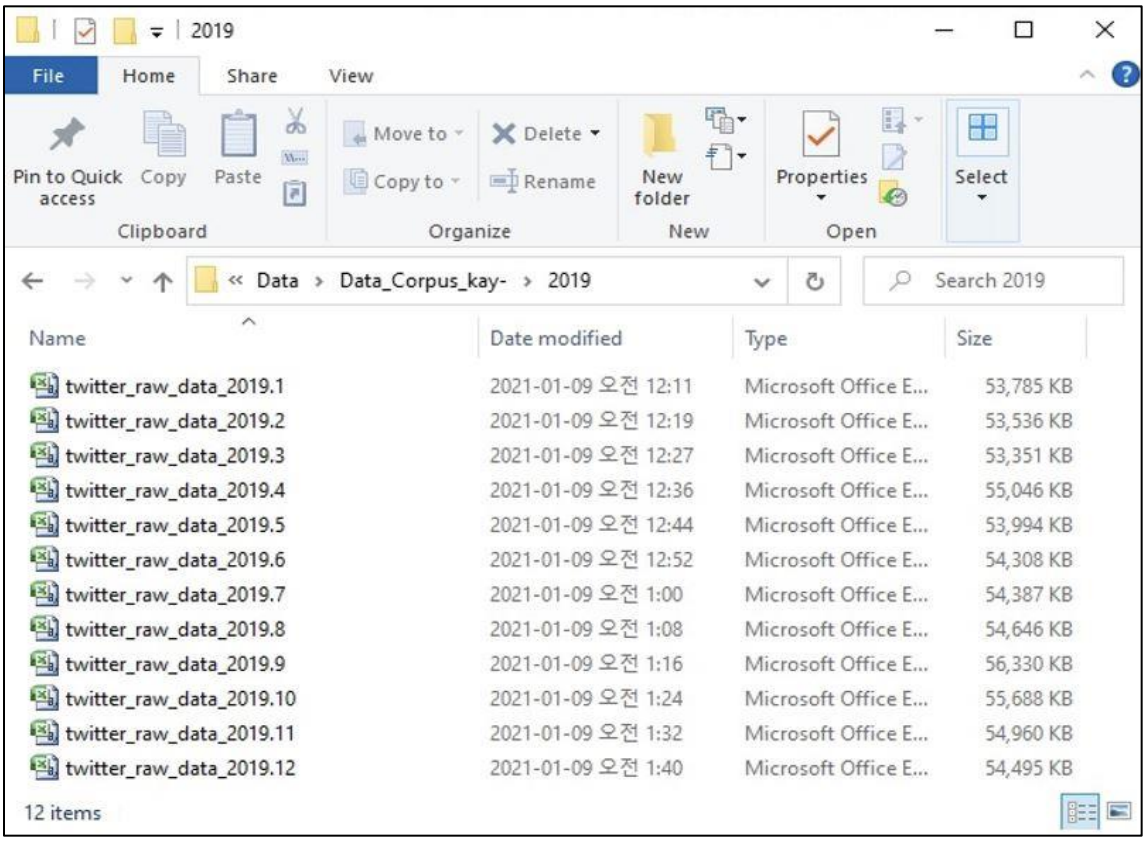


Figure 3.2 Scraped tweets saved in the form of month-to-month files. This example indicates the 2019 corpus out of ten yearly corpora from 2010 to 2019 of the neologism *kay-*.

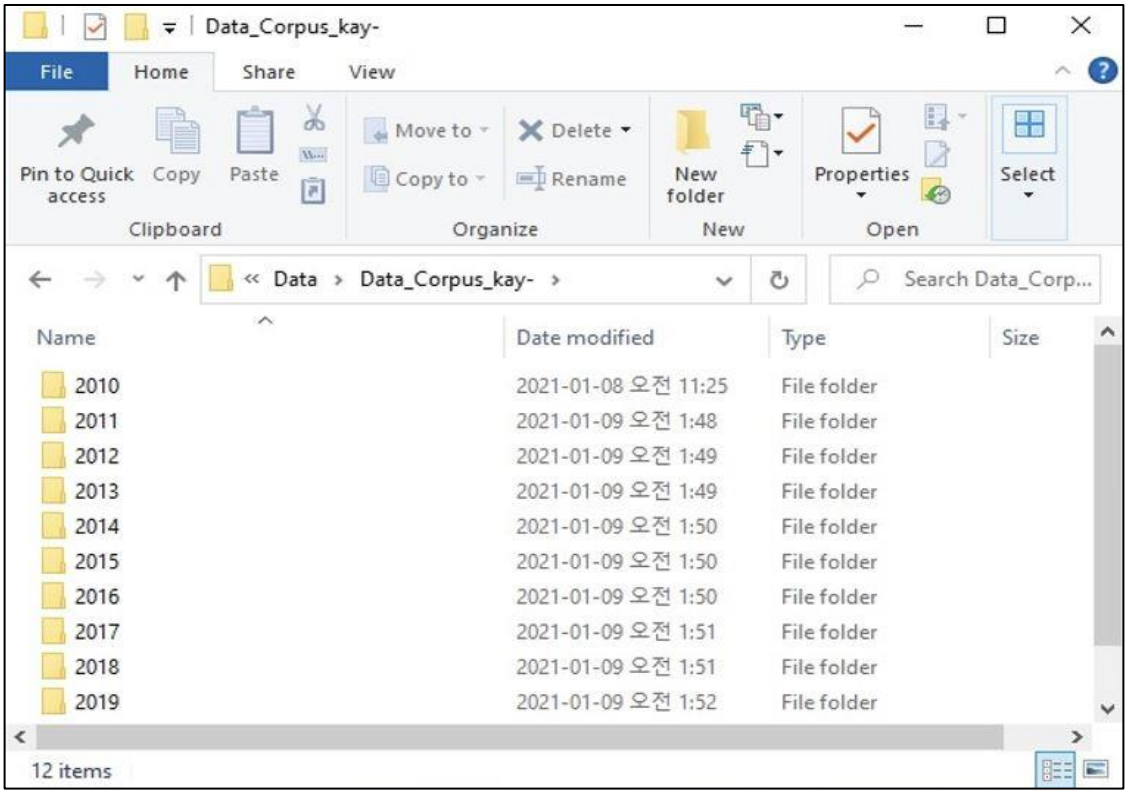


Figure 3.3 Monthly files grouped by year. This example shows ten yearly corpora from 2010 to 2019 of the neologism *kay-*.

```
>>> from konlpy.tag import Okt
>>> okt = Okt()

>>> print(okt.morphs(u'단독입찰보다 복수입찰의 경우'))
['단독', '입찰', '보다', '복수', '입찰', '의', '경우']

>>> print(okt.nouns(u'유일하게 항공기 체계 종합개발 경험을 갖고 있는 KAI는'))
['항공기', '체계', '종합', '개발', '경험']

>>> print(okt.phrases(u'날카로운 분석과 신뢰감 있는 진행으로'))
['날카로운 분석', '날카로운 분석과 신뢰감', '날카로운 분석과 신뢰감 있는 진행', '분석', '신뢰', '진행']

>>> print(okt.pos(u'이것도 되나욬ㅋㅋ', norm=True, stem=True))
[('이', 'Determiner'), ('것', 'Noun'), ('도', 'Josa'), ('되다', 'Verb'), ('ㅋㅋ', 'KoreanParticle')]
```

Figure 3.4 Example code using Okt

The Natural Language Toolkit (NLTK) in Python is a suite of libraries and programs for natural language processing but it is not designed for the Korean language. Therefore, KoNLPy is used for data preprocessing. KoNLPy is a Python package for natural language processing of the Korean language. It has various Korean morpheme analyzer tools. Among them, Open Korean Text (Okt) is employed in this study. With Okt, it is possible to split data into morphemes, tag morphemes with their parts of speech, reduce morphemes with inflections to their stems, and only extract data that a researcher needs from an entire dataset.

Figure 3.4 shows example code[5] using Okt. The first two lines function to bring Okt. The third line carries out splitting the data into morphemes, the fourth and fifth lines perform only extracting nouns and phrases from the data respectively, and the last line conducts tagging each morpheme with its part of speech after splitting the data into morphemes, normalizing those morphemes, and making the morphemes with inflections their stems. The results are shown in square brackets following the command.

3.1.2.2 Data Preprocessing

A raw corpus of tweets scraped from Twitter is not user-friendly because it contains unnecessary things for this study such as punctuation marks and other foreign languages, has no annotation on grammatical information like parts of speech, and lacks consistency in text. Therefore, data preprocessing is required to derive significant information from the raw corpus more efficiently. Data preprocessing comprises the following six processes: tokenization, normalization, stemming, cleaning, tagging, and removal of stop words. In this study, the four processes, i.e., tokenization,

[5] The source of this example code is as follows: https://konlpy.org/ko/latest/api/konlpy.tag/#okt-class.

normalization, stemming, and tagging are performed by means of the Korean morpheme analyzer Okt in KoNLPy, which is an open source Korean tokenizer developed by Will Hohyon Ryu.

Tokenization is the process of splitting the entire text into tokens. Depending on researchers, tokens can be words, phrases, sentences, and so on. In this study, the entire text was split into morphemes. Normalization involves changing to lowercase/uppercase, expanding abbreviations, and so on so that different forms of the same word can be recognized as the same word. The normalization of Korean is different from that of English because, for example, there is no distinction between lowercase and uppercase in the Korean language. In the normalization of Korean, different variations from the same word created by the use of different vowels or addition of unnecessary consonants are converted into their original form. For instance, variations from the Korean greeting meaning 'hello/hi', *annyenghaseyyos* (*안녕하세욧*) and *annyenghaseyyong* (*안녕하세용*) are converted into their original form *annyenghaseyyo* (*안녕하세요*). They are cases where additional consonants are added to the original form. Stemming is the process of reducing a word to its stem or root form. To take the same example, the Korean greeting *annyenghaseyyo* (*안녕하세요*) is a polite form so it is reduced to its stem *annyenghata* (*안녕하다*). Cleaning is the process of correcting or removing incomplete/incorrect/irrelevant data (e.g., punctuation marks, numbers, and other foreign languages).

Tagging is the process of tagging each token with parts of speech (e.g., noun, verb, adjective, or adverb). The POS tagger in the tag package of Okt splits the entire text into morphemes, tags each morpheme with its part of speech, normalizes variations of the morphemes, and reduces the morphemes with inflections to their stems or root forms[6]. The POS tagger has the

[6] e.g. 비가 온다. NM: Nominative case particle
pi-ka *o-n-ta.* IN: Indicative mood suffix
rain-NM come-IN-DC DC: Declarative sentence-type suffix (See the next page.)

following twenty-four tags: Noun, Verb, Adjective, Adverb, Determiner, Modifier, Conjunction, Exclamation, Josa, PreEomi, Eomi, Prefix, VerbPrefix, Suffix, Punctuation, Foreign, Alpha, Number, Unknown, Korean Particle, Hashtag, ScreenName, Email, and URL (some of the tags such as Noun and Verb are actual parts of speech while others such as Hashtag and Email are not). The number, kind, and name of POS tags vary depending on morpheme analyzer tools. The information of parts of speech is very important for the analysis of semantic prosody of the neologism *kay-* because it makes it possible to extract only adjectives and verbs where the neologism *kay-* is attached from tweets.

Stop words are a set of commonly used words which have very little meaning, for example, like *and*, *the*, and *a*/*an* in English. The removal of stop words allows users to focus on more critical words. For instance, a search engine allows us to do so by presenting more pages about relatively more important words (i.e., content words) than commonly used words (i.e., function words) when we type phrases or sentences into the search engine. In order to make a corpus containing more critical words and obtain clearer results from a corpus, the process of removing stop words is very crucial.

3.2 Artificial Intelligence

Artificial intelligence (AI) has been paid much attention today. AI refers to intelligence displayed by machines. It simulates human intelligent behaviors by means of computers and trains computers to perform intelligent tasks such as learning, decision-making, and judgment (Da Xu, Lu & Li

---

The example at the bottom of the preceding page is the one mentioned in Section 3.1, meaning 'it is raining.'. Let us apply tagging by means of Okt to this example. The first eojul *pi-ka* is split into two morphemes *pi* and *ka*. They are labeled "noun" and "josa" respectively. Josa is a Korean postposition. Korean postpositions are known as case markers. The morpheme *ka* is used to indicate that the preceding noun is subject. The second eojul *o-n-ta* is reduced to its stem *ota* and labeled "verb".

2021). AI has permeated many various aspects of our life at high speed, replacing existing manual work with automatic processing. AI applications include autonomous vehicles such as drones and self-driving cars (e.g., Tesla), virtual assistants understanding human speech (e.g., Alexa or Siri), advanced web search engines (e.g., Google search), computer programs playing games such as chess and Go (e.g., AlphaGo), recommendation systems (e.g., systems in Amazon, Netflix, or YouTube), image recognition, medical diagnosis, creating art such as poetry and stories, spam filtering, predicting flight delays, and so on.

Two terms frequently mentioned in AI are machine learning and deep learning. Machine learning is part of AI and deep learning is a subset of machine learning on the basis of artificial neural networks. Machine learning is the study of computer algorithms that can learn, predict, and improve their performance through experience, i.e., past information or data. The first definition of machine learning was given by Arthur Samuel, an American pioneer in AI and computer gaming, in 1959. According to him, machine learning is defined as "the field of study that gives computers the ability to learn without being explicitly programmed". Machine learning mainly depends on methods and models based on statistics and probability theory to solve practical problems.

Machine learning is classified into three broad categories depending on the nature of output or feedback available to the learning algorithm: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is learning a function that maps an input to an output through example pairs consisting of inputs and their desired outputs. In other words, learning algorithms analyze labeled training datasets which are composed of example pairs and create the function which can map new examples to their correct outputs (i.e., determine the class labels for unseen instances or predict outcomes accurately).

Unsupervised learning involves learning patterns from unlabeled, unclassified, or uncategorized data. Learning algorithms take a dataset that only contains inputs and find structure in the dataset on their own without any external guidance or intervention. The learning algorithms identify similarities or commonalities in the dataset and react based on whether such similarities or commonalities are in each new piece of dataset or not. Clustering or grouping belongs to an unsupervised learning task.

Reinforcement learning concerns training learning algorithms to interact with a complex and dynamic environment to achieve a certain goal such as playing a game against an opponent or driving a vehicle. The learning algorithms learn solutions to problems by trial and error. That is, through rewards or penalties for the actions they perform, the learning algorithms learn what actions they should take so as to maximize the total reward. This is conceptually similar to genetic programming.

Deep learning is a subfield of machine learning related to artificial neural networks. Artificial neural networks refer to computing systems inspired by neural networks that constitute a biological brain. Artificial neural networks are built like a biological brain, with artificial neuron nodes connected to each other. Each connection transmits a signal to other neurons and the neurons receiving the signal process and send it to other neurons connected to them. One process of transmitting and receiving a signal constitutes one layer. As the process builds up, multiple layers are formed. The adjective “deep” in deep learning implies the use of multiple layers. The use of multiple layers in the network enables deep learning algorithms to perform tasks that machine learning algorithms cannot easily perform. Deep learning algorithms are trained in supervised, semi-supervised, or unsupervised learning.

The computational techniques to be employed in this study are algorithms based on deep learning. Word2vec and Latent Dirichlet Allocation are trained in unsupervised learning while long short-term memory is trained in supervised learning. The model of long short-term memory is used for binary classification and its performance is assessed through classification evaluation metrics. The classification evaluation metrics include accuracy, F1-score, a confusion matrix, a precision-recall curve, the area under the precision-recall curve (PR AUC or AUC-PR), among others.

3.2.1 Natural Language Processing

As we have to learn a foreign language to communicate with people in a country where that foreign language is used, we have to learn a computer programming language to communicate with a computer. This is because a computer cannot understand natural language, i.e., human language. In order to have computers process human language, natural language processing is necessary. As a subfield of computer science and artificial intelligence, natural language processing is concerned with programming computers to process and analyze large-scale natural language data. Applications of natural language processing include machine translation, information retrieval, question answering, and text mining.

Research on natural language processing started in the 1950s, and the introduction of machine learning algorithms in the late 1980s revolutionized natural language processing. In the 2010s, as deep neural-network-based machine learning algorithms became prevalent, approaches founded on deep neural networks were applied to natural language processing, replacing statistical natural language processing with neural natural language processing. Accordingly, the latest natural language processing systems have been designed on the basis of deep learning algorithms.

The following subsections introduce three neural natural language processing techniques, which are to be used for the analysis of semantic change of the three neologisms.

3.2.1.1 Word2vec and Cosine Similarity

Word2vec is a technique that uses a neural network model to represent words from a large corpus as vectors of real numbers. The biggest advantage of word2vec is that it considers the meaning of words when mapping words to vectors. That is, semantically similar words have similar vectors. Computers cannot understand human language so it is required to convert it to numbers so that they can process it. One of the ways to change letters to numbers (i.e., vectors) is one-hot encoding. This technique represents words as binary vectors composed of "0" and "1". For example, to represent the sentence *I like an apple* as one-hot vectors, first, the technique assigns an index to each word: *I*: 0, *like*: 1, *an*: 2, *apple*: 3. Next, it assigns "1" to the position for the index number and "0" to the others as follows: *I* = [1,0,0,0], *like* = [0,1,0,0], *an* = [0,0,1,0], *apple* = [0,0,0,1]. However, this technique does not consider word meaning at all. Furthermore, it gets more inefficient when the size of a corpus gets large. The example sentence consists of four words so the number of dimensions of the one-hot vectors is four. When a corpus of 1,000 words is represented as one-hot vectors, the vectors have 1,000 dimensions with one "1" and nine hundred ninety-nine "0"s for each word. This demands much storage space on the computer.

Word embedding is a method developed to overcome those limitations of one-hot encoding. It represents words as real-valued vectors. One of the techniques to perform word embedding is word2vec, which represents words as real-valued vectors under the distributional hypothesis that words with similar contexts have similar meanings. Word2vec uses context words of the target words to convert the target words into distributed representations (i.e., real-valued vectors). Accordingly, words which have similar meanings have similar vectors because of their similar

contexts and those vectors are located close to one another in the vector space. Also, as vectors have real numbers[7], the dimension of vectors can be reduced regardless of the vocabulary size. This makes the process of converting words to vectors more efficient.

The process of converting words to vectors involves the utilization of either of two model architectures: continuous bag-of-words (CBOW) or skip-gram (the latter is utilized in the semantic change analysis of *leyal* using word2vec and cosine similarity because it is known to produce better results than the former). In the architecture of CBOW, the model predicts the target word from its surrounding context words using pairs of (context word, target word). For example, if a given sentence is *the kids chase each other around the red table in the living room* and the window size is "2", the model has pairs like ([the, chase], kids), ([the, table], red), ([red, in], table), and so on. The CBOW architecture assumes that the order of context words does not affect prediction. Contrastively, in the skip-gram architecture, the model predicts its surrounding context words for a given word. To take the same example as CBOW, the model predicts the context words [the, chase] when the word *kids* is given, [the, table] for the word *red*, [red, in] for the word *table*, and so on. Word2vec utilizes either of these two architectures to learn relations among words in linguistic data.

[7] Unlike one-hot vectors, vectors from word2vec have real numbers (e.g., *apple* = [-0.1, -0.2, 0.0], *orange* = [-0.1, 0.5, 0.0], *car* = [2.1, 0.3, 1.5]). Please note that *apple* and *orange* have similar vectors. The values of these examples were created to help those who do not have background information about word embedding to understand real-valued vectors better. They are not actual values from word2vec.

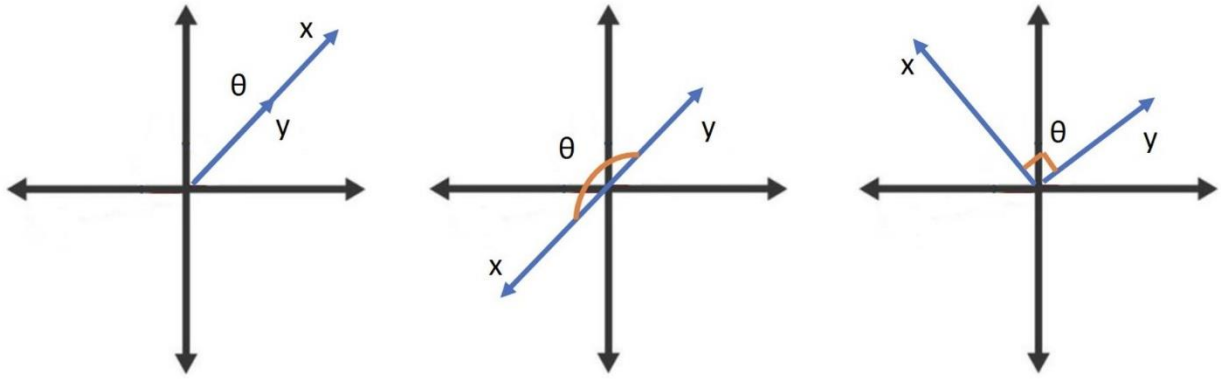


Figure 3.5 Cosine similarity. In the three figures, x and y mean vectors. From left, the first figure shows similar vectors ($\cos(\theta) = 1$), the second figure shows opposite vectors ($\cos(\theta) = -1$), and the third figure shows orthogonal vectors ($\cos(\theta) = 0$).

