Ph.D. DISSERTATION

Word Embedding-Based Semantic Analysis of English Loanwords in Japanese and Korean

단어임베딩을 이용한 일본어와 한국어에서의 영어 외래어 의미분석

BY

Akihiko Yamada

February 2021

DEPARTMENT OF LINGUISTICS
COLLEGE OF HUMANITIES
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Word Embedding-Based Semantic Analysis of English Loanwords in Japanese and Korean

단어임베딩을 이용한 일본어와 한국어에서의 영어 외래어 의미분석

BY

Akihiko Yamada

February 2021

DEPARTMENT OF LINGUISTICS
COLLEGE OF HUMANITIES
SEOUL NATIONAL UNIVERSITY

Word Embedding-Based Semantic Analysis of English Loanwords in Japanese and Korean

단어임베딩을 이용한 일본어와 한국어에서의 영어 외래어 의미분석

지도교수 신 효 필

이 논문을 언어학박사 학위논문으로 제출함

2021년 2월

서울대학교 대학원

언어학과 언어학전공

야마다 아키히코

야마다 아키히코의 언어학박사 학위논문을 인준함

2021년 2월

위 원 장: 남승호

부위원장: 신효필

위 원: 김문형

위 원: 유현조

위 원: 정선우

Abstract

Through cultural exchanges with foreign countries, a lot of foreign words have entered another country with a foreign culture. These foreign words, **loanwords**, have broadly prevailed in languages all over the world.

Historical linguistics has actively studied the loanword because loanword can trigger the linguistic change within the recipient language. Loanwords affect existing words and grammar: native words become obsolete, foreign suffixes and words coin new words and phrases by combining with the native words in the recipient language, and foreign prepositions are used in the recipient language. Loanwords themselves also undergo language changes–morphological, phonological, and semantic changes–because of linguistic constraints of recipient languages through the process of integration and adaptation in the recipient language. Several fields of linguistics–morphology, phonology, and semantics–have studied these changes caused by the invasion of loanwords.

Mainly loanwords introduce to the recipient language a completely new foreign product or concept that can not be expressed by the recipient language words. However, people often use loanwords for giving prestigious, luxurious, and academic images. These sociolinguistic roles of loanwords have recently received particular attention in sociolinguistics and pragmatics.

Most previous works of loanwords have gathered many examples of loanwords and summarized the linguistic change patterns. Recently, corpus-based quantitative studies have started to statistically reveal several linguistic factors such as the word length influencing the successful integration and adaptation of loanwords in the recipient language. However, these frequency-based researches have difficulties quantifying the complex semantic information. Thus, the quantitative analysis of the loanword semantic phenomena has remained undeveloped.

This research sheds light on the quantitative analysis of the semantic phenomena

of loanwords using the Word Embedding method. Word embedding can effectively convert semantic contextual information of words to vector values with deep learning methods and big language data. This study suggests several quantitative methods for analyzing the semantic phenomena related to the loanword. This dissertation focuses on three topics of semantic phenomena related to the loanword: **Lexical competition**, **Semantic adaptation**, and **Social semantic function and the cultural trend change**.

The first study focuses on the lexical competition between the loanword and the native synonym. Frequency can not distinguish the types of a lexical competition: *Word replacement* or *Semantic differentiation*. Judging the type of lexical competition requires to know the context sharing condition between loanwords and the native synonyms. We apply the geometrical concept to modeling the context sharing condition. This geometrical word embedding-based model quantitatively judges what lexical competitions happen between the loanwords and the native synonyms.

The second study focus on the semantic adaptation of English loanwords in Japanese and Korean. The original English loanwords undergo semantic change (semantic adaptation) through the process of integration and adaptation in the recipient language. This study applies the transformation matrix method to compare the semantic difference between the loanwords and the original English words. This study extends this transformation method for a contrastive study of the semantic adaptation of English loanwords in Japanese and Korean.

The third study focuses on the social semantic role of loanwords reflecting the current cultural trend in