Semantic similarity between words can be calculated through cosine similarity (CS), which is a metric used to measure similarity between vectors. In mathematics and physics, vectors refer to quantities that have both magnitude and direction. CS uses the cosine of the angle between two vectors to measure the similarity between them. When two vectors are in the same direction (i.e., the angle between them is 0°), the CS value is "1". When two vectors are in the opposite direction (i.e., the angle between them is 180°), the CS value is "-1". When two vectors are perpendicular (i.e., the angle between two vectors is 90°), the CS value is "0". Figure 3.5 shows each case. The closer to "1" the CS value is, the more similar the two vectors are. The CS value quantifying similarity between two vectors indicates how semantically similar two words corresponding to the two vectors are to each other.

3.2.1.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative statistical model that discovers latent semantic structures in an extensive body of text. It classifies words by topics. In LDA, a topic is not an actual topic like the subject of an essay but an abstract topic, that is, a cluster of similar words not

containing any explanation of how the topic is developed. This means that we cannot gain topics in the form of words and have to infer them from words in each cluster. LDA is founded on Bayesian inference, which is a method of statistical inference where the probability for a hypothesis is updated through the continuous addition of available information according to Bayes' theorem.

Let us look at how LDA works. Every word in a document is randomly allocated to topics (the number of topics has to be decided by a researcher). This allocation is random so it is not correct in fact. However, with the information of random allocation, it is possible to calculate how the topics are distributed in documents (TD) and how words are distributed with those topics (WD). After such computations, with the first word excluded from the documents, TD and WD are calculated again. The recalculated values are used to calculate the probability that the first word belongs to each of the topics. The first word is allocated to the topic which has the highest probability among them. This process is applied to every word from the second word to the last one in the documents. It does not end in the first round but repeats enough so that every word can be correctly classified. As the probability that each word belongs to a particular topic is continuously updated, the classification of words by topics gets more accurate.

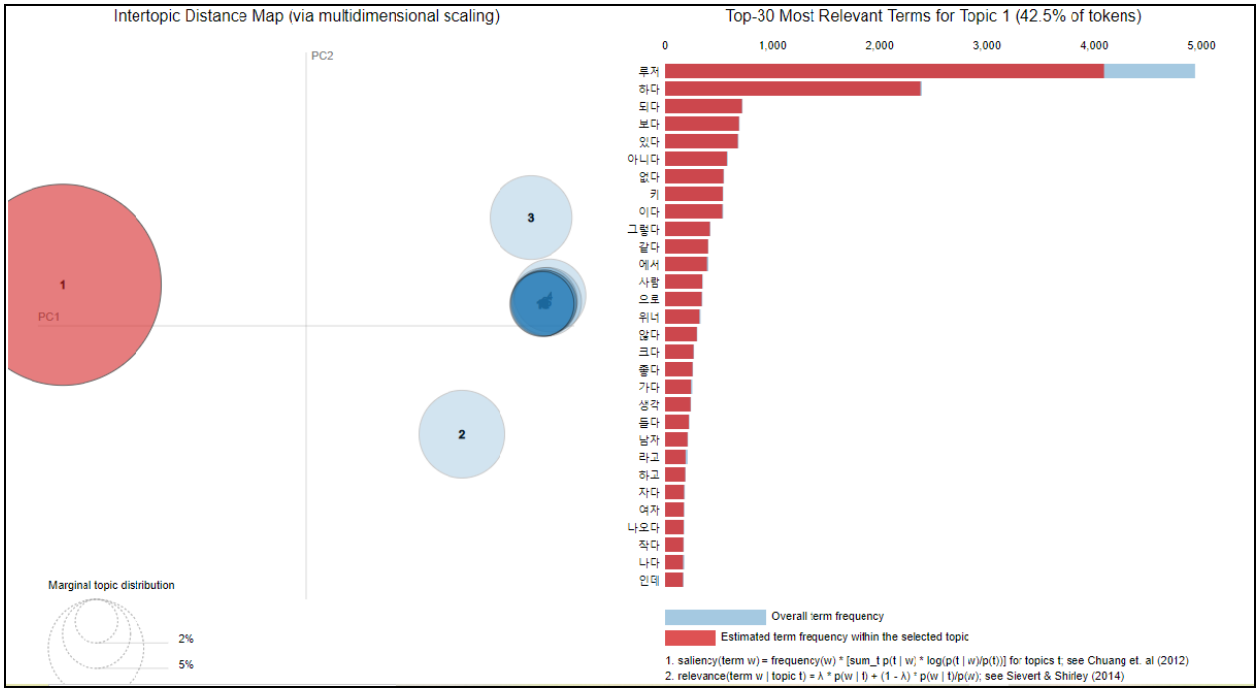


Figure 3.6 LDA result from 12,562 tweets written in 2010 containing *lwuce* as a keyword

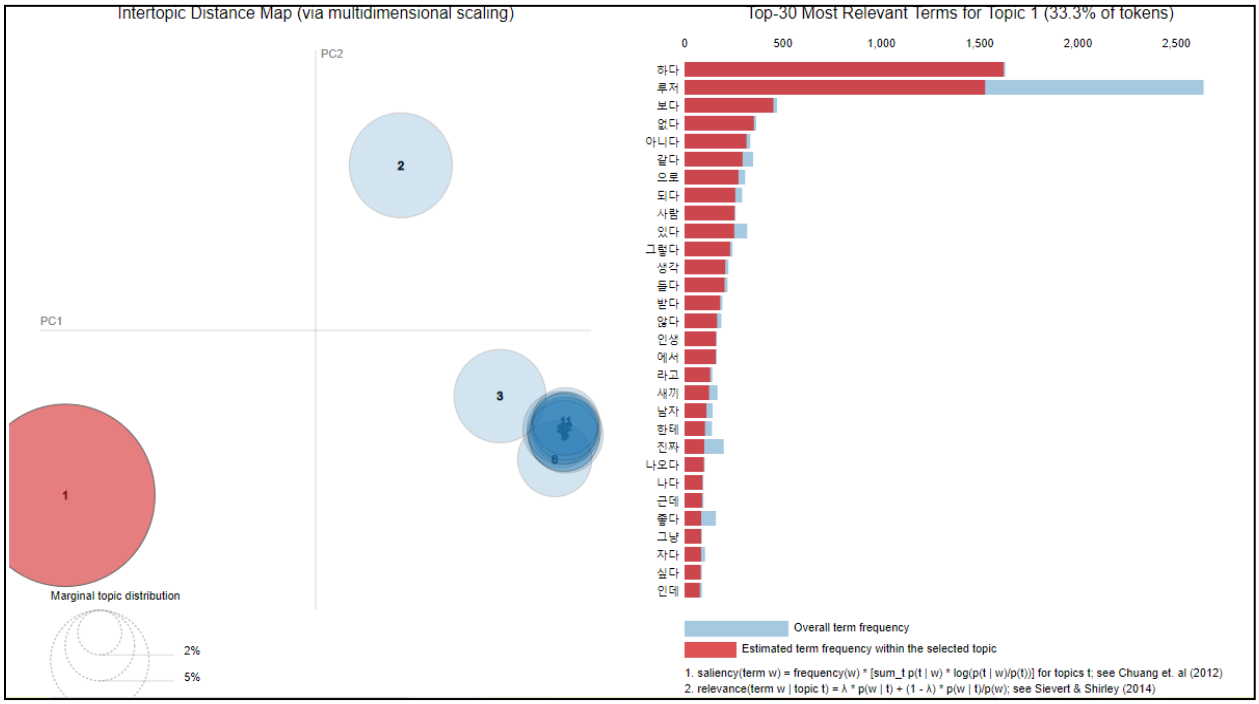


Figure 3.7 LDA result from 13,461 tweets written in 2019 containing *lwuce* as a keyword

Figure 3.6 and Figure 3.7 are the LDA results from 12,562 tweets written in 2010 and 13,461 tweets written in 2019 respectively. The tweets were collected with *lwuce* set as a keyword. The circles on the left side mean clusters of words, i.e., topics. The size of each circle indicates the proportion that the topic accounts for in the documents. Clicking on each topic shows the list of words related to that topic on the right side. Both Figure 3.6 and Figure 3.7 show a list of the top thirty words in Topic 1 that accounts for the highest proportion in the documents. The bar graphs indicate the frequency of each of the thirty words within Topic 1 and their frequencies in the entire document.

As we can see in Figure 3.6 and 3.7, LDA does not present what topics are so we have to infer them. However, it is not easy to infer the topics from words inside each cluster. Also, because LDA does not consider semantic similarity between words, words which have similar meanings can be allocated to different topics. That is, LDA classifies words using not word meanings but the information of how topics and words are distributed in documents. This can cause an issue with topic coherence. Accordingly, classification results from LDA can be different from those from humans. A way to assess the performance of LDA is to use a topic coherence measure, which is an evaluation method for topic models. It indicates how consistent words in clusters are in terms of meaning by measuring the semantic similarity among top words in a cluster for each cluster. It represents the average of topic coherence values from clusters. In this study, C_v topic coherence, one of the most popular topic coherence metrics, is used to evaluate the LDA performance.

3.2.1.3 Long Short-term Memory

Long short-term memory (LSTM) is a special type of recurrent neural network (RNN). A RNN is a class of artificial neural networks designed to handle and learn time series data where the sequence of data is important. Because a RNN can use information it processed in the past to

process future input, it has a wide range of applications. For instance, in natural language processing, it has been applied to text summarization, sentiment classification, spam detection, machine translation, and so on (e.g., Moholkar et al. 2019, Chandra & Khatri 2019, Kumar, Rao & Premnath 2020). Those applications require processing and learning long sequences of words. With linguistic data, a RNN captures and learns temporal features for words by considering the sequence of words.

However, a traditional RNN loses much information from past input when the sequence is long. To overcome this limitation, LSTM was developed. As an improved model of a traditional RNN, LSTM can better preserve information from past input. It is known to produce more accurate results when the gap between the relevant information and the point where it is needed is very large. In the following example *I traveled to New York last week. There were lots of things to see. The food was great, too. While staying there, my friend called me to ask where I was. I said I was in* _____, the blank requires information about location and that information can be predicted not from adjoining words but from the first sentence. As this example shows, when predicting the next word requires not neighboring words but the previous context, LSTM is useful in that it can make a good prediction from the past information well preserved in it (the processes of LSTM taking words, converting them into vectors, and handling those vectors are based on calculus and linear algebra).

The reason why LSTM produces better results than the traditional RNN is that LSTM has four components which the traditional RNN lacks: forget gate, input gate, cell state, output gate. They perform specific functions. The forget gate is in charge of determining how much previous information to forget. It removes some of the previous information which is unnecessary. The input gate decides how much current information to preserve. It is concerned with remembering

current information. The cell state is computed through the sum of results from the forget and input gates. Depending on each value from the two gates, how much previous and current information to preserve is finally determined. The output gate is concerned with determining how much current information to forget in the process of the next information with the cell state. Thanks to these four components, LSTM is able to process time series data more accurately.

3.2.2 Model Evaluation Metrics

The model of LSTM is trained to sort words into positive and negative for the analysis of semantic prosody of the neologism *kay-*. That is, LSTM is used as a binary classification model in this study. Evaluation metrics to assess the performance of classification models include accuracy, F1-score, a confusion matrix, a precision-recall curve, the area under the precision-recall curve (PR AUC or AUC-PR). The following subsections cover two evaluation metrics among them, i.e., a confusion matrix and a precision-recall curve (the notion of AUC-PR is briefly mentioned with the explanation of a precision-recall curve) because the two evaluation metrics are to be employed to evaluate the performance of LSTM. In machine learning, it is very important to build a generalized model which can correctly predict unseen data. For such a generalized model, it is required to estimate the accuracy of a model on unseen data through evaluation metrics. Accordingly, a confusion matrix and a precision-recall curve are used to assess how accurately LSTM can predict unseen data.

3.2.2.1 Confusion Matrix

A confusion matrix is a specific table layout which represents the prediction results of classification. The rows of the matrix stand for actual values while the columns stand for predicted values. Table 3.3 indicates a confusion matrix with two class labels, where the top left is True

Positive (TP), the top right is False Negative (FN), the bottom right is True Negative (TN), and the bottom left is False Positive (FP).

TP means instances where a model classifies an actual positive sample as positive, FN means instances where a model classifies an actual positive sample as negative, TN means instances where a model classifies an actual negative sample as negative, and FP means instances where a model classifies an actual negative sample as positive.

Each prediction belongs to one of these four outcomes. All correct predictions are placed in the diagonal of the table (TP and TN) and values outside the diagonal represent prediction errors. A confusion matrix has an advantage of making it easy to visually examine prediction results. In the analysis of the neologism *kay-*, a confusion matrix is used to evaluate the performance of LSTM by showing how LSTM predicted the given data.

Table 3.3 Confusion matrix with two class labels

		Predicted Values	
		Positive	**Negative**
Actual Values	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

3.2.2.2 Precision-Recall Curve

To understand the notion of precision-recall curve, we should know what precision and recall are. Precision refers to the fraction which True Positive accounts for out of the instances that a model classifies as Positive, which is related to correctness. It shows how accurately a model predicts Positive and is known as Positive Predictive Value (PPV). In contrast, recall refers to the proportion of Positive instances which a model has found out of actual Positive samples, which is

associated with completeness. It is known as sensitivity. How to calculate precision and recall is as follows (please refer to the above confusion matrix with two class labels):

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP + FP}} \quad \& \quad \textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP + FN}}$$

Both precision and recall are focused on actual Positive samples and instances that a model classifies as Positive. They have different viewpoints on Positive. Precision looks into Positive from the perspective of a model while recall looks into Positive from the perspective of actual Positive samples. They serve to complement each other and the higher the values of precision and recall are, the better a model is. They are known to be appropriate for data the classes of which are imbalanced.

The training dataset used to train the LSTM model consists of positive and negative words/phrases but the ratio of positive words/phrases to negative ones is not balanced. Because the number of positive words/phrases is fewer, it is needed to check whether the model has been well trained in terms of positive words/phrases (the minority group). The application of precision and recall to the minority group is as follows:

$$\textbf{Precision} = \frac{\textbf{962}}{\textbf{962 + 9}} = \textbf{0.99} \quad \& \quad \textbf{Recall} = \frac{\textbf{962}}{\textbf{962 + 5}} = \textbf{0.99}$$

(TP: 962, TN: 1950, FP: 9, FN: 5)

A total of 2,926 words/phrases consists of 1,959 negative words/phrases (TN + FP) and 967 positive words/phrases (TP + FN). The number of positive words/phrases that the model has detected is 962 (TP) among the 967 positive words/phrases (FN means the case where the model

classifies positive words/phrases as “negative”). The number of words/phrases that the model classifies as “positive” is 971 (TP + FP). However, 962 words/phrases (TP) are actually positive and 9 words/phrases (FP) are negative words/phrases (FP means the case where the model classifies negative words/phrases as “positive”). The precision and recall values show that the model can predict the minority group, i.e., positive words/phrases accurately (99% accuracy) and detect almost every positive word/phrase (99% detection). These values come from the last fold among five folds so they are different from the result of normalized confusion matrix based on all the five folds in Section 6.5.2 (I used the last fold as an example for simplification and the detailed information on five folds is covered in Section 6.3.1).

A precision-recall curve is a plot of the y-axis of precision against the x-axis of recall. It shows the trade-off between precision and recall for a predictive model. The area under the precision-recall curve (AUC-PR) is an indicator showing a model’s performance. Both a precision-recall curve and the value of AUC-PR are obtained by means of a Python library. The higher the AUC-PR value is, the higher both precision and recall are. High precision and high recall mean that a model returns accurate results and detects a majority of actual Positive samples respectively.

CHAPTER 4

DISTRIBUTIONAL CORPUS ANALYSIS OF KOREAN NEOLOGISM *LEYAL*

4.1 Data Collection

The neologism *leyal* is mainly used to represent 'really' in informal settings, in particular on the Internet by young people. In order to collect data of the neologism *leyal*, I chose Korean Twitter. According to Shi (2020), the Korean language on Twitter is close to spoken Korean. More ungrammatical expressions and slang are used on Korean tweets, compared to written Korean in formal writings such as editorials and news articles. Also, Korean tweets are composed of short and less refined sentences. Given that Korean tweets have the characteristics of spoken Korean and the neologism *leyal* is slang, I chose Twitter as a source of data for the neologism *leyal*. Specifically, I scraped 15,500 Korean tweets that included *leyal* as a keyword per month from 2010 to 2019 (except for 2010, 2015, 2016, and 2019) using snscrape in Python. The number of tweets scraped for each year was 186,000, except for the four atypical years. For the four years that are exceptions, 175,953 tweets, 183,892 tweets, 178,569 tweets, and 177,202 tweets were scraped respectively with the numbers of monthly collected tweets uneven. The four years are estimated to have insufficient numbers of tweets about *leyal*. On the basis of the scraped tweets, I constructed ten yearly Twitter corpora corresponding to each of the ten years from 2010 to 2019.

4.2 Data Preprocessing

The ten yearly Twitter corpora went through the following six steps for data preprocessing:

- **Step 1**: The same tweets scraped several times were reduced to one tweet.
- **Step 2**: Unnecessary things for this study such as numbers, punctuation marks, and other foreign languages were removed from tweets.
- **Step 3**: Once again duplicate tweets were reduced to one tweet in order to filter retweets, which contain the same messages under different user IDs. Retweets are not filtered in the first step because of different user IDs so this step was added.
- **Step 4**: In tweets containing *leyal* and *matulitu* consecutively, when there was spacing between them, the spacing was removed. This is because the spacing causes the word *leyalmatulitu* representing the Spanish football club 'Real Madrid' to be recognized as two separate words.
- **Step 5**: Each sentence in tweets was divided into morphemes and the morphemes were converted to their stems.
- **Step 6**: Stop words were removed from the tweets.

I carried out the above six steps in Python version 3.7. For the fifth step, I employed the Korean morpheme analyzer Okt in the Python package of KoNLPy. For the sixth step, a list of Korean stop words containing 605 stop words[8] was applied in order to do an analysis based on content words. Stop words are words which have very little meaning while they are commonly used, for

[8] The number of English stop words is 421, which comes from the Brown stop words list known as the standard stop list (please note that the set of 421 stop words is one of many stop words lists and there is no single set of English stop words). The reasons why the number of Korean stop words is more than that of English stop words are inferred from the following two facts: (i) the Korean language has more words than English (https://en.wikipedia.org/wiki/List_of_dictionaries_by_number_of_words for more detailed information); (ii) compared to English, Korean verb conjugations are much more complex and there are more various verb endings (Korean verb conjugations depend upon the verb tense, aspect, mood, and the social relation between speakers and listeners). In natural language processing, the number of stop words can vary depending on how analysts define them.

example, like *the* and *a*/*an* in English. Table 4.1 shows the number of morphemes[9] in each of the ten preprocessed Twitter corpora after the six steps.

Table 4.1 Number of morphemes for each Twitter corpus

Year	Number of morphemes
2010	2,660,854
2011	2,632,336
2012	2,544,611
2013	2,414,255
2014	2,536,169
2015	2,563,663
2016	2,580,151
2017	2,779,693
2018	2,980,972
2019	3,012,924

4.3 Methodology

In order to observe how the two meanings of the neologism *leyal* have developed, I measured semantic similarity between the neologism and two alternative Korean words representing each of the two meanings (*cincca* representing 'really' and *leyalmatulitu* representing 'Real Madrid'). This is because which word between *cincca* and *leyalmatulitu* has higher semantic similarity to *leyal* indicates which meaning of *leyal* is more dominant. In this study, semantic similarity between two

[9] In Chapter 3, it was mentioned that eojuls are used to describe the size of Korean corpora because Korean corpora consist of eojuls. However, the preprocessed Twitter corpora in this study are composed of morphemes since each sentence in tweets was divided into morphemes in the process of preprocessing. The reason why not eojuls but morphemes are used as tokens in this study is to remove stop words such as particles and endings from the corpora (in eojuls, particles and endings are attached to content words so it is not easy to remove them). Accordingly, the number of morphemes is used to describe the size of the preprocessed Twitter corpora.

words is measured using morphemes[10] in the contexts of the two words. According to the distributional hypothesis that words with similar contexts have similar meanings, morphemes which co-occur with each of the two words are compared. The more morphemes the two words share, the more similar their meanings are. The comparison of their co-occurring morphemes between *leyal* and the two alternative words is performed through the DCA approach. That is, the vectors of *leyal*, *cincca*, and *leyalmatulitu* from word2vec are compared. Word2vec represents words as vectors under the distributional hypothesis (the context words of the target words are involved in the process of converting the target words into vectors) so the comparison of vectors implies an indirect comparison of lexical items which co-occur with the target words. If the values of two vectors are similar to each other, it means that their co-occurring words are similar as well as the two vectors are semantically similar. Also, the vectors are located close to one another in the vector space.

To be specific, there are two model architectures that word2vec utilizes to produce distributed representations (i.e., vectors) of words: continuous bag-of-words (CBOW) and skip-gram. I chose skip-gram which learns linguistic data in the way of predicting its context words for a given word. The preprocessed Twitter corpora were given to the skip-gram model and the model was set to learn its ten context words for a given word[11] (five words before and after the given word). Words whose frequency was less than ten were excluded from training. After training the skip-gram model to learn the preprocessed Twitter corpora, the semantic similarity between words was calculated through cosine similarity (CS), which is a metric used to measure similarity between vectors. The value of CS between two vectors indicates how semantically similar the two

[10] The ten preprocessed Twitter corpora are composed of morphemes so analyses employing word2vec and cosine similarity are based on morphemes. However, as it does not matter to clearly distinguish the two terms (morpheme and word), they are used interchangeably in this study.
[11] Every word in the corpora is used as a given word.

vectors (i.e., words) are to each other on the basis of the cosine of the angle between the two vectors. The closer to "1" the value of CS is, the more similar the meanings of the two words are. For each of the ten Twitter corpora, I investigated CS values between *leyal* and *cincca* and those between *leyal* and *leyalmatulitu*.

4.4 Results

Table 4.2 indicates the values of CS between *leyal* and *cincca* (CS1) and those between *leyal* and *leyalmatulitu* (CS2) for each year from 2010 to 2019. The CS values between *leyal* and *cincca* are closer to "1" for all the ten years while none of those between *leyal* and *leyalmatulitu* are beyond "0.306". This means that the meaning of the neologism *leyal* is more similar to that of *cincca*, which implies that the neologism *leyal* has been more used to represent 'really' than 'Real Madrid'.

Figure 4.1 shows how the values of CS have changed across time. The CS values between *leyal* and *cincca* relatively remain stable whereas those between *leyal* and *leyalmatulitu* more frequently and more significantly rise and fall. This indicates that *leyal* representing 'really' has been consistently used but the use of *leyal* as 'Real Madrid' is relatively unstable and has not been strongly entrenched[12].

The change in the CS values over time shows that each meaning of *leyal* subtly changes from year to year. However, in terms of the relation between the two meanings, there is no change because the meaning 'really' has been more dominant than 'Real Madrid' throughout the ten years.

[12] It can be argued that the CS values between *leyal* and *leyalmatulitu* are affected by the number of fans of the Spanish football club Real Madrid. The more fans the club has, the more tweets about the club there are on Twitter. The number of tweets from those fans can affect the CS values but the fluctuating values by the number mean that the meaning of 'Real Madrid' has not been strongly entrenched yet. If it had been entrenched, we could have seen almost consistent CS values as 'really'.

That is, we can say that there is no semantic change from the perspective of the relationship between the two meanings.

Table 4.2 Values of cosine similarity between *leyal* and *cincca* (CS1) and those between *leyal* and *leyalmatulitu* (CS2) for each year

Year	CS1	CS2
2010	0.763	0.263
2011	0.780	0.266
2012	0.782	0.306
2013	0.788	0.239
2014	0.840	0.290
2015	0.792	0.228
2016	0.801	0.169
2017	0.807	0.227
2018	0.835	0.225
2019	0.832	0.203

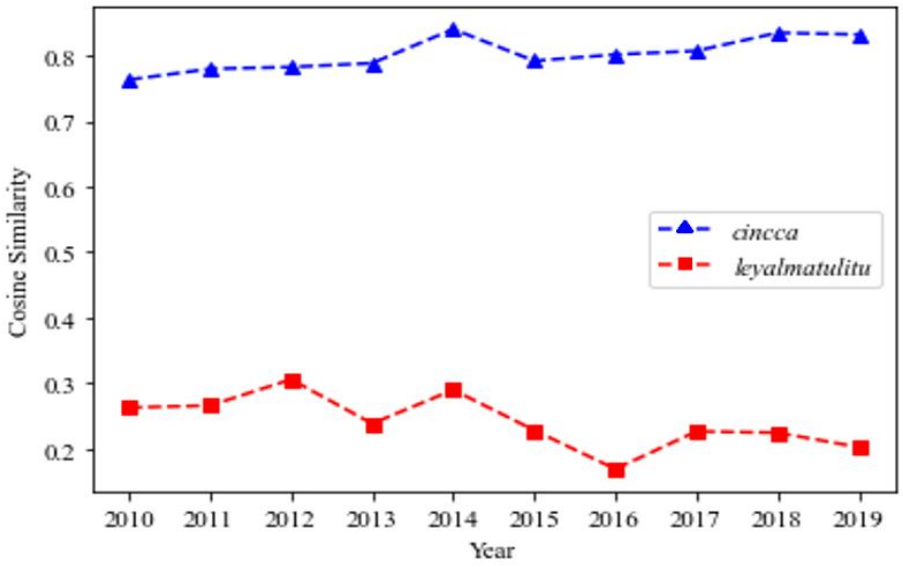
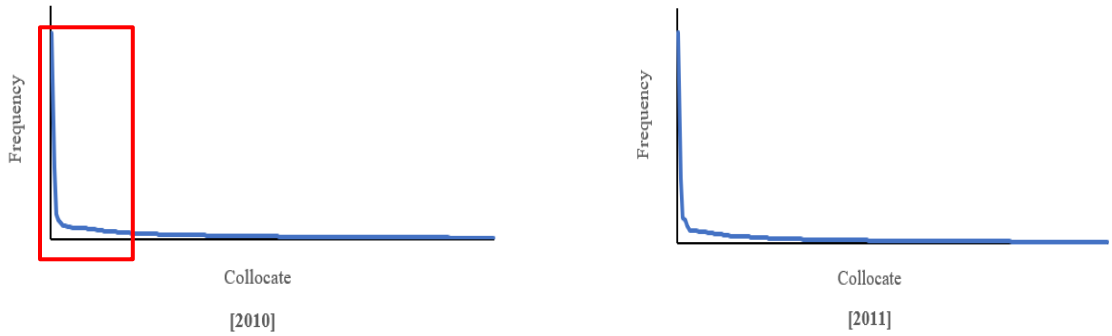


Figure 4.1 Change in the values of cosine similarity between *leyal* and *cincca* and those between *leyal* and *leyalmatulitu* over time

4.5 Method Validation

As the internal workings of word2vec are not transparent, we can doubt whether word2vec has been well trained and represented words as vectors based on their meanings. In particular, since word2vec is trained in unsupervised learning, there are no metrics to evaluate its performance. Therefore, it is needed to check whether word2vec has been well trained and the results from word2vec are reliable. For method validation, I decided to use a collocation analysis in corpus linguistics. Compared to word2vec, the method of analyzing collocates is simpler, clearer, and more intuitive (despite these advantages, why word2vec is preferable is presented in the following section). Unless the results from word2vec and those from the collocation analysis agree, we have no choice but to raise a question about the DCA approach and the results from it.

Collocates for the collocation analysis were obtained by means of LancsBox, which is a software tool for corpus analysis. I looked for collocates within three words of *leyal* that occurred a minimum of five times and set the Log Dice threshold value to "7.0". Log Dice is one of the association measures used to identify collocations. Because it operates on a scale with a fixed value, it is possible to directly compare Log Dice across different corpora (Gablasova, Brezina & McEnery 2017). The fixed maximum value is "14" and the closer the value is to "14", the stronger the association.



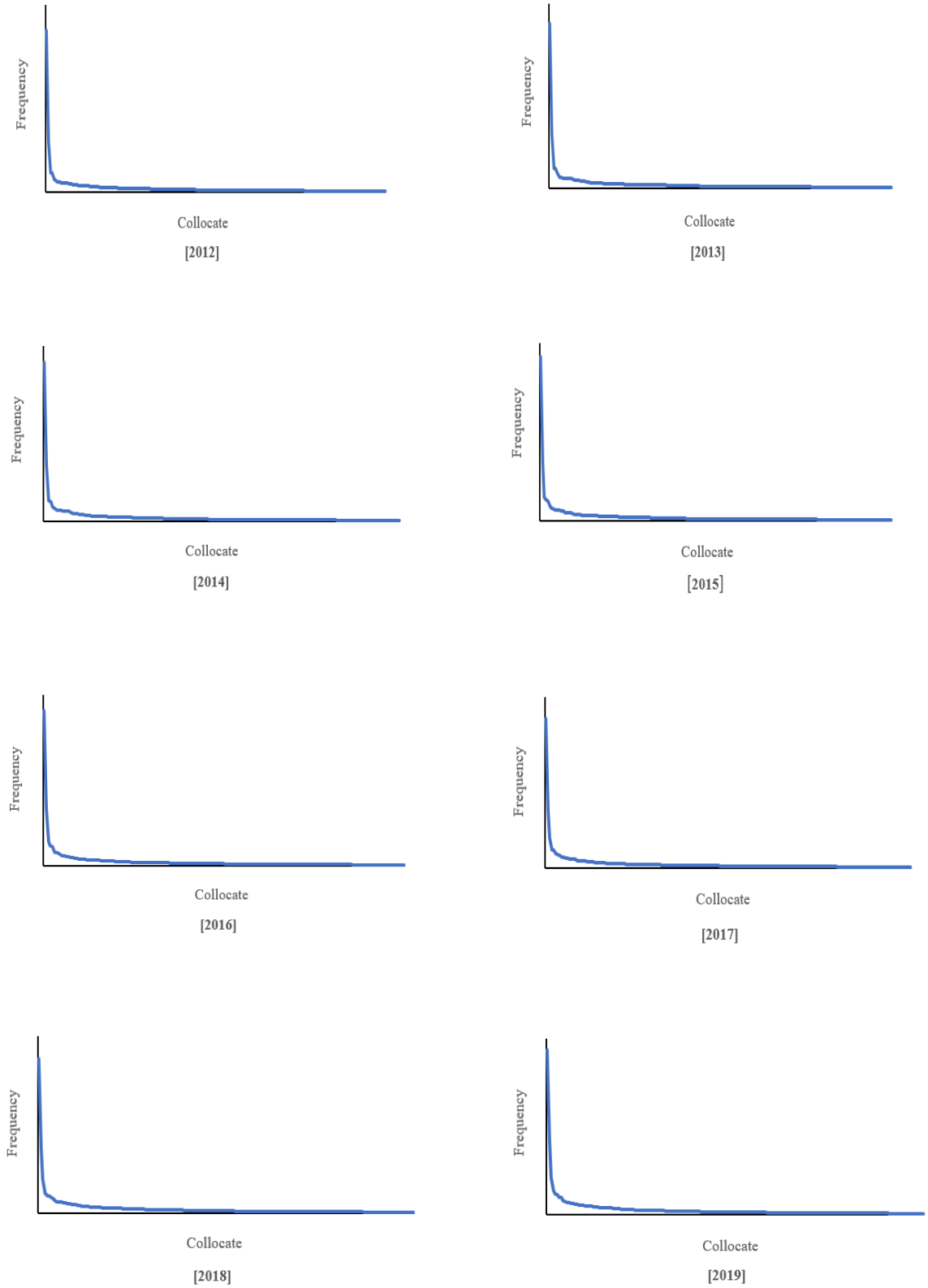


Figure 4.2 Distributional frequency profiles of collocates of *leyal* from each corpus

To begin with, I obtained collocates of *leyal* for each corpus and then arranged them according to their frequency (the number of collocates is about 200 for each corpus). Their frequency profiles showed an "A-curve" as expected (Figure 4.2). I extracted the top thirty collocates[13] from each corpus and compared the order of them among corpora (the box on the 2010 A-curve in Figure 4.2 shows where the top thirty collocates from the 2010 corpus are located and the location is true of the other A-curves). If there is a significant change in the order of the top thirty collocates across time, it means that the meaning of *leyal* and the way people use *leyal* have changed. To observe the change of the order based on the frequency more systematically, I investigated how the order of the top thirty collocates from the 2010 corpus varies across time. Table 4.3 indicates how the frequencies of the top thirty collocates from the 2010 corpus change depending on corpora (the collocates that do not have "English meaning" belong to either a case where there is no corresponding English word/morpheme or a case where it is hard to infer the meaning without contexts).

[13] The top collocates with high frequency are the ones that we hear and use frequently. Because they are essential collocates characterizing the ones of *leyal*, their change means a significant change to the way people use *leyal* and the meaning of *leyal*.

Table 4.3 Frequencies of the top thirty collocates of *leyal* from the 2010 corpus and their change across time

English meaning	2010 Collocate	Frequency	2011	2012	2013	2014	2015	2016	2017	2018	2019
Really	레알	176831	179274	172432	160215	163407	166130	162932	174787	174533	168342
To do	하다	60943	54655	54309	53356	58493	59492	60529	67041	73390	75741
To see/watch/look	보다	21376	21744	20452	19343	21277	21312	20623	21443	23072	24363
To exist/have	있다	16123	14744	13696	13937	15338	15134	14582	14765	16423	18098
Really	진짜	13555	19427	20757	19457	20703	23744	25920	34601	38240	37916
To become/be	되다	11880	10866	10300	9924	10705	10373	10155	10305	11224	12014
Not to be	아니다	11212	10836	10960	10531	11817	12861	14221	16192	19441	21023
All	다	10259	11116	11606	10887	11774	12179	13487	16485	17604	17532
To be good	좋다	10189	10124	9716	9398	10259	10974	10148	10222	10940	10613
Not to exist/have	없다	9732	9947	9647	9838	11221	10978	11053	11676	13105	13335
-	님	9458	7126	6860	7303	7547	8310	7908	8922	9426	9342
To be so	그렇다	9361	9076	8043	7720	8197	7994	7686	7477	8277	8512
Not	안	9341	9810	9566	9546	10785	11170	11487	12924	14057	14402
To go	가다	9338	9387	8094	7230	6770	6913	6953	7948	8156	8167
My	내	9186	10457	10792	9772	10394	10286	10833	11704	11595	11246
One	한	9099	8432	8042	7982	8717	8342	8461	8749	9739	10451
-	임	8671	7730	6674	5992	5546	5159	5934	8242	8839	8727
Today	오늘	8609	7708	6637	5264	5046	5177				
To come	오다	8112	7684	7072	7127	7788	8738	9174	12272	12484	12067
Thing	거	8036	7558	8407	9062	13332	19989	21650	21503	19977	20375
I	난	7343	6536	5370							
-	당	7073									
By the way/But	근데	7042	8765	8595	8156	7871	7648	7063	7496	8293	9185
To eat	먹다	6973	6405		4787	5351	5495			6471	7062

-	이야	6591	7145	6368	5059						
Really/Truly	정말	6582									
To break out	돋다	6370									
Cannot	못	6204	5996	5940	5701	6550	6868	7254	8077	8827	8710
-	들다	6199	5809	5281	5093	5265	5473	5819	6381	7310	7397
To sleep	자다	5925									

In Table 4.3, the second column is the top thirty collocates from the 2010 corpus and the third column shows their frequency. The first collocate is the node word *leyal* (its dominant meaning is given in the table). For visualization of the change, I used the lightness and darkness of a color. This visualization method comes from Kretzschmar (2021). Three colors were used. The darkest color is used for frequency ranging from 50,000 to under 200,000, the second darkest color for frequency ranging from 10,000 to under 40,000, and the next color for frequency under 10,000. The darker the color of a cell is, the more frequently the collocate is used. The white cells indicate a case where the collocate from the 2010 corpus is not found in the top thirty collocates from another corpus. For example, the top thirty collocates from the 2011 corpus include four different collocates from those of the 2010 corpus.

Table 4.3 shows that nineteen collocates among the top twenty collocates from the 2010 corpus are used with high frequency in the other corpora and the change of the order of the collocates is not significant. Because the change of the order occurs in the range of the same twenty collocates, it is considered as slight. To make sure that the change of the order is slight, I examined collocates in the top thirty from each corpus which are not in the top thirty from the 2010 corpus. About 78.5% of those collocates belong to the fifty most frequent collocates (i.e., the top fifty) from the 2010 corpus, which supports that the order of collocates has not changed by a wide margin. Overall, the dark cells remain dark and the light cells remain light throughout the ten years. This means that there is no great change in frequencies of the main context words of *leyal*. The results from the analysis of Table 4.3 imply that there has been no significant semantic change of *leyal* over the ten years as well as the way people use *leyal* has been consistent.

Next, in order to look into which meaning of *leyal* is more dominant between ‘really’ and ‘Real Madrid’, I compared the top thirty collocates of *leyal* from the 2010 corpus with those of

cincca and *leyalmatulitu* from the same corpus. Intriguingly, the top thirty collocates of *leyal* and *cincca* are the same, which means that the main context words of the two words are similar to each other. From the same collocates, it is inferred that their meanings are similar and the neologism *leyal* has been used to represent 'really'. Since it has been found that there has been no semantic change of *leyal* over the ten years, the result from the 2010 corpus is expected to be true of the other corpora. Therefore, it can be said that the neologism *leyal* has been mainly used to represent 'really' throughout the ten years.

In addition, Table 4.4 shows the top thirty collocates of *leyalmatulitu* and *leyal* from the 2010 corpus and their frequency (the collocates that do not have "English meaning" belong to either a case where there is no corresponding English word/morpheme or a case where it is hard to infer the meaning without contexts). The *leyalmatulitu* collocates in bold type are the ones that overlap with the *leyal* collocates. Although the overlapping collocates are used with high frequency, about two third of the *leyalmatulitu* collocates include different collocates from the *leyal* collocates. This demonstrates that the usage of *leyal* is different from that of *leyalmatulitu* and their meanings are quite different from each other.

Table 4.4 Top thirty collocates of *leyalmatulitu* and *leyal* from the 2010 corpus and their frequency

English meaning	**Collocates of *leyalmatulitu***	**Frequency**	**Collocates of *leyal***	**Frequency**
To do	하다	60943	레알	176831
To see/watch/look	보다	21376	하다	60943
To exist/have	있다	16123	보다	21376
To become/be	되다	11880	있다	16123
Not to be	아니다	11212	진짜	13555
To be good	좋다	10189	되다	11880
To go	가다	9338	아니다	11212
One	한	9099	다	10259
Today	오늘	8609	좋다	10189
To come	오다	8112	없다	9732

-	당	7073	님	9458
Word/End	말	5809	그렇다	9361
-	인	5728	안	9341
-	전	5062	가다	9338
Work/One/Day	일	4482	내	9186
Now	지금	4433	한	9099
Not to know	모르다	4006	임	8671
Among/During	중	3945	오늘	8609
-	분	3791	오다	8112
-	지	3357	거	8036
-	시	3335	난	7343
-	면	3299	당	7073
-	라	3077	근데	7042
To be right/be hit	맞다	2946	먹다	6973
-	수	2888	이야	6591
Real Madrid	레알마드리드	2426	정말	6582
To like	좋아하다	2223	돋다	6370
Soccer	축구	2132	못	6204
To kick/wear/be cold/be full	차다	2112	들다	6199
FC Barcelona	바르샤	1993	자다	5925

To sum up, the collocation analysis shows the same results as word2vec and cosine similarity: (i) the meaning of *leyal* subtly changes with time (the subtle change in the order of collocates from year to year indicates this); (ii) the meaning ‘really’ is more dominant than ‘Real Madrid’ and *leyal* has been mainly used to represent ‘really’; (iii) there is no semantic change in terms of the relation between the two meanings. The same results support that the DCA approach is suitable for lexical semantic change research.

4.6 Discussion

The analysis of the two meanings of the neologism *leyal* shows that there is no semantic change from the perspective of the relationship between the two meanings. The CS values between *leyal* and *cincca* and those between *leyal* and *leyalmatulitu* indicate that people have consistently used the neologism *leyal* to represent ‘really’ over the past ten years while the use as ‘Real Madrid’ has

been constantly inactive. There has been no great change in the way people use the neologism *leyal*, which means that the use of *leyal* as 'really' is perceived as normal. These results from word2vec and cosine similarity are shown in the collocation analysis as well.

The agreement of results between the DCA approach and collocation analysis gives us a question of why we should use word2vec despite the same result. Because each method has its own advantage, the best answer is that analysts should choose an appropriate method depending on the focus of their analysis. To mention the advantage of a collocation analysis, it can show analysts how the word under scrutiny is actually used. Table 4.3 in the preceding section shows what is happening with the collocates of *leyal*, i.e., the usage of *leyal*. Also, through the frequencies of collocates, we can infer whether the frequency of *leyal* has increased over time. The change of the frequency of a neologism is important information. As mentioned in Section 1.1, the frequency of a new word is low at first but it increases over time. With the high frequency, the new word is gradually used in different contexts, resulting in change in its meaning.

From a collocation analysis, we can see which collocates have come up to or dropped out of the top-ranked collocates. The specific information of change in the frequency and order of collocates allows analysts to observe the subtle semantic change of the target word. This is the biggest advantage of a collocation analysis, compared to the analysis employing the DCA approach. The analysis using word2vec and cosine similarity simply tells us whether semantic change has occurred and which is more dominant among the meanings of the target word. Therefore, for observation of the subtle semantic change and the specific usage of the target word, a collocation analysis is recommended.

However, there are no clear criteria defining what is a significant or slight change in the order of collocates. Also, it is hard to link the degree of change in the order of collocates to the

degree of change in meaning and usage clearly and objectively. If the quantification of the relation between the two is established, it will be able to improve a collocation analysis. Another disadvantage of a collocation analysis is that it relies on the use of corpus analysis software tools. Those tools cannot consider word meaning so they cannot indicate the semantic similarity between words. Corpus analysis tools use statistical association measures to display collocates. It is the analyst's job to connect the meaning of the target word with its collocates and analyze how the meaning of the target word has changed using the collocates. This manual analysis costs analysts more time and energy than the DCA approach.

In contrast, the biggest advantage of analysis employing word2vec is that it can consider word meaning. Word2vec produces vectors reflecting word meaning. On the basis of such vectors, cosine similarity quantitatively indicates the degree of semantic similarity between words. CS values make it easier and more convenient to access word meaning and do lexical semantic research. The CS values between the neologism *leyal* and the two alternative words have allowed me to track the semantic change of the neologism *leyal* more efficiently. This efficiency will serve as a considerable advantage when an analyst needs to analyze a huge amount of data. Therefore, for the analysis of large-scale data spanning a very long period of time, the use of word2vec and cosine similarity is recommended. However, because it is impossible to obtain specific and detailed information on the usage of the target word, additional analyses using other techniques or collocation analyses are required for the specific and detailed information.

In conclusion, the use of a collocation analysis for method validation is significant in that it is the first attempt applying a method in corpus linguistics to method validation of the DCA approach. In the fields of computational linguistics and computer science, word similarity judgments, intrinsic evaluations like measuring a trajectory's smoothness (Bamler & Mandt 2017),

attested shifts generated by historical linguists (Hamilton, Leskovec & Jurafsky 2016), and synthetic tasks (Kulkarni et al. 2015, Rosenfeld & Erk 2018) have been employed to evaluate the performance of distributional models such as word2vec. Methods related to corpus linguistics have never been tried so far. Therefore, it is worth doing per se to use a collocation analysis to assess the performance of word2vec and perform method validation of word2vec. This attempt will contribute to linking the DCA approach with corpus linguistics.

CHAPTER 5

DISTRIBUTIONAL CORPUS ANALYSIS OF KOREAN NEOLOGISM *LWUCE*

5.1 Data Collection

Like the neologism *leyal*, the neologism *lwuce* is also a non-standard word and mainly used in informal settings. Therefore, to collect data of *lwuce*, I scraped tweets including *lwuce* as a keyword by means of snscrape in Python. However, it was not possible to scrape tweets evenly for each month so I collected tweets unevenly per month from 2010 to 2019, within the range of a minimum number of 759 and a maximum number of 5,536. Also, I could not scrape the same number of tweets as for the neologism *leyal*. The maximum number of tweets that I could scrape for a year was 27,387. Table 5.1 shows the number of tweets scraped for each year. I used the scraped tweets to construct ten yearly Twitter corpora from 2010 to 2019.

Table 5.1 Number of tweets scraped for each year

Year	**Number of tweets**
2010	19,411
2011	27,387
2012	26,058
2013	22,113
2014	16,655
2015	25,803
2016	17,957
2017	21,447
2018	21,260
2019	19,639
Total	**217,730**

5.2 Data Preprocessing

For a clearer analysis, I decided to divide each of the ten yearly corpora in Section 5.1 into two subcorpora depending on the two meanings of *lwuce*, resulting in a total of twenty subcorpora (the ten yearly corpora are composed of tweets containing *lwuce* as a keyword). Moreover, in order to gain better results from Latent Dirichlet Allocation (LDA), I decided to extract nouns and numbers from the twenty subcorpora and implement LDA only with those nouns and numbers (i.e., I removed all the other parts of speech). Implementing LDA only with nouns is one of the ways to improve the performance of LDA (https://towardsdatascience.com/6-tips-to-optimize-an-nlp-topic-model-for-interpretability-20742f3047e2). The reason why I decided to extract numbers as well is that I wanted to include "180" in the analysis. To be specific, each of the ten yearly Twitter corpora went through the following six steps for data preprocessing:

- **Step 1**: The same tweets scraped several times were reduced to one tweet.
- **Step 2**: Unnecessary things such as numbers, punctuation marks, and other foreign languages were removed, with only the Korean alphabet left. However, in tweets containing "180", numbers were not removed so that "180" could remain. This is because "180" is a very important number for the new meaning of the neologism *lwuce* (as the overall frequencies of the other numbers in the tweets containing "180" were low, their presence was not considered to affect results so they were not removed).
- **Step 3**: Duplicate tweets were once again reduced to one tweet in order to filter retweets.

- **Step 4**: The tweets were divided into two groups: the existing meaning subcorpus and the new meaning subcorpus (if a tweet contains any of the following seven morphemes[14], it was classified as the new meaning subcorpus: *180*, *khi*, *iha*, *kkalchang*, *seynthi*, *seynchi*, and *sseynthi*).
- **Step 5**: Each sentence in tweets in the two subcorpora was split into morphemes. The morphemes were converted to their stems and tagged with their parts of speech.
- **Step 6**: Based on the tag information, only numbers and nouns which are not included in stop words were extracted from the two subcorpora.

I performed all the above six steps in Python version 3.7. For the fifth step, I employed the morpheme analyzer tool Okt in KoNLPy. Regarding stop words in the sixth step, I used the same list of stop words as used in the data preprocessing of *leyal*. Table 5.2 indicates the number of morphemes in each of the ten yearly corpora after preprocessing.

Table 5.2 Number of morphemes in each of the ten yearly corpora after preprocessing

Year	Number of morphemes
2010	159,687
2011	202,095
2012	182,721
2013	135,418
2014	94,958
2015	163,649
2016	103,154
2017	125,708
2018	130,409
2019	128,720

[14] The morpheme *khi* means ‘stature/height’, *iha* ‘below’, and *kkalchang* ‘shoe insole’. The latter three morphemes are variants of *seynthimithe*. All of them mean ‘centimeter’.

5.3 Methodology

To examine how the new and existing meanings of the neologism *lwuce* have developed, I employed LDA. The choice of this technique was because there are no other words representing the new meaning of *lwuce* in the Korean language. As it was impossible to compare vectors between similar words by means of word2vec, I decided to investigate change in the context words of *lwuce* directly. To look into the context words more efficiently, I chose LDA. However, since LDA classifies lexical items without considering their meanings, I focused on the number of topics, rather than inferring topics from lexical items in clusters.

As people use a specific word frequently, the number of tweets including that word as a keyword will gradually increase. Compared to a smaller number of tweets, the increased number of tweets contains more contexts of the keyword, leading the number of topics of contexts to increase. Therefore, if the number of topics from the new meaning subcorpus increases over time, we can say that the use of the new meaning has increased. On the contrary, if the number of topics decreases across time, it means that the use of the new meaning has decreased. The change in the use of each meaning will be able to show which meaning is more dominant and how each meaning has developed.

In order to find the number of topics from each of the preprocessed subcorpora (i.e., the ten new meaning subcorpora and the ten existing meaning subcorpora), I used C_v topic coherence. A topic coherence measure is used to check whether LDA has properly classified lexical items. It indicates how consistent words in a cluster are in terms of meaning by measuring the semantic similarity of words. C_v topic coherence is one of the most popular topic coherence measures and it ranges between 0 and 1. The value ".55" is considered to be satisfactory and the value between ".65" and ".8" are regarded as good. The values ".9" and "1" are rarely observed. When words are

compound words such as *United States* and *United Kingdom* or words are identical, the two rare values can be observed.

I investigated the number of topics satisfying "beyond the minimum topic coherence value .55" for each of the twenty subcorpora. The numbers were heuristically selected. I set fifty as the limit of trials for each subcorpus and selected the first number meeting the requirement on topic coherence within the fifty trials. Table 5.3 indicates the topic coherence value for each of the twenty subcorpora (the values have been rounded to the nearest hundredth). The topic coherence value from the 2013 new meaning subcorpus is not beyond .53. Despite fifty trials, I could not obtain the number of topics producing a value beyond .55. None of the values from the new meaning subcorpora are at or above .65 (good level) and only in three cases out of the ten existing meaning subcorpora we can see such values. In order to improve the reliability of LDA results, it is required to analyze the numbers of topics from values above .65.

Table 5.3 Topic coherence value for each of the twenty subcorpora

Year	New meaning subcorpus	Existing meaning subcorpus
2010	.56	.62
2011	.61	.60
2012	.63	.63
2013	.53	.63
2014	.60	.67
2015	.59	.62
2016	.59	.66
2017	.60	.65
2018	.56	.58
2019	.57	.64

5.4 Results

Table 5.4 indicates the numbers of topics from the two subcorpora for each year and their relative proportions ("ES" means the existing meaning subcorpus, "NS" means the new meaning subcorpus, and the proportion values have been rounded to the nearest tenth). Figure 5.1 shows how the proportions have changed over time. The proportion from the existing meaning subcorpus is always higher than that from the new meaning subcorpus. This implies that the existing meaning of *lwuce* is more dominant than the new meaning throughout the ten years. That is, there is no semantic change in terms of the dominance relation between the two meanings. Also, the proportion from the new meaning subcorpus gradually decreased from 2010 to 2017 (although it slightly increased in 2014) but increased in 2018 by a wide margin and stayed the same in 2019. This means that the use of the new meaning gradually decreased but increased in 2018. The proportion values of the new and existing meanings in 2015 show that the most significant change in the use of the two meanings occurred in 2015. The difference of proportions between 2014 and 2015 is greatest. The second most significant change occurred in 2018. In contrast to the previous years showing a gradual decrease in the use of the new meaning, the use of the new meaning increased and that of the existing meaning decreased in 2018.

Table 5.4 Numbers of topics from the two subcorpora for each year and their proportions

Year	Number of topics from ES (NES)	Number of topics from NS (NNS)	Proportion of NES (%)	Proportion of NNS (%)
2010	50	30	62.5	37.5
2011	55	29	65.5	34.5
2012	52	27	65.8	34.2
2013	49	20	71.0	29.0
2014	45	19	70.3	29.7
2015	51	16	76.1	23.9
2016	46	14	76.7	23.3
2017	48	14	77.4	22.6
2018	47	18	72.3	27.7
2019	47	18	72.3	27.7

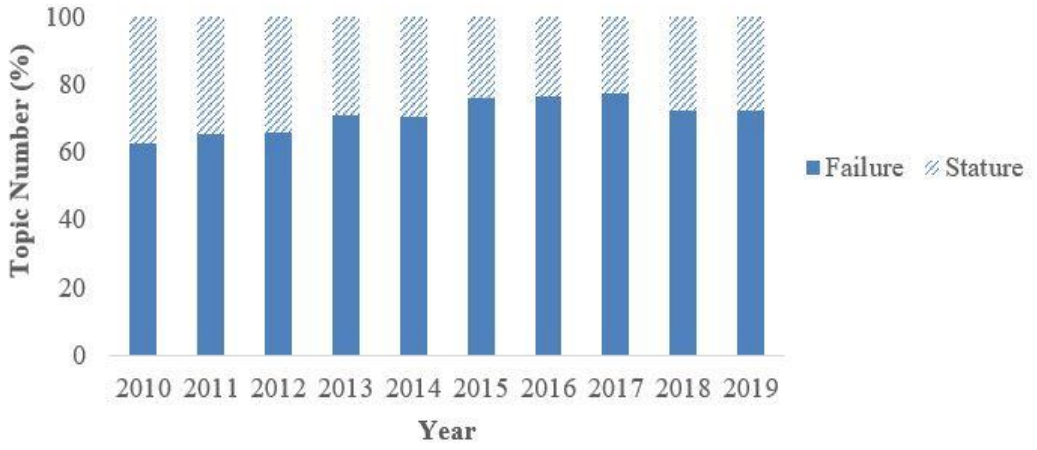
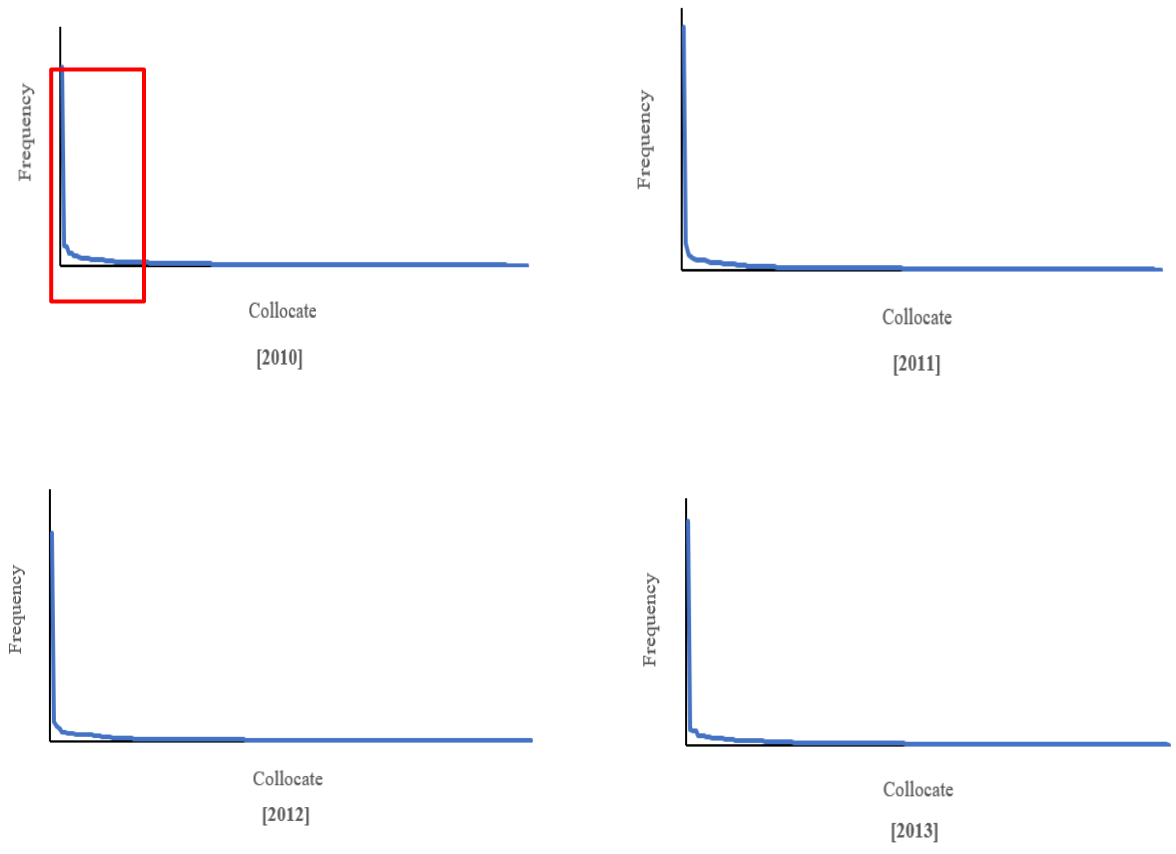


Figure 5.1 Change in the proportions of topic numbers for each subcorpus

5.5 Method Validation

In order to assess whether the results from LDA are reliable, I performed a collocation analysis. To begin with, I obtained the collocates of *lwuce* from each of the ten yearly corpora only consisting of nouns and numbers (i.e., two subcorpora combined for each year). The collocates of *lwuce* were obtained in the same way as *leyal*. I used LancsBox, considered three words to the left and right of *lwuce* with a minimum frequency of "5", and set the Log Dice threshold value to "7.0".

After obtaining collocates for each corpus, I arranged them according to their frequencies and checked whether their frequency profiles show A-curves. Figure 5.2 indicates the distributional frequency profiles of collocates from each corpus (the number of collocates is about 200 per corpus). I selected top thirty collocates from each corpus (the box on the 2010 A-curve in Figure 5.2 shows where the top thirty collocates are located and the location is true of the other A-curves) and observed how the orders of number and two words essential to the definition of the new meaning of *lwuce* have changed across the ten yearly corpora. The number is *180* and the two words are *namca* and *khi*. *Namca* means 'man' and *khi* represents 'stature/height'.



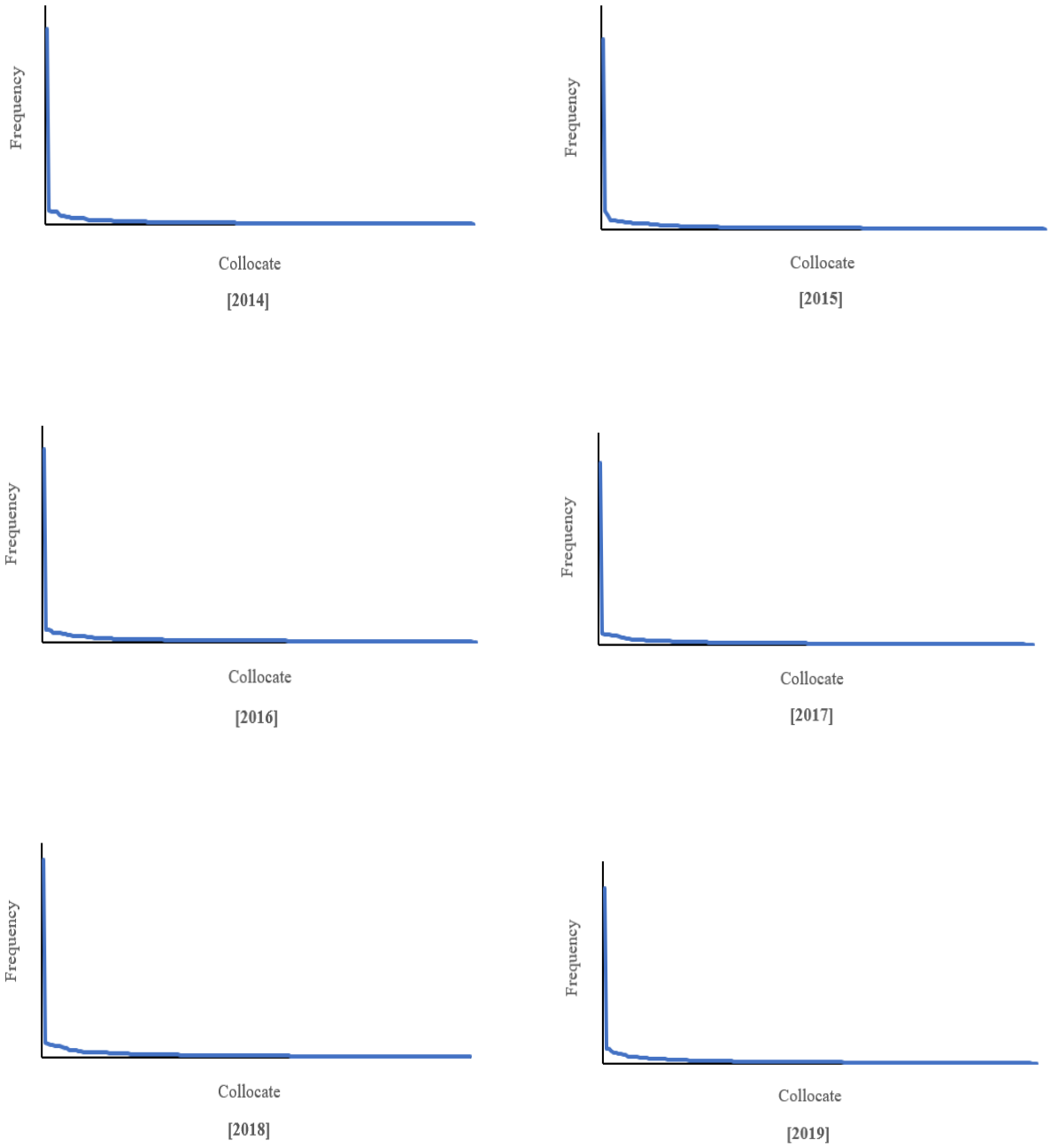


Figure 5.2 Distributional frequency profiles of collocates of *lwuce* from each corpus

Table 5.5 Top thirty collocates from each corpus

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
루저	루저	루저	루저	루저	루저	루저	루저	루저	루저
난	난	난	난	내	빅뱅	외톨이	내	내	내
키	키	키	키	키	외톨이	내	진짜	진짜	진짜
사람	내	내	내	난	시티	말	난	못	사람
전	사람	말	말	감	진짜	사람	외톨이	사람	거
위너	위너	사람	감	말	말	난	말	외톨이	말
말	전	위너	사람	사람	내	생각	못	말	못
내	말	오늘	위너	오늘	난	진짜	사람	거	새끼
생각	오늘	못	오늘	생각	베베	거	거	난	생각
남자	티켓팅	전	뭐	진짜	거	키	생각	새끼	난
오늘	못	남자	못	거	사람	남자	오늘	생각	뭐
녀	진짜	티켓팅	진짜	못	생각	못	더	인생	인생
임	생각	뭐	더	더	노래	여자	새끼	안	더
여자	뭐	더	일	남자	키	더	감	오늘	오늘
뭐	남자	진짜	티켓팅	뭐	더	새끼	인생	더	안
더	거	생각	생각	위너	오늘	뭐	뭐	남자	남자
180	더	여자	거	180	척	안	남자	뭐	남
거	임	임	전	여자	겁쟁이	오늘	안	여자	그냥
그냥	여자	거	남자	안	뭐	빅뱅	키	그냥	감
일	지금	180	분	새끼	위	180	여자	남	키
세상	180	그냥	안	티켓팅	감	인생	애	존나	애
수	그냥	지금	임	전	남자	척	수	개	여자
못	일	분	180	그냥	못	그냥	전	척	외톨이
지금	수	뀀	콘	막	180	감	개	키	임
발언	인생	일	그냥	날	안	전	척	애	개
진짜	날	안	위	일	여자	위너	일	수	보고
정말	안	보고	콘서트	수	그냥	수	위너	보고	존나
안	보고	수	벵	분	개	지금	지금	위너	게
아이폰	기분	기분	수	보고	지금	겁쟁이	그냥	지금	클럽
분	상루	인생	지금	인생	전	애	존나	겁쟁이	일

Table 5.5 is top thirty collocates from each corpus. It shows how the orders of the three collocates *khi*, *namca*, and *180* change over time (because the focus of this table is to indicate

change in the order of the three collocates, the English meaning of each collocate is skipped). The darkest color is used for *khi*, the second darkest color for *namca*, and the third darkest color for *180*. The collocate *khi* is consistently ranked third for four years from 2010 to 2014. Since 2015, its order fluctuates out of the tenth rank. The other two collocates *namca* and *180* also fluctuate out of the tenth rank. The collocate *namca* is contained in the twenty-fifth rank every year while *180* is not shown in the thirtieth rank for three years from 2017 to 2019. The change in the order of the three collocates shows that the use of the new meaning has significantly decreased since 2015.

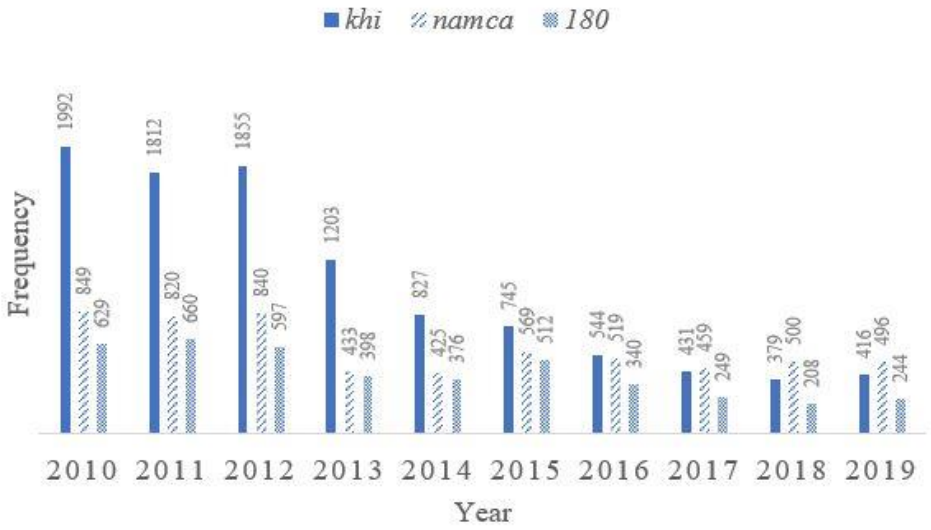


Figure 5.3 Change in the frequencies of *khi*, *namca*, and *180* across time

Figure 5.3 shows how the frequencies of the three collocates *khi*, *namca*, and *180* change across time (for the frequency of *180* from 2017 to 2019, I looked into collocates beyond the thirtieth rank). The frequency of *khi* continues to decrease except for the period from 2011 to 2012. The frequencies of *namca* and *180* repeatedly rise and fall by a small margin but compared to 2010,

their frequencies decreased in 2019. Among the three collocates, the frequency of *khi* decreased most significantly. The overall decrease in the frequencies of the three collocates supports that the use of the new meaning decreased.

More specifically, I looked into the top eleven collocates for each corpus (the first collocate is the node word so I examined eleven collocates). To find out to which meaning each collocate is connected[15], I used collocates of *lwuce* from the subcorpora. To take an example of the 2010 corpus, I obtained top eleven collocates from each of the three corpora (i.e., the 2010 corpus, the 2010 new meaning subcorpus, and the 2010 existing meaning subcorpus). The top eleven collocates from each corpus have been selected from the list of collocates arranged according to their frequencies. Next, I classified the eleven collocates from the 2010 corpus into three groups using different colors. The darkest color was used for overlapped collocates between the two subcorpora, the second darkest color for collocates from the existing meaning subcorpus, and the third darkest color for collocates from the new meaning subcorpus. This process was repeated for each year.

[15] There are too many words related to the existing meaning as the existing meaning is used in so many different contexts. People can use *lwuce* when they do not have money, do not have a good job, lose a game, cannot get a concert ticket, and so on. The existing meaning is used with various topics. Therefore, the existing meaning is not limited to particular words. In a collocation analysis, it is not easy to find which collocates are associated with the existing meaning. Also, it is not easy to find which collocates are related to the new meaning among other collocates except for "man", "stature", and "180".

Table 5.6 Eleven collocates classified into three groups for each year

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
루저 (Loser) 20792	루저 (Loser) 27824	루저 (Loser) 24781	루저 (Loser) 18129	루저 (Loser) 12519	루저 (Loser) 21228	루저 (Loser) 12487	루저 (Loser) 15452	루저 (Loser) 14718	루저 (Loser) 13872
난 (I) 2062	난 (I) 3097	난 (I) 2384	난 (I) 1317	네 (My) 863	빅뱅 (Bigbang) 2091	외톨이 (Loner/Outsider) 848	네 (My) 1017	네 (My) 1123	네 (My) 1172
키 (Stature) 1992	키 (Stature) 1812	키 (Stature) 1855	키 (Stature) 1203	키 (Stature) 827	외톨이 (Loner/Outsider) 1593	네 (My) 805	진짜 (Really) 904	진짜 (Really) 1057	진짜 (Really) 1147
사람 (Person) 1331	네 (My) 1513	네 (My) 1564	네 (My) 1196	난 (I) 819	서터 (City) 1018	말 (Word) 725	난 (I) 888	못 (Not) 964	사람 (Person) 1006
전 (I) 1305	사람 (Person) 1312	말 (Word) 1100	말 (Word) 768	감 (-) 799	진짜 (Really) 1009	사람 (Person) 637	외톨이 (Loner/Outsider) 873	사람 (Person) 911	거 (Thing) 900
윈너 (Winner) 1129	윈너 (Winner) 1214	사람 (Person) 1060	감 (-) 766	말 (Word) 606	말 (Word) 997	난 (I) 635	말 (Word) 831	외톨이 (Loner/Outsider) 902	말 (Word) 894
말 (Word) 1077	전 (I) 1190	윈너 (Winner) 1034	사람 (Person) 765	사람 (Person) 554	내 (My) 977	생각 (Thought) 599	못 (Not) 825	말 (Word) 882	못 (Not) 775
내 (My) 897	말 (Word) 1153	오늘 (Today) 956	윈너 (Winner) 733	오늘 (Today) 532	난 (I) 914	진짜 (Really) 596	사람 (Person) 780	거 (Thing) 874	새끼 (Bastard) 760
생각 (Thought) 849	오늘 (Today) 1129	못 (Not) 912	오늘 (Today) 668	생각 (Thought) 472	빼빼 (Bae bae) 893	거 (Thing) 574	거 (Thing) 747	난 (I) 784	생각 (Thought) 721
남자 (Man) 849	티켓팅 (Ticketing) 991	전 (I) 869	뭐 (What) 611	진짜 (Really) 468	거 (Thing) 830	키 (Stature) 544	생각 (Thought) 616	새끼 (Bastard) 726	난 (I) 706
오늘 (Today) 781	못 (Not) 960	남자 (Man) 840	못 (Not) 604	거 (Thing) 446	사람 (Person) 786	남자 (Man) 519	오늘 (Today) 611	생각 (Thought) 690	뭐 (What) 573

Table 5.6 shows eleven collocates classified into three groups for each year (the meanings of collocates which have homonyms were inferred from random selection of some tweets and the collocate which does not have its English meaning is a case where it is hard to infer the Korean meaning without a number of contexts containing the collocate). The number below each collocate

means its frequency. The existing meaning collocates are always more than the new meaning collocates. This demonstrates that the existing meaning is more dominant than the new meaning throughout the ten years. For five years from 2010 to 2014, the collocate *khi* representing 'stature/height' is ranked third. However, in 2015, the number of the new meaning collocates is one and it is ranked eleventh. In 2016, the number of the new meaning collocates is two and they are ranked tenth and eleventh respectively. For the other three years from 2017 to 2019, there is no new meaning collocate in the eleven collocates. The change in the order and number of the new meaning collocates shows that the use of the new meaning has decreased over time.

Concerning the existing meaning collocates, the number of those collocates had gradually increased for six years from 2010 to 2015. In 2016, the number of the collocates decreased from seven to five. The number remained the same for two years from 2017 to 2018. In 2019, the number decreased from five to four. While the LDA analysis shows that the use of the existing meaning decreased in 2018, the collocation analysis shows that it decreased starting in 2016. However, in 2015, the number of the existing meaning collocates most greatly increased. This fact and the demoted rank of the new meaning collocate *khi* in 2015 demonstrate that the most significant change in the use of *lwuce* occurred in 2015, which agrees with the result from LDA.

To sum up, the collocation analysis shows that the first two results among the following four ones from LDA are true and the fourth one is partially true: (i) the existing meaning is more dominant than the new meaning throughout the ten years; (ii) in 2015, the use of the new meaning most significantly decreased and that of the existing meaning most significantly increased; (iii) the use of the new meaning increased and that of the existing meaning decreased in 2018; (iv) the use of the new meaning gradually decreased but increased in 2018. Therefore, not every result from LDA is reliable. The employment of LDA for lexical semantic research is not yet recommended

but the comparison of results between the collocation analysis and the LDA analysis demonstrates that LDA has enough potential to be used for such research.

5.6 Discussion

Compared to the number of tweets collected for the analysis of *leyal*, that of tweets for *lwuce* is too small. For *lwuce*, the maximum number of tweets for a year is 27,387 but for *leyal*, 186,000 tweets were collected per year. The small amount of data makes it difficult to generalize results. Therefore, for generalization of the results from LDA and the collocation analysis, a larger number of tweets should be collected.

With regard to the LDA results, they show that there is no change in terms of the relationship between the new and existing meanings of *lwuce*. The existing meaning has always been more frequently used than the new meaning. The investigation of the new meaning shows that the use of the new meaning continued to decrease until 2017. The decrease implies that the use of *lwuce* as "a man's stature" was active when the neologism was created (it was created in late 2009) and was a hot issue. The reason for the decrease might be public antipathy to the combination between the new meaning and the existing meaning.

Some of the LDA results were verified by the collocation analysis used for method validation of LDA but others were not. The disagreement of some results between LDA and the collocation analysis shows that LDA is not an adequate method to study word meaning. LDA still needs improvement for better results but this study has confirmed that it has enough potential to be employed for lexical semantic research.

An advantage of the LDA analysis is quantification. I used the number of topics to track how the two meanings of *lwuce* have developed. The change in the number of topics for each meaning represents change in the use of each meaning. The more the number of topics is, the more

the use. Because quantification allows us to see how much change occurred by means of numbers, we can observe language change more clearly.

However, LDA has a limitation. The limitation is related to the reason why I did not analyze lexical items in clusters. It is not easy to infer what topics are from lexical items in clusters. This is because not every lexical item in a cluster is semantically relevant to the others. Since LDA classifies lexical items based not on their meanings but statistics, the classification results can be different from classification by humans.

One way to overcome this limitation is to employ topic models which can classify documents considering word meaning. An example of those topic models is Gaussian LDA. Gaussian LDA uses word2vec to group semantically similar words into topics. Gaussian LDA employs vectors with word meanings from word2vec to classify words into topics. Das, Zaheer & Dyer (2015) demonstrate that the topic coherences of results from Gaussian LDA are higher than those from LDA.

Also, it would be helpful to use formal language which consists of grammatical sentences as input data. I found that implementing LDA with nouns and numbers still leaves semantically irrelevant lexical items in clusters. Furthermore, I found that it is impossible to remove all the irrelevant lexical items from clusters as long as I use Twitter data. Twitter data contains many ungrammatical expressions and sentences. In addition, as people violate rules for word spacing, it is hard to parse sentences in tweets. Okt in KoNLPy cannot correctly split ungrammatical sentences and slang into morphemes and tag them with their parts of speech. Those incorrect morphemes affect classification by LDA. Therefore, the use of formal language such as news articles and editorials is recommended.

Compared to LDA, an advantage of a collocation analysis is that it shows the actual usage of *lwuce*. Through change in the frequencies of its collocates related to the new meaning, we can infer whether the use of the new meaning has increased or decreased. Moreover, we can see what words *lwuce* is mainly used with. Also, we can see how those words change over time. Which collocates have come up to or dropped out of the top-ranked collocates shows which meaning of *lwuce* is more dominant and how each meaning has developed. The change in the frequency and order of collocates allows analysts to figure out the semantic change of the target word obviously.

However, a collocation analysis does not tell us the degree to which each meaning of *lwuce* changes every year exactly. For example, the three collocates related to the new meaning (*180*, *namca*, and *khi*) behave differently from each other in terms of the order. Since we need the information of overall change combining the three collocates, the difference gives no help. In other words, the bigger the difference is, the harder it is to figure out the overall trend. If the order of a certain collocate decreases but that of another collocate increases, it is not easy to make a conclusion. In light of this, quantification is required and the DCA approach will be able to complement a collocation analysis.

As mentioned in Section 4.6, the use of a collocation analysis for method validation of the DCA approach is meaningful because methods related to corpus linguistics have never been tried so far. Because a collocation analysis is based on the intuitive and transparent analysis of linguistic data, its results are clear and obvious. Therefore, a collocation analysis will be able to serve as a suitable method for the validation of the DCA approach, especially models trained in unsupervised learning in artificial intelligence.

CHAPTER 6

DISTRIBUTIONAL CORPUS ANALYSIS OF KOREAN NEOLOGISM *KAY-*

6.1 Data Collection

To collect the data of the neologism *kay-*, I employed Korean Twitter because the neologism *kay-* is slang used in informal settings. As there are many homonyms of the neologism *kay-* in the Korean language (e.g., *kay* representing 'piece', *kay* representing 'dog', *kay* representing 'lid', and so on), it was not easy to scrape tweets only containing the neologism *kay-*. Therefore, I decided to scrape tweets mixed with the neologism and those homonyms by means of snscrape in Python and extract words needed for the analysis of the neologism in the process of data preprocessing. To be specific, I collected about 35,000 tweets including *kay* as a keyword per month from 2010 to 2019 (except for the months of January and February in 2010 and the month of April in 2011). Concerning the three months that are exceptions, I scraped 19,872 tweets, 30,271 tweets, and 6,045 tweets respectively because I could not scrape the amount of data that I aimed for despite several trials. Using the tweets scraped per year, I constructed ten yearly Twitter corpora which correspond to each of the ten years from 2010 to 2019.

6.2 Data Preprocessing

The process of data preprocessing was applied to the Twitter dataset (i.e., the ten Twitter corpora) and the KNU Korean Sentiment Lexicon dataset. For the Twitter dataset, adjectives and verbs where the neologism *kay-* is attached were extracted from the Twitter corpora through

preprocessing. For the KNU Korean Sentiment Lexicon dataset, training and validation datasets to the model of long short-term memory were made through preprocessing.

6.2.1 Twitter Dataset

Each of the ten Twitter corpora went through the process of preprocessing consisting of the following six steps:

- **Step 1**: Duplicate tweets were reduced to one tweet to deal with cases where the same tweet was repeatedly scraped.
- **Step 2**: Unnecessary things for this study (e.g., numbers, punctuation marks, other foreign languages, and so on) were removed from tweets, leaving only the Korean alphabet in each sentence.
- **Step 3**: Duplicate tweets were reduced to one tweet once again to deal with retweets containing the same contents under different user IDs.
- **Step 4**: Every sentence in each tweet was divided into eojuls, which are separated by spacing.
- **Step 5**: Eojuls where *kay* appears were only extracted from tweets.
- **Step 6**: Given that the neologism *kay-* is attached to adjectives or verbs in contrast with the existing prefix *kay-* attached to nouns, adjectives and verbs immediately following *kay-* were only extracted from the eojuls after they were converted into their stems.

Table 6.1 An example for each step in the preprocessing of Twitter dataset except for step 3

Step	Example
1	*@sweetie_doll 장비 사다보면 귀족에서 **개**망할것 같아요*
2	*장비 사다보면 귀족에서 **개**망할것 같아요*
4	*장비, 사다보면, 귀족에서, **개**망할것, 같아요*
5	***개**망할것*
6	*망하다*

Table 6.2 Number of morphemes for each Twitter corpus in step 3

Year	Number of morphemes
2010	8,111,873
2011	6,723,103
2012	6,507,313
2013	5,605,623
2014	4,845,912
2015	5,601,009
2016	6,322,393
2017	7,113,176
2018	7,438,726
2019	7,364,402

I carried out all the above steps in Python version 3.7 and 3.8. For the sixth step, I employed KoNLPy, specifically Okt, to split the entire text into morphemes and annotate the morphemes with their parts of speech. Table 6.1 shows an example for each step excluding the third one (in the examples, the syllables represented in bold type are *kay*-) and Table 6.2 indicates the number of morphemes in each of the ten Twitter corpora in the third step. The corpora in the third step are final Twitter corpora which are only composed of the Korean alphabet. The parts of speech of the homonyms of the neologism *kay*- are mostly nouns which can stand alone. In other words, those homonyms are not attached to the following adjectives or verbs, leaving spacing between themselves and the following adjectives or verbs. Therefore, it was assumed that the extraction of adjectives and verbs from eojuls in which there is no spacing between *kay*- and the immediately

following adjective or verb guarantees that *kay-* before the extracted adjectives and verbs is the neologism *kay-*.

6.2.2 KNU-KSL Dataset

To train the model of long short-term memory (LSTM) to classify words as positive or negative, training and validation datasets for the model are required. For the training and validation datasets, the KNU Korean Sentiment Lexicon (KNU-KSL) was used. KNU-KSL is a Korean sentiment dictionary which includes a list of domain-independent sentiment words. It was constructed by Park et al. (2018). They used their deep learning model based on bidirectional long short-term memory to classify glosses contained in the Standard Korean Language Dictionary (SKLD) as either positive or negative meaning, and then extracted positive words and phrases from the glosses categorized as positive meaning and negative words and phrases from the glosses categorized as negative meaning. KNU-KSL consists of 9,795 negative words and phrases, 151 neutral words and phrases, and 4,839 positive words and phrases. Every word and phrase in KNU-KSL is on a five-level scale: -2 (very negative), -1 (negative), 0 (neutral), 1 (positive), and 2 (very positive). KNU-KSL contains sentiment information about emoticons and coined words frequently used on the Web as well as tells what words and phrases are positive or negative. For the creation of training and validation datasets to the LSTM model, the dataset of KNU-KSL went through the following steps:

- **Step 1**: The five-level scale assigned to words and phrases was reduced to a binary scale composed of positive and negative, with the neutral words and phrases removed. Negative and very negative words and phrases were labeled with “0” and positive and very positive words and phrases with “1”.

- **Step 2**: Every word and phrase in KNU-KSL was split into morphemes and the morphemes were reduced to their stems or root forms.
- **Step 3**: Those stems or root forms were converted into the sequences of integers based on frequency.
- **Step 4**: Because the lengths of the sequences of integers were different depending on phrases, the sequences were all adjusted to the same length (i.e., four in this study). If the length of a sequence is below the adjusted length, "0" was added to the sequence through padding[16].

I performed all of the above processes in Python version 3.7 and 3.8. For the second step, I used Okt in KoNLPy and for the third step, I employed a Python library for artificial neural networks called Keras. Table 6.3 shows an example for each step (the used example phrase means 'suffering from poverty'). The preprocessed KNU-KSL dataset was partitioned into training and validation datasets. Partitioning the dataset was computationally carried out.

Table 6.3 An example for each step in the preprocessing of KNU-KSL dataset

Step	Example
1	가난에 쪼들려서 -2 → 0
2	가난, 에, 쪼들리다 → 0
3	[50, 454, 16] → 0
4	[50, 454, 16, 0] → 0

[16] Padding in RNN refers to making all the sequences fit a given standard length by padding or truncating some sequences.

6.3 Methodology

6.3.1 Training LSTM

LSTM trained in a supervised fashion needs training and validation datasets. In this study, those datasets have the form of pairs which consist of a word and a class label ("0" representing negative or "1" representing positive). The training dataset is for training the model while the validation dataset is for improving the model and checking if the model has been well trained. Simply put, the validation dataset corresponds to a practice test. The well-trained model through training and validation is used to classify the actual data, i.e., the adjectives and verbs extracted from the Twitter corpora, into positive and negative.

One of the most important things in making the validation dataset is to make sure that the validation dataset is composed of different words from the training dataset. This is because it is needed to check if the model excessively corresponds to the training dataset. If the model corresponds too exactly to the training dataset (this is called "overfitting" in machine learning), it cannot correctly sort unseen words in the validation dataset and further it fails to classify unseen words in the actual dataset accurately. Therefore, it is critical to contain unseen words in the validation dataset to check how the model responds to unseen words.

Moreover, class imbalance should be considered. As the ratio of negative words and phrases to positive ones in the KNU-KSL dataset is 2:1, the KNU-KSL dataset is imbalanced data. If the ratio is not considered in partitioning the KNU-KSL dataset into training and validation datasets, in an extreme case, only negative words and phrases might be assigned to the training dataset and only positive words and phrases might be assigned to the validation dataset. In this case, the LSTM model cannot make a correct prediction of the validation dataset because the model

has never been trained with positive words and phrases. To avoid this ratio issue and the overfitting issue mentioned above, I applied stratified k-fold cross-validation.

To be specific, I set the value of k to “5”, leading the preprocessed KNU-KSL dataset to be randomly partitioned into five subsamples that have the same size. As each of the five subsamples takes turns to function as the validation dataset, the other four subsamples function as the training datasets. For each validation dataset, its training dataset gets different (Figure 6.1). Because stratified k-fold cross-validation enables every subsample to be used in training and validation, it prevents the model from excessively corresponding to a specific training dataset. Also, it allows the ratio in every subsample to be the same as the ratio in the original KNU-KSL dataset.

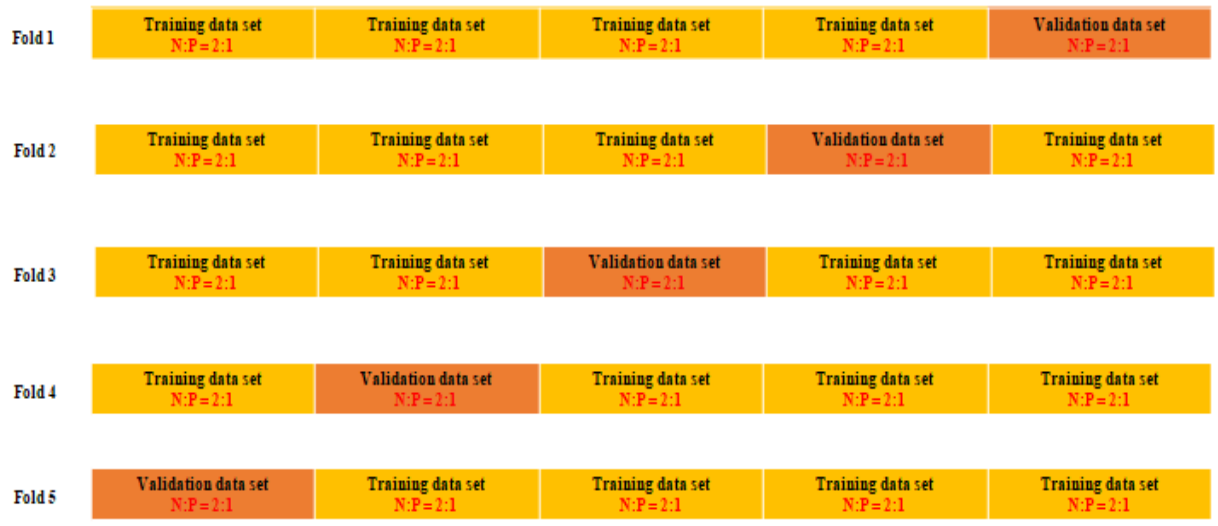


Figure 6.1 Stratified 5-fold cross-validation applied to KNU-KSL. In each fold, the preprocessed KNU-KSL dataset is partitioned into training and validation datasets. “N:P = 2:1” in the figure means the ratio of negative words and phrases to positive ones is 2:1.

For each fold, the number of words and phrases in the training dataset was 11,708. It takes the model a long time to process all of them at once so they were divided into a hundred and eighty-

three groups with one group composed of sixty-four words and phrases (batch size = 64). A total of a hundred and eighty-three iterations were required for all the groups to be passed through (iterations = 183). The LSTM model was updated every iteration, making the model better fit the training dataset.

Moreover, for each fold, early stopping was applied. Early stopping refers to one of the constraints imposed on a model to prevent it from overfitting. When the performance of a model does not improve any more during the process of training, the model stops before it begins to overfit. Early stopping plays an important role in determining the number of times a model has to be trained for optimal performance.

6.4 Results

To find out how the semantic prosody of the neologism *kay-* has changed over time, I counted the token frequencies of positive and negative adjectives and verbs where the neologism *kay-* is attached for each year. Specifically, the trained LSTM model was used to sort the adjectives and verbs extracted from each Twitter corpus into positive "1" and negative "0". Based on such classification, I computed the sum of the token frequencies of every positive word and that of the token frequencies of every negative word for each year. Because the number of tweets collected per year was different from each other, I compared the relative proportions of the total token frequencies of positive and negative words. Table 6.4 shows the proportions of positive and negative words on the basis of their total token frequencies for each year. Figure 6.2 indicates that the proportion which positive words make up has gradually increased over time. This finding demonstrates that the number of positive words where the neologism *kay-* is attached has been increasing and the semantic prosody of the neologism *kay-* is shifting from negative toward

positive. The shift of the semantic prosody supports the aphorism of usage-based linguistics that meaning is use by showing that the meaning of a lexical item is determined by how people use the lexical item.

Table 6.4 Proportions of positive and negative words based on their token frequencies for each year

Year	Proportion of positive words	Proportion of negative words
2010	26.77	73.23
2011	36.50	63.50
2012	42.17	57.83
2013	43.03	56.97
2014	42.96	57.04
2015	45.98	54.02
2016	46.31	53.69
2017	47.73	52.27
2018	50.06	49.94
2019	53.96	46.04

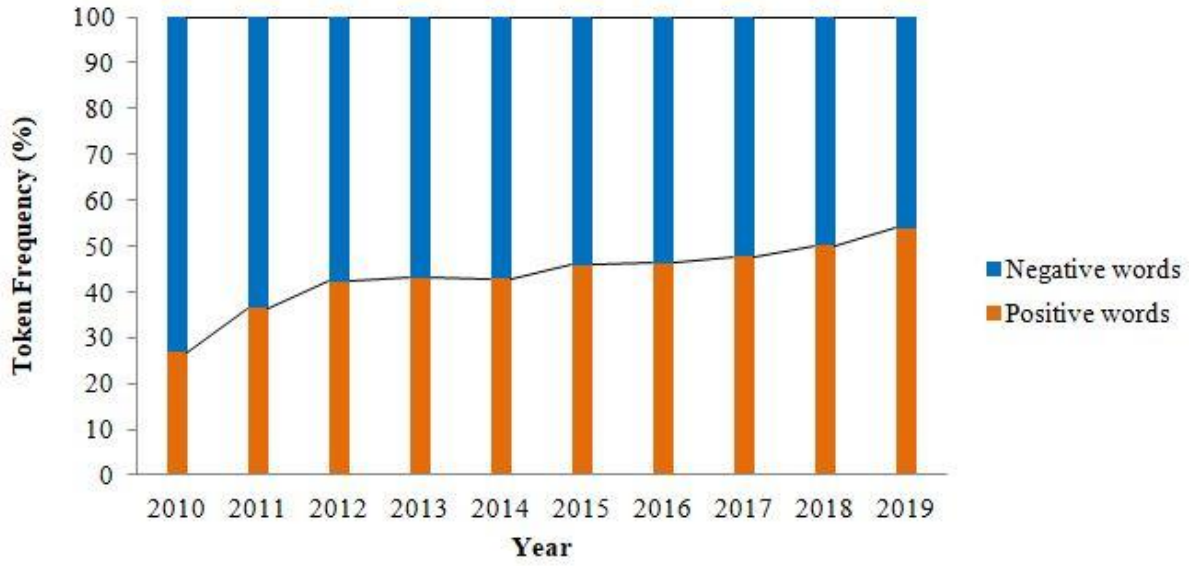


Figure 6.2 Change in the proportions of positive and negative words from 2010 to 2019

6.5 Evaluation

This section presents three metrics to evaluate the performance of the LSTM model: (i) validation accuracy, (ii) a confusion matrix, and (iii) a precision-recall curve. The three metrics are applied to the validation datasets. The results from the three metrics demonstrate that the LSTM model has been trained enough to produce reliable findings with the actual dataset.

6.5.1 Validation Accuracy

Validation accuracy indicates how accurately the LSTM model can predict validation datasets. Here, "predict" means classifying words into positive and negative. Table 6.5 shows the values of validation accuracy from the five validation datasets and their average value. The average value is "0.98188", which implies that the accuracy of the model predicting the validation datasets is about 98 percent on average. Additionally, Table 6.6 shows the values of training accuracy from the five training datasets and their average value. The average value is "0.99038", which means that the accuracy of the model predicting the training datasets is about 99 percent on average. Both of the validation and training accuracy values show that the model has been well trained.

Table 6.5 Validation accuracy for each fold and their average validation accuracy

Fold	Validation accuracy
F1	0.9351
F2	0.9870
F3	0.9935
F4	0.9969
F5	0.9969
Average validation accuracy	0.98188

Table 6.6 Training accuracy for each fold and their average training accuracy

Fold	Training accuracy
F1	0.9842
F2	0.9868
F3	0.9943
F4	0.9933
F5	0.9933
Average training accuracy	0.99038

6.5.2 Confusion Matrix

A confusion matrix is a measure to evaluate the performance of a classification model. In the matrix, each row stands for actual classes while each column stands for predicted classes. Figure 6.3 shows a normalized confusion matrix, which makes it easier to visually interpret the model performance by representing the values in the matrix (i.e., the numbers of classes) as percentages. The top left means True Negative (TN), the top right is False Positive (FP), the bottom right is True Positive (TP), and the bottom left is False Negative (FN). The values in the main diagonal indicate the probability that the model will correctly predict. TN indicates that the probability of the model classifying actual negative words as “negative” is 0.99 (99 percent). TP shows that the probability of the model classifying actual positive words as “positive” is about 0.96 (96 percent). These results also support that the model has been well trained.

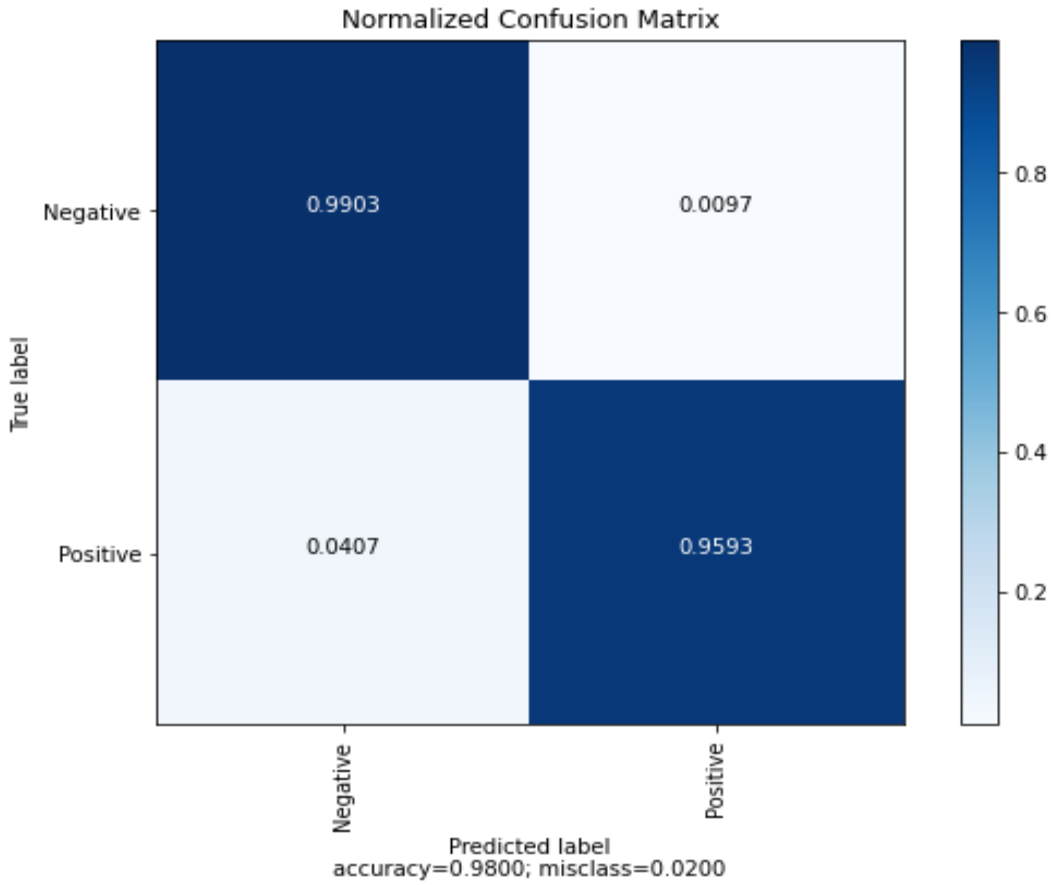


Figure 6.3 Normalized confusion matrix

6.5.3 Precision-recall Curve

A precision-recall curve shows whether a classification model correctly predicts a minority group when classes (in this study, “0” and “1”) are imbalanced in a dataset. Because positive words and phrases (the instances of minority group) are fewer than negative words and phrases (the instances of majority group) in the training and validation datasets, it is required to check whether the model has been well trained in terms of positive words and phrases. In the precision-recall curve in this study, “precision” indicates the capability of the model to accurately predict positive words and phrases and “recall” indicates the capability of the model to detect positive words and phrases out

of the actual positive words and phrases. High precision and high recall make the most ideal model. The precision-recall curve is a plot of the y-axis of precision against the x-axis of recall. The area under the precision-recall curve (AUC-PR) is an indicator showing the performance of the model. The closer to "1" the value of AUC-PR is, the better the predictive model is. Figure 6.4 shows five precision-recall curves from the five folds and their average curve. The average curve shows that its AUC-PR value is "0.9942" (Overall AUC in the figure), which implies that the LSTM model can detect almost all of the actual positive words and phrases and predict positive words and phrases with nearly 100 percent accuracy regarding the validation data sets. These results also demonstrate that the model has been well trained.

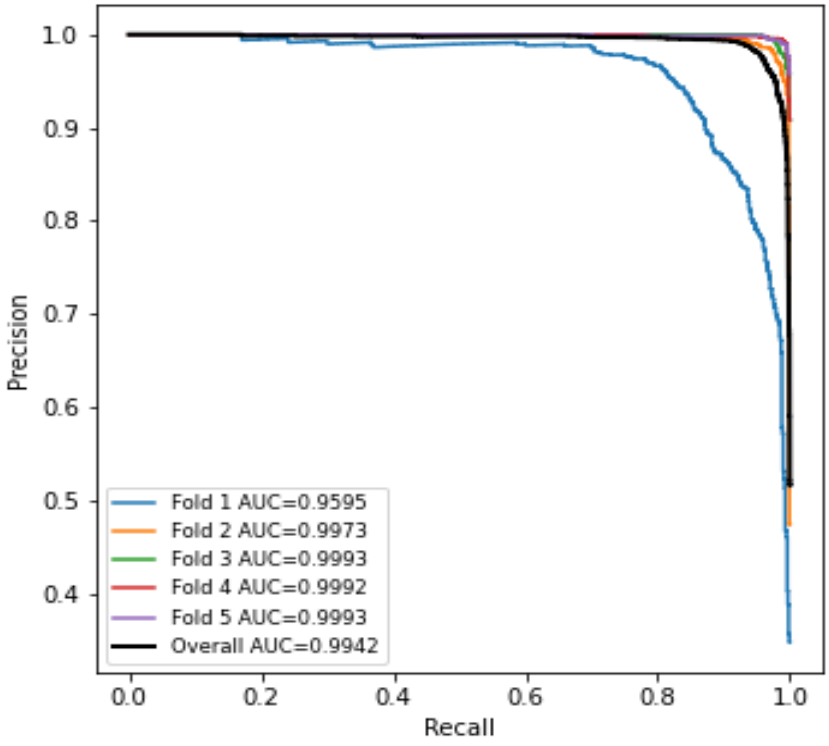


Figure 6.4 Five precision-recall curves and their average curve with the values of AUC-PR

6.6 Method Validation

The employment of LSTM for the analysis of semantic prosody still needs method validation, although the three metrics to assess the performance of the LSTM model demonstrate that the model has been well trained. This is because the evaluation is based on the validation datasets from KNU-KSL. In other words, it is required to evaluate the performance of the model with the actual data (i.e., the adjectives and verbs extracted from the twitter corpora). If there had been a sort of answer sheet showing the extracted adjectives and verbs and their class labels ("0" or "1"), I could have compared the answer sheet with the classification result from the model. However, there was no such thing as an answer sheet so I devised another option.

I decided to select and analyze twenty words for each year because it was impossible to classify every adjective and verb manually. To begin with, before choosing the twenty words, I arranged the adjectives and verbs extracted from each corpus according to their frequency and checked whether the words are the ones following the neologism *kay-*. Contrary to my assumption, most of the verbs were words related to the homonyms of *kay* (in particular, *kay* used as a counting unit for things). Also, the extraction of adjectives was not accurate. Due to the inclusion of wrong verbs and adjectives, it turned out that the semantic prosody result in Section 6.4 might be incorrect (I should have preprocessed the data more carefully).

To find out how the semantic prosody of the neologism *kay-* has developed exactly, I removed the wrong adjectives and verbs related to the homonyms from the above word list. Also, I removed adjectives and verbs which are neutral because the LSTM model classified neutral words as positive or negative. Next, I selected the top twenty words and assessed whether the

LSTM model accurately classified them by sentiment. The classification of two adjectives[17] was different from my classification result. After correcting their sentiment classification, I added up the token frequencies of words in each category (positive and negative) and represented the values as percent for comparison among corpora.

Table 6.7 shows the proportions of positive and negative words based on their token frequencies for each corpus. Figure 6.5 indicates that the proportion of positive words has increased overall (this result is unexpectedly the same as the result from the LSTM analysis including wrong adjectives and verbs). The shift from negative toward positive clearly explains the many encounters of combinations of the neologism *kay-* and a positive word. The change of semantic prosody will be able to support usage-based linguistics as empirical evidence proving that the meaning of a word is determined by how people use the word.

Table 6.7 Proportions of positive and negative words based on their token frequencies for each corpus

Year	Positive words	Negative words
2010	38.22	61.78
2011	53.12	46.88
2012	55.72	44.28
2013	60.52	39.48
2014	62.97	37.03
2015	69.87	30.13
2016	71.67	28.33
2017	70.07	29.93
2018	66.77	33.23
2019	70.06	29.94

[17] The two adjectives are *쩔다* (to be excellent) and *부럽다* (to be jealous). LSTM classified the former into "negative" and the latter into "positive". They occurred in the top twenty throughout the ten years.

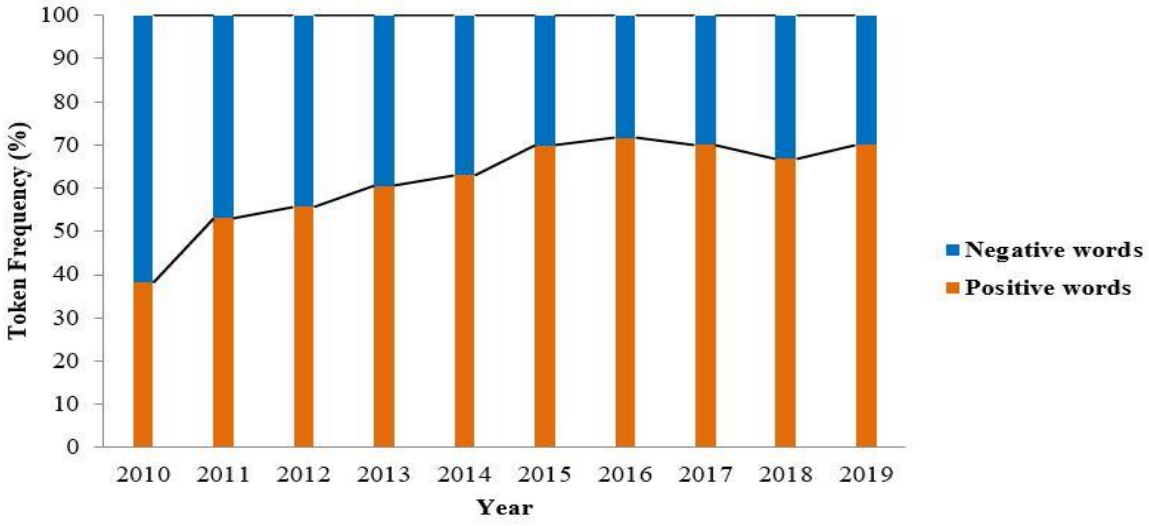


Figure 6.5 Change in the proportions of positive and negative words from 2010 to 2019

To find out on which part of speech the above result (Figure 6.5) is based, I looked into the remaining words after removing wrong words for each corpus. Table 6.8 indicates the top twenty words before and after removing wrong adjectives and verbs for the 2010 corpus. The words in bold type are verbs. Before the removal, the numbers of adjectives and verbs were four and sixteen respectively. However, after the removal, most of the remaining words are adjectives (seventeen adjectives and three verbs).

Table 6.8 Top twenty words before and after removing wrong adjectives and verbs for the 2010 corpus

Top twenty before removing	English meaning	Frequency	Top twenty after removing	English meaning	Frequency
하다	To do	3445	부럽다	To be jealous	397
같다	To be like	1954	웃기다	To make sb laugh	341
먹다	To eat	1125	쩔다	To be excellent	281
짜다	To squeeze/knit	970	피곤하다	To be tired	273
있다	To exist/have	935	망하다	To fail	248
치다	To strike	765	좋다	To be good	213
되다	To become/be	615	춥다	To be cold	210
이다	To be	561	짜증나다	To be annoyed	205
가다	To go	502	귀엽다	To be cute	172
남다	To remain	486	바쁘다	To be busy	155
쓰다	To write/use	428	많다	To be many/much	153
찌다	To steam/gain weight	420	발리다	To be defeated	139
키우다	To raise (a child or animals)	420	빡치다	To be angry	132
부럽다	To be jealous	397	재밌다	To be fun	118
넘다	To exceed/jump over	379	힘들다	To be tough	106
웃기다	To make sb laugh	341	덥다	To be hot	85
나오다	To come out/appear	292	어렵다	To be hard/difficult	82
쩔다	To be excellent	281	무섭다	To be fearful	79
피곤하다	To be tired	273	이쁘다	To be pretty	74
꾸다	To dream/borrow	252	털리다	To be stolen/exhausted	74

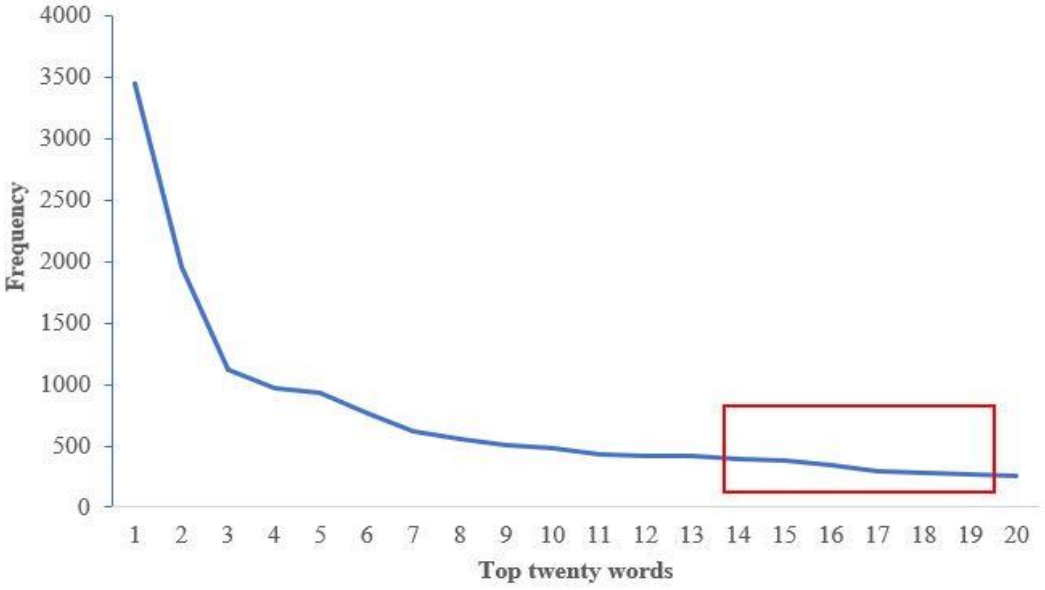


Figure 6.6 Frequency profile of the top twenty words before the removal

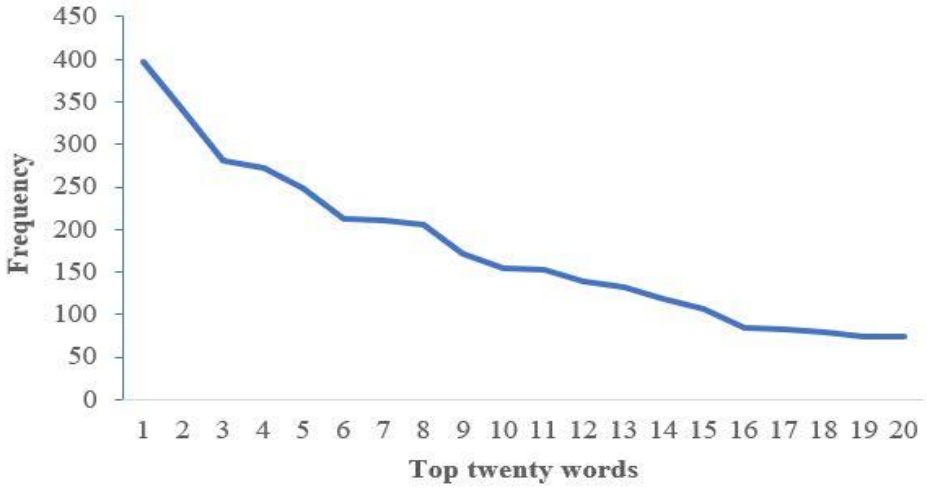


Figure 6.7 Frequency profile of the top twenty words after the removal

Figure 6.6 is the frequency profile of the top twenty words before the removal. It shows an "A-curve" (the top twenty words are represented as numbers). The first four words of the top twenty after the removal are located at the long tail of A-curve (see the box). This shows that the

top twenty words before and after the removal are obviously different from each other. Figure 6.7 is the frequency profile of the top twenty words after the removal. It also shows an "A-curve", although it is shallower than the A-curve before the removal. However, the important thing is the shape of the curve, not the depth of the curve (Kretzschmar 2015: 86). Figure 6.6 and Figure 6.7 support that the distributional frequency profiles of linguistic features always follow the A-curve.

Table 6.9 Numbers of the remaining adjectives and verbs within the top twenty for each corpus

Year	**Remaining adjectives**	**Remaining verbs**
2010	17	3
2011	19	1
2012	19	1
2013	19	1
2014	18	2
2015	18	2
2016	19	1
2017	19	1
2018	19	1
2019	18	2

Table 6.9 shows that the numbers of the remaining adjectives and verbs within the top twenty for each corpus. The number of remaining adjectives is much more than that of remaining verbs for all the ten years. This means that the result (Figure 6.5) is based on adjectives. Another thing that we can infer from Table 6.9 is that the neologism *kay-* is used with particular verbs. Among them, the most frequent verb is 웃기다('to make sb laugh'), which occurs throughout the ten years.

Table 6.10 Difference of classification results between the DCA approach and an analyst

Word	English meaning	Part of speech from DCA	Sentiment from DCA	Part of speech from analyst	Sentiment from analyst
부럽다	To be jealous	Adjective	Positive	Adjective	Negative
웃기다	To make sb laugh	Verb	Positive	Verb	Positive
쩔다	To be excellent	Verb	Negative	Adjective	Positive
피곤하다	To be tired	Adjective	Negative	Adjective	Negative
망하다	To fail	Adjective	Negative	Adjective	Negative
좋다	To be good	Adjective	Positive	Adjective	Positive
춥다	To be cold	Adjective	Negative	Adjective	Negative
짜증나다	To be annoying	Adjective	Negative	Adjective	Negative
귀엽다	To be cute	Adjective	Positive	Adjective	Positive
바쁘다	To be busy	Adjective	Negative	Adjective	Negative
많다	To be many	Adjective	Positive	Adjective	Positive
발리다	To be defeated	Verb	Negative	Verb	Negative
빡치다	To be angry	Adjective	Negative	Adjective	Negative
재밌다	To be fun	Adjective	Positive	Adjective	Positive
힘들다	To be tough	Adjective	Negative	Adjective	Negative
덥다	To be hot	Adjective	Negative	Adjective	Negative
어렵다	To be hard/difficult	Adjective	Negative	Adjective	Negative
무섭다	To be fearful	Adjective	Negative	Adjective	Negative
이쁘다	To be pretty	Adjective	Positive	Adjective	Positive
털리다	To be stolen/exhausted	Verb	Negative	Verb	Negative

Table 6.10 shows the difference of classification results between the DCA approach and an analyst (see the underlined words). They are the results of analysis of the top twenty after the removal from the 2010 corpus. In terms of word class, the classification of one word did not agree with mine. The Korean morpheme analyzer Okt classified the word *쩔다* as "verb" but I classified it as "adjective". This word is slang meaning 'to be excellent' and the combination of *kay-* and the slang word has been more frequently used among teenagers.

In terms of sentiment analysis, the sentiment of two words *쩔다* ('to be excellent') and *부럽다* ('to be jealous') was different from my classification. LSTM classified *쩔다* as "negative" and *부럽다* as "positive". Because the slang word *쩔다* is not contained in the training dataset from KNU-KSL, LSTM can make a wrong prediction of that word. However, LSTM classified *부럽다* as a positive word although it is contained in KNU-KSL as a negative word. Except for the slang word, given that one word out of twenty words is a wrong prediction, it is estimated that about 5% of the total sentiment predictions from LSTM are wrong.

6.7 Discussion

The analysis of the semantic prosody of the neologism *kay-* using LSTM has shown several problems: (i) the classification result of slang was not accurate; (ii) LSTM classified neutral words as positive or negative because it was trained to classify words into positive and negative; (iii) wrong adjectives and verbs related to the homonyms of *kay* were extracted from the corpora. However, these problems can be improved. The inclusion of information on slang in Korean morpheme analyzer tools and training datasets to LSTM will be able to address the first problem. With regard to the second problem, not removing neutral words and phrases from KNU-KSL could address the problem. That is, if LSTM is trained to classify words into positive, neutral, and negative, it will be able to produce more accurate classification results. Concerning the third problem, the employment of techniques related to word-sense disambiguation could be a solution. From this study, it was found that the idea of using word spacing cannot be a proper way to discriminate the target word from its homonyms because people tend to violate rules for word spacing in informal writings. A manual analysis is most exact but it is inefficient when it is needed

to handle a huge amount of data. Therefore, the application of techniques disambiguating word senses[18] would be an optimal solution.

Despite the three problems, the result before correction (Figure 6.2) is not that different from the result after correction (Figure 6.5). Both of them show that the semantic prosody of the neologism *kay-* is shifting from negative toward positive. The comparable results demonstrate that LSTM has worked properly. Also, they show that all of those errors related to the training data did not have a great effect on the actual result. Although the training data before correction has wrong adjectives and verbs and neutral words, it produces a similar result, which implies that the proportion of wrong words in the training data is not great enough to affect the actual result. In other words, if training data includes a sufficient amount of data essential to the actual result, the inclusion of wrong data does not affect the actual result significantly.

The advantage of the DCA approach is that it can overcome some limitations of previous studies. One of the limitations is that semantic prosody is contingent on analysts' interpretation. That is, semantic prosody is the translation of collocate profiles or concordances by analysts. Interpretation involves analysts' subjectivity to a certain degree. Accordingly, the semantic prosody of the same lexical item can vary depending on analysts. Considering that the process of classifying co-occurrences of the target lexical item into favorable and unfavorable is not straightforward, some studies propose methods of quantifying prosodies (e.g., Osgood, Suci & Tannenbaum 1957, Dilts & Newman 2006). Also, others make reference to previous studies sorting lexical items into positive and negative (e.g., Ahmadian, Yazdani & Darabi 2011). However, these are not radical resolutions. Especially, the methods of quantifying prosodies need further studies because they are limited to a particular part of speech.

[18] Studies on Korean word-sense disambiguation have been actively performed (e.g., Jeong & Park 2015, Han et al. 2017, Nguyen et al. 2018, Kim & Kwon 2021).

Another limitation is that previous studies did not suggest efficient methods to deal with large-scale data. According to Louw (1993), semantic prosody is a consistent aura of meaning with which a form is imbued by its collocates. Because the process of a form being imbued with meaning from its collocates requires a lengthy period, semantic prosody research presupposes collection and analysis of a huge amount of data spanning a long period of time. A diachronic analysis over a long period of time is very important for the analysis of semantic prosody. However, few studies have introduced new methods to handle large-scale data more efficiently. As we can see in the section 2.3.1.3, many studies have focused on the analysis of semantic prosody per se, rather than suggesting methods to process large-scale data with minimum wasted effort.

As a solution to both of the limitations from previous studies, I propose an artificial intelligence method. The artificial intelligence method can address the subjectivity issue and handle large datasets efficiently. As the most state-of-the-art technique founded on deep learning, LSTM can resolve the issue over subjective interpretation by classifying words into positive and negative without human subjective judgment (the use of the same LSTM model produces the same results regardless of analysts). Also, LSTM makes it easier to investigate the development of semantic prosody over a long period of time by allowing researchers to process large-scale data efficiently, compared to a manual analysis.

To sum up, the use of LSTM requires analysts to make training data carefully (the ways to improve the three problems found in this study will be able to help to make more accurate training data). Moreover, the comparable results before and after correcting the training data demonstrate that LSTM works properly. Furthermore, it has turned out that if training data to LSTM includes data essential to the actual result and the essential data is enough to unveil the actual result, the inclusion of wrong data does not have a great effect on the actual result. Lastly, the DCA approach

using LSTM is able to help to address the limitations of previous studies. Through the analysis of the semantic prosody of the neologism *kay-*, this study has confirmed the possibility that the DCA approach can be applied to semantic prosody research.

CHAPTER 7

CONCLUSION

7.1 Overview

This study examines how the meanings of three Korean neologisms *leyal*, *lwuce*, and *kay-* have changed over ten years through the analysis of their distributional behaviors in Korean Twitter data. To analyze those distributional behaviors more efficiently, I used distributional corpus analysis (DCA). DCA is an approach which delves into context words using corpora and computational techniques in natural language processing. For the neologism *leyal*, I scraped 15,500 tweets including *leyal* as a keyword per month from 2010 to 2019 (except for 2010, 2015, 2016, and 2019) using snscrape in Python. After preprocessing the Twitter data, I trained word2vec with the preprocessed data. To find out how the two meanings of the neologism *leyal* have developed over the past ten years and which meaning is more predominant between 'really' and 'Real Madrid', I used cosine similarity (CS). Through CS, I measured the semantic similarity between *leyal* and two alternative words representing the two meanings of *leyal* (i.e., *cincca* representing 'really' and *leyalmatulitu* representing 'Real Madrid') and looked into how the CS values between *leyal* and *cincca* (CS1) and those between *leyal* and *leyalmatulitu* (CS2) have changed. The results show that the CS1 values are always higher than the CS2 values. That is, the CS1 and CS2 values subtly change from year to year but the overall trend (CS1 higher than CS2) lasts throughout the ten years. This means that 'really' is more dominant than 'Real Madrid', there is no semantic change in terms of the relation between the two meanings, and the neologism *leyal*

has been more used to represent ‘really’. These findings agree with results from the collocation analysis used for method validation of word2vec and cosine similarity.

For the neologism *lwuce*, I utilized Latent Dirichlet Allocation (LDA) to find out how the new and existing meanings of *lwuce* have developed over time and which meaning holds a predominant position. Because there is no word that can represent the new meaning of *lwuce* among existing words, I chose LDA. Specifically, I used snscrape in Python to collect tweets including *lwuce* as a keyword from 2010 to 2019 with the number of monthly tweets uneven (the minimum number of monthly tweets is 759 and the maximum number is 5,536). In the process of preprocessing, I divided the twitter data into two groups depending on the meanings for each year and turned each group into a subcorpus which only consists of nouns and numbers. After preprocessing, I employed LDA to examine the number of topics for each of the twenty subcorpora (i.e., the ten new meaning subcorpora and the ten existing meaning subcorpora). The change in the number of topics over time shows that (i) the existing meaning is more dominant than the new meaning throughout the ten years, (ii) the use of the new meaning sharply decreased and that of the existing meaning sharply increased in 2015, (iii) the use of the new meaning increased and that of the existing meaning decreased in 2018, and (iv) the use of the new meaning gradually decreased but increased in 2018. The collocation analysis used for method validation of LDA demonstrates that the first two results are true and the fourth one is partially true.

For the neologism *kay-*, I probed its connotational and attitudinal meaning. As the number of combinations of the neologism and a positive word has been increasing these days, I explored how the semantic prosody of the neologism *kay-* has changed over time. I scraped around 35,000 tweets including *kay* as a keyword per month from 2010 to 2019 (except for January and February in 2010 and April in 2011) using snscrape in Python. In the process of preprocessing, I extracted

adjectives and verbs where the neologism *kay-* is attached from the Twitter data. After training long short-term memory (LSTM) to sort words into positive and negative with the preprocessed KNU Korean Sentiment Lexicon, I employed the trained LSTM model to classify the extracted adjectives and verbs into positive and negative. I observed the change of semantic prosody by investigating how the total token frequency of the classified positive words and that of the classified negative words change from year to year. The result shows that the proportion of positive words has been increasing. That is, the semantic prosody of the neologism *kay-* has shifted from negative toward positive over the past ten years. However, it turned out that this result might not be reliable through method validation. Because wrong adjectives and verbs related to the homonyms of *kay* were included in the analysis, it cannot be said that the result shows the semantic prosody of the neologism *kay-*. However, the analysis after removing wrong words and correcting wrong classification showed the same result as the LSTM analysis.

The change of semantic prosody supports the aphorism of usage-based linguistics that meaning comes out of language use (Tomasello 2009) by demonstrating that the meaning of a lexical item is determined by how people use the lexical item. That meaning is determined by language use implies that experience is essential to language acquisition (Tomasello 2003) because it means that we need continuous exposure to language experience to acquire meaning. As empirical evidence supporting the aphorism that meaning is use, the change of the semantic prosody of *kay-* will be able to make a contribution to strengthening the theoretical basis of usage-based linguistics.

7.2 Significance

This study made several “first attempts”. First, this work is the first study using artificial intelligence and Korean social media data to analyze the distributional behaviors of Korean

neologisms and track their semantic change over time. It shows how distributional corpus analysis (DCA) can be used to explore semantic change of Korean words. Because there are few studies employing the DCA approach, I expect that this study could encourage linguists to use the DCA approach, in particular artificial intelligence techniques, to do research on Korean lexical semantics.

Secondly, this work is the first study showing DCA from the perspective of corpus linguistics. It covers the processes of data collection and preprocessing in detail as well as analysis methods. It is a first attempt to connect the DCA approach with corpus linguistics. It would be helpful for corpus linguists who want to apply the DCA approach to their own corpora but do not have any background in that approach. This study will lead corpus linguists to explore linguistic phenomena using the DCA approach.

Thirdly, this work established specific methods to validate the DCA approach using a collocation analysis in corpus linguistics for the first time. It shows that the approach of applying A-curve frequency profiles to collocation analysis works well for method validation of the DCA approach. This attempt is worth doing per se because methods related to corpus linguistics and language as a complex system have never been tried so far. Compared to techniques in the DCA approach whose internal workings are not transparent, a collocation analysis is clear and obvious. Therefore, a collocation analysis will be able to serve as a good method for the validation of the DCA approach. The use of a collocation analysis is expected to stimulate interdisciplinary research between corpus linguistics and natural language processing.

This pioneer study making several “first attempts” will function as a foundational study upon which further DCA studies can build in corpus linguistics. Although this study was carried out with the Korean language, the methodology can be applied to other languages as well. The

application of the DCA approach to lexical semantic research of various languages could lead to active DCA research, resulting in the development of the DCA approach.

7.3 Conclusion

The biggest benefit that we can get by capitalizing on the DCA approach is efficiency. Analyzing a huge amount of linguistic data manually demands much time and energy. However, an analysis using computational techniques in natural language processing saves time and costs. As the entire process from data collection to preprocessing to analysis is performed through computer code, analysts can obtain and analyze data more conveniently and faster. Another advantage of the DCA approach is quantification. It helps researchers to analyze data in a clear and objective way. The analysis using word2vec and cosine similarity shows how semantically similar *leyal* is to the two alternative words through quantification. The analysis employing LDA uses the number of topics to show how the uses of the two meanings of *lwuce* have changed. Because quantification allows us to see language change by means of numbers, we can observe language change more clearly. Also, the subjectivity of interpreting results can be excluded.

However, the analysis by a computer has lower accuracy than analysis by a human. Since computers cannot understand human language, it is required to convert it to numbers so that they can process it. Computers handle not "language" but "number" so their analysis cannot be perfectly the same as a human's analysis. In the case of LDA, it has two big problems: (i) not every word in a cluster semantically relates to the others and (ii) topics are difficult to infer from clusters. This is because LDA uses statistics to classify words without considering word meaning. This study demonstrates that the DCA approach still needs method validation. We need to check whether the employed models have been well trained and their results are reliable. Validation is necessary even for models trained in supervised learning like LSTM, although metrics to assess their performance

show that their performance is good. Through the method validation of LSTM, it was found that wrong words were included in the data, which might have been overlooked without such a validation process.

When we use the DCA approach, the most important thing is that we should not rely blindly on its results. We should doubt them. The current level of computer analysis of linguistic data falls short of human analysis. An analysis depending on statistical processes cannot explain every aspect of language. Language is not numeric. Language should not be approached like mathematical problems which can be answered by the application of math formulas. Because there are many variables and exceptions in language, all of those things should be considered to understand language properly. That is where the expertise of linguists comes in. Computer scientists and natural language processing engineers as well as analysts using natural language processing techniques should keep this in mind. I do not argue that the DCA approach is unreliable so we should not use it. My point is that we should employ techniques in the DCA approach more prudently and accept its results more carefully.

REFERENCES

Adolphs, Svenja & Ronald Carter. 2002. Point of view and semantic prosodies in Virginia Woolf's To the Lighthouse. *Poetica* 58. 7-20.

Ahmadian, Moussa, Hooshang Yazdani & Ali Darabi. 2011. Assessing English Learners' Knowledge of Semantic Prosody through a Corpus-Driven Design of Semantic Prosody Test. *English Language Teaching* 4(4). 288-298.

Alcaraz-Mármol, Gema & Jorge Soto Almela. 2016. The semantic prosody of the words inmigración and inmigrante in the Spanish written media: A corpus-based study of two national newspapers. *Revista Signos* 49(91). 145-167.

Alrajhi, Mashael. 2019. The Semantic Prosody and Semantic Preference of Maximizers in Saudi EFL Writings. *International Journal of Language and Linguistics* 6(3). 30-39.

Arnon, Inbal & Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62(1). 67-82.

Atkins, Beryl T. Sue & Beth Levin. 1995. Building on a corpus: A linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography* 8(2). 85-114.

Atkins, Beryl T. Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1). 1-16.

Bamler, Robert & Stephan Mandt. 2017. Dynamic word embeddings. *Proceedings of the 34th International Conference on Machine Learning*. 380-389.

Bar-Yam, Yaneer. 2009. General features of complex systems. In L. Douglas Kiel (ed.),

Knowledge management, organizational intelligence and learning, and complexity: Vol. 1, 43-95. EOLSS Publishers.

Begagić, Mirna. 2013. Semantic preference and semantic prosody of the collocation make sense. *Jezikoslovlje* 14(2-3). 403-416.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243-257.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Bod, Rens, Jennifer Hay & Stefanie Jannedy (eds.). 2003. *Probabilistic linguistics*. MIT Press.

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6. 213-234.

Bublitz, Wolfram. 1996. Semantic prosody and cohesive company: somewhat predictable. *Leuvense Bijdragen*: *Tijdschrift voor Germaanse Filologie* 85(1-2). 1-32.

Burkette, Allison. 2001. The story of chester drawers. *American Speech* 76(2). 139-157.

Burkette, Allison. 2009. The lion, the witch, and the armoire: Lexical variation in case furniture terms. *American Speech* 84(3). 315-339.

Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form: Vol. 9: Typological studies in language*. John Benjamins Publishing.

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5). 425-455.

Bybee, Joan. 2002. Sequentiality as the basis of constituent structure. In Talmy Givón & Bertram F. Malle (eds.), *The evolution of language out of pre-language*: *Vol. 53: Typological studies in language*, 109-134. John Benjamins Publishing.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711-733.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford University Press.

Bybee, Joan. 2010. *Language, usage, and cognition*. Cambridge University Press.

Bybee, Joan & Paul Hopper (eds.). 2001. *Frequency and the emergence of linguistic structure*: *Vol. 45: Typological studies in language*. John Benjamins Publishing.

Bybee, Joan & Sandra A. Thompson. 2000. Three frequency effects in syntax. *Berkeley Linguistics Society* 23. 65-85.

Caines, Andrew. 2012. You talking to me? Corpus and experimental data on the zero auxiliary interrogative in British English. In Stefan Th. Gries & Dagmar Divjak (eds.), *Volume 1: Frequency Effects in Language Learning and Processing*, 177-206. De Gruyter Mouton.

Caldwell-Harris, Catherine, Jonathan Berant & Shimon Edelman. 2012. Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics. In Dagmar Divjak & Stefan Th. Gries (eds.), *Volume 2: Frequency effects in language representation*, 165–194. De Gruyter Mouton.

Chandra, Arijit & Sunil Kuma Khatri. 2019. Spam SMS filtering using recurrent neural network and long short term memory. *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, 118-122. IEEE.

Chater, Nick & Christopher D. Manning. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Science* 10(7). 335-344.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1968. *Language and mind.* New York: Harcourt Brace Jovanovich.

Chomsky, Noam. 1980. Rules and representations. *Behavioral and Brain Sciences* 3. 1-61.

Chomsky, Noam. 1986. *Knowledge of language*. Berlin: Praeger.

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22-29.

Church, Kenneth Ward, William Gale, Patrick Hanks & Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik (ed.), *Lexical acquisition: exploiting on-line resources to build a lexicon*, 115-164. Hillsdale, NJ: Lawrence Erlbaum.

Church, Kenneth Ward, William Gale, Patrick Hanks, Donald Hindle & Rosamund Moon. 1994. Lexical substitutability. In Beryl T. Sue Atkins & Antonio Zampolli (eds.), *Computational approaches to the lexicon*, 153-177. Oxford University Press.

Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. University of Chicago Press.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Da Xu, Li, Yang Lu & Ling Li. 2021. Embedding blockchain technology into IoT for security: A survey. *IEEE Internet of Things Journal* 8(13). 10452-10473.

Das, Rajarshi, Manzil Zaheer & Chris Dyer. 2015. Gaussian lda for topic models with word embeddings. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 795-804.

de Bot, Kees, Marjolijn Verspoor & Wander Lowie. 2005. Dynamic systems theory and applied linguistics: The ultimate “so what”? *International Journal of Applied Linguistics* 15(1). 116-118.

Del Tredici, Marco, Raquel Fernández & Gemma Boleda. 2019. Short-term meaning shift: a

distributional exploration. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2069-2075. The Association for Computational Linguistics.

Dilts, Philip & John Newman. 2006. A note on quantifying 'good' and 'bad' prosodies. *Corpus Linguistics and Linguistic Theory* 2(2). 233-242.

Dobric, Nikola. 2013. *Theory and practice of corpus-based semantics*. BoD–Books on Demand.

Dörnyei, Zoltán. 2009. *The psychology of second language acquisition*. Oxford University Press.

Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2). 143-188.

Ellis, Nick C. 2009. Optimizing the input: Frequency and sampling in usage-based and form-focused learning. In Michael H. Long & Catherine J. Doughty (eds.), *The handbook of language teaching*, 139-158. Chichester: Wiley-Blackwell.

Firth, John Rupert. 1935. The Technique of Semantics. *Transactions of the Philological Society* 34(1). 36-73.

Firth, John Rupert. 1957a. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*. Oxford: Blackwell.

Firth, John Rupert. 1957b. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Firth, John Rupert. 1968. A synopsis of linguistic theory 1930-55. In Frank Robert Palmer (ed.), *Selected Papers of J. R. Firth 1952-59*, 168-205. London: Longmans.

Francis, W. Nelson. 1992. Language Corpora BC. In Jan Svartvik (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4—8 August 1991*, 17-31.

Francis, W. Nelson & Henry Kučera. 1979. Brown corpus manual. *Letters to the Editor* 5(2). 7.

Frermann, Lea & Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics* 4. 31-45.

Gablasova, Dana, Vaclav Brezina & Tony McEnery. 2017. Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67(S1). 155-179.

Geeraerts, Dirk. 2002. The theoretical and descriptive development of lexical semantics. In Leila Behrens & Dietmar Zaefferer (eds.), *The lexicon in focus: competition and convergence in current lexicology*, 23-42. Peter Lang Verlag.

Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford University Press.

Gilquin, Gaëtanelle. 2003. Causative get and have: so close, so different. *Journal of English Linguistics* 31(2). 125-148.

Givón, Talmy. 1995. *Functionalism and grammar*. John Benjamins Publishing.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press.

Gulordava, Kristina & Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*. 67-71. Association for Computational Linguistics.

Halliday, Michael Alexander Kirkwood. 1966. Lexis as a linguistic level. In Charles Ernest Bazell, John Cunnison Catford, Michael Alexander Kirkwood Halliday & Robert Henry Robins (eds.), *In memory of J. R. Firth*, 148-163. London: Longman.

Halliday, Michael Alexander Kirkwood & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1489-1501.

Han, Kijong, Sangha Nam, Jiseong Kim, YoungGyun Hahm & Key-Sun Choi. 2017. Unsupervised Korean word sense disambiguation using CoreNet. *Proceedings of the 29th Annual Conference on Human and Cognitive Language Technology*. 153-158.

Harris, Zellig S. 1954. Distributional structure. *Word* 10(2-3). 146-162.

Hatfield, Sudarat Leerabhandh. 2005. *Lexical variation of Chiangmai dialect in Chiangmai Province in Thailand.* University of Georgia Dissertation.

Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39(6). 1041-1070.

Hellrich, Johannes & Udo Hahn. 2017. Exploring diachronic lexical semantics with JeSemE. *Proceedings of ACL 2017, System Demonstrations*. 31-36.

Hoey, Michael. 1991. *Patterns of lexis in text.* Vol. 299. Oxford: Oxford University Press.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Hoover, Sandra Elaine. 2001. *Lexical variation and change in farming words: 1970-2001*. University of Georgia Dissertation.

Hopper, Paul. 1987. Emergent grammar. *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*. 139-157.

Hopper, Paul & Elizabeth Closs Traugott. 1993. *Grammaticalization*. Cambridge University Press.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Ernst Klett Sprachen.

Hunston, Susan & Geoffrey Thompson (eds.). 1999. *Evaluation in text: Authorial stance and the construction of discourse*. Oxford University Press.

Jatowt, Adam & Kevin Duh. 2014. A framework for analyzing semantic change of words across time. *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. 229-238. IEEE Press.

Jeong, Hanjo & Byeonghwa Park. 2015. Korean word sense disambiguation using dictionary and corpus. *Journal of Intelligence and Information Systems* 21(1). 1-13.

Johnson, Ellen. 1996. *Lexical change and variation in the southeastern United States in the twentieth century*. University of Alabama Press.

Jones, Susan & John Sinclair. 1974. English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie* 24. 15-61.

Kang, Beom-mo. 2008. Building corpora and making use of frequency (statistics) for linguistic descriptions. *Journal of Korealex* 12. 7-40.

Katz, Jerrold J. & Jerry A. Fodor. 1963. The structure of a semantic theory. *Language* 39(2). 170-210.

Kennedy, Graeme. 1991. Between and through: The company they keep and the functions they serve. In Karin Aijmer & Bengt Altenberg (eds.), *English corpus linguistics: Studies in honour of Jan Svartvik*, 95-110. London: Longman.

Kim, Minho & Hyuk-Chul Kwon. 2021. Word sense disambiguation using prior probability estimation based on the Korean WordNet. *Electronics* 10(23). 2938.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. ACL.

Kjellmer, Göran. 1994. *A Dictionary of English Collocations: Based on the Brown Corpus: in three volumes*. Vol. 1. Oxford University Press.

Kjellmer, Göran. 2003. Synonymy and corpus work: on almost and nearly. *ICAME Journal* 27. 19-27.

Kretzschmar, William A., Jr. 2009. *The linguistics of speech*. Cambridge University Press.

Kretzschmar, William A., Jr. 2015. *Language and complex systems*. Cambridge University Press.

Kretzschmar, William A., Jr. Virginia G. McDavid, Theodore K. Lerud & Ellen Johnson. 1993. *Handbook of the linguistic atlas of the Middle and South Atlantic States*. University of Chicago Press.

Kretzschmar, William A., Jr. 2021. Complex systems for corpus linguists. *ICAME Journal* 45(1). 155-177.

Krug, Manfred. 2003. Frequency as a determinant in grammatical variation and change. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 7-67. Berlin: Mouton de Gruyter.

Kübler, Sandra & Heike Zinsmeister. 2015. *Corpus linguistics and linguistically annotated corpora*. Bloomsbury Publishing.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*. 625-635. International World Wide Web Conferences Steering Committee.

Kumar, D. Ganesh, M. Kameswara Rao & K. Premnath. 2020. A Recurrent Neural Network Model for Spam Message Detection. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 1042-1045. IEEE.

Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the*

mind. University of Chicago Press.

Langacker, Ronald W. 1987a. *Foundations of cognitive grammar: Theoretical prerequisites*. Vol. 1. Stanford University Press.

Langacker, Ronald W. 1987b. An introduction to cognitive grammar. *Cognitive Science* 10(1). 1-40.

Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar: Descriptive application*. Vol. 2. Stanford University Press.

Langacker, Ronald W. 2000. A dynamic usage-based model. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 1-63. Stanford: SLI Publications.

Larsen-Freeman, Diane. 1997. Chaos/complexity science and second language acquisition. *Applied Linguistics* 18(2). 141-165.

Larsen-Freeman, Diane & Lynne Cameron. 2008. *Complex systems and applied linguistics*. Oxford University Press.

Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman & Timothy Baldwin. 2012. Word sense induction for novel sense detection. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 591-601. Association for Computational Linguistics.

Le, Thi Hien Thao & Sung Yeon Kim. 2018. A corpus analysis of collocational behaviors of near-synonymous adjectives. *Multimedia-Assisted Language Learning* 21(4). 181-210.

Lehecka, Tomas. 2015. Collocation and colligation. *Handbook of pragmatics online*. Benjamins.

Lindblom, Björn, Peter MacNeilage & Michael Studdert-Kennedy. 1984. Self-organizing processes and the explanation of phonological universals. In Brian Butterworth, Bernard

Comrie & Östen Dahl (eds.), *Explanations for language universals*, 181-204. Berlin/New York/Amsterdam: Mouton.

Liu, Dilin. 2010. Is it a chief, main, major, primary, or principal concern?: A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics* 15(1). 56-87.

Liu, Dilin & Maggie Espino. 2012. Actually, Genuinely, Really, and Truly: A corpus-based Behavioral Profile study of near-synonymous adverbs. *International Journal of Corpus Linguistics* 17(2). 198-228.

Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 157-176. John Benjamins Publishing.

Louw, Bill. 1997. The role of corpora in critical literary appreciation. In Anne Wichmann, Steven Fligelstone, Tony McEnery & Gerry Knowles (eds.), *Teaching and language corpora*, 240-251. New York: Addison Wesley Longman.

Louw, Bill. 2000. Contextual prosodic theory: Bringing semantic prosodies to life. In Chris Heffer & Helen Sauntson (eds.), *Words in context*: *A tribute to John Sinclair on his retirement*, 48-94. Birmingham: ELR.

Lyons, John. 1963. *Structural Semantics*. Oxford: Blackwell.

McEnery, Tony & Anita Wilson. 2001. *Corpus linguistics: an introduction*. Edinburgh University Press.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Mello, Heather Lee. 2013. *Analysis of language variation and word segmentation for a corpus of*

Vietnamese blogs: a sociolinguistic approach. University of Georgia Dissertation.

Mitchell, Melanie. 2009. *Complexity: A guided tour*. Oxford University Press.

Moholkar, Kavita, Krupa Rathod, Krishna Rathod, Mritunjay Tomar & Shashwat Rai. 2019. Sentiment Classification Using Recurrent Neural Network. In S. Balaji, Álvaro Rocha & Yi-Nan Chung (eds.), *Intelligent Communication Technologies and Virtual Mobile Networks*, 487-493. Springer, Cham.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41(2). 1-69.

Nguyen, Quang-Phuoc, Anh-Dung Vo, Joon-Choul Shin & Cheol-Young Ock. 2018. Effect of word sense disambiguation on neural machine translation: A case study in Korean. *IEEE Access* 6. 38512-38523.

Osgood, Charles Egerton, George J. Suci & Percy H. Tannenbaum. 1957. *The measurement of meaning*. Urbana: University of Illinois press.

Park, Sang-Min, Chul-Won Na, Min-Seong Choi, Da-Hee Lee & Byung-Won On. 2018. Bi-LSTM 기반의 한국어 감성사전 구축 방안 [KNU Korean sentiment lexicon: Bi-LSTM-based method for building a Korean sentiment lexicon]. *Journal of Intelligence and Information Systems* 24(4). 219-240.

Partington, Alan. 1991. *A corpus-based study of the collocational behaviour of amplifying intensifiers in English*. Unpublished M.A. thesis. University of Birmingham.

Partington, Alan. 1998. *Patterns and meanings: using corpora for English language research and teaching*. John Benjamins Publishing.

Partington, Alan. 2004. Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1). 131-156.

Phoocharoensil, Supakorn. 2021. Semantic prosody and collocation: A corpus study of the near-synonyms persist and persevere. *Eurasian Journal of Applied Linguistics* 7(1). 240-258.

Porzig, Walter. 1934. Wesenhafte Bedeutungsbeziehungen [essential meaning relations]. *Beiträge zur Geschichte der Deutschen Sprache und Literatur* 58. 70-97.

Romberg, Alexa R. & Jenny R. Saffran. 2010. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(6). 906-914.

Rosenfeld, Alex & Katrin Erk. 2018. Deep neural models of semantic shift. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 474-484.

Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. *Proceedings of the EACL 2009 Workshop on Geometrical Models of Natural Language Semantics*. 104-111.

Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2011. Tracing semantic change with latent semantic analysis. In Kathryn Allan & Justyna A. Robinson (eds.), *Current methods in historical semantics*, 161-183. De Gruyter Mouton.

Sankoff, David, Rejean Lessard & Nguyen Ba Truong. 1977. Computational linguistics and statistics in the analysis of the Montreal French Corpus. *Computers and the Humanities*. 185-191.

Schneider, Edgar W. 1997. Chaos theory as a model for dialect variability and change? In Alan Robert Thomas (ed.), *Issues and methods in dialectology*, 22-36. Bangor: University of Wales.

Shi, Chungkon. 2020. 계량적 방법을 이용한 트위터 언어의 특징 연구 – 구어와 문어의 언어 양상을 중심으로 [A study on the characteristics of Twitter language using quantitative methods - focusing on the linguistic aspects of spoken and written languages]. *한국어문교육* 31. 111-142.

Sinclair, John (ed.). 1987a. *Looking up: An account of the COBUILD project in lexical computing*. London/Glasgow: Collins ELT.

Sinclair, John. 1987b. Collocation: a progress report. In Ross Steele & Terry Threadgold (eds.), *Volume 2: Language topics: Essays in honour of Michael Halliday*, 319-331. John Benjamins Publishing.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, John. 1996. The search for units of meaning. *Textus: English Studies in Italy* 9(1). 75-106.

Sinclair, John. 2003. *Reading concordances: An introduction*. Pearson/Longman.

Stewart, Dominic. 2010. *Semantic prosody: A critical evaluation*. Routledge.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(1). 23-55.

Stubbs, Michael. 2001a. On inference theories and code theories: Corpus evidence for semantic schemas. *Text & Talk* 21(3). 437-465.

Stubbs, Michael. 2001b. *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell publishers.

Szymanski, Terrence. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 448-453.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change.

Talmy, Leonard. 1983. How language structures space. In Herbert L. Pick Jr. & Linda P. Acredolo (eds.), *Spatial orintation: theory, research, and application*, 225-282. New York: Plenum Press.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Vol. 6. John Benjamins Publishing.

Tognini-Bonelli, Elena. 2004. Working with corpora: Issues and insights. In Caroline Coffin, Ann Hewings & Kieran O'Halloran (eds.), *Applying English grammar: Functional and corpus approaches*, 11-24. London: Arnold.

Tomasello, Michael. 2003. *Constructing a language: a usage-based theory of language acquisition*. Harvard University Press.

Tomasello, Michael. 2009. The usage-based theory of language acquisition. In Edith L. Bavin (ed.), *The Cambridge handbook of child language*, 69-87. Cambridge University Press.

Van Geert, Paul. 1991. A dynamic systems model of cognitive and language growth. *Psychological Review* 98(1). 3-53.

Weisgerber, Johann Leo. 1927. Die Bedeutungslehre: ein Irrweg der Sprachwissenschaft? [Semantic theory: a wrong direction in linguistics?]. *Germanisch-Romanische Monatsschrift* 15. 161-183.

Whitsitt, Sam. 2005. A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics* 10(3). 283-305.

Widdowson, Henry G. 2000. On the limitations of linguistics applied. *Applied Linguistics* 21(1). 3-25.

Wijaya, Derry Tanti & Reyyan Yeniterzi. 2011. Understanding semantic change of words over

centuries. *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural DiversiTy on the Social Web.* 35-40.

Xiao, Richard & Tony McEnery. 2006. Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics* 27(1). 103-129.

Zhang, Rui-Hua. 2013. A corpus-based study of semantic prosody change: The case of the adverbial intensifier. *Concentric*: *Studies in Linguistics* 39(2). 61-82.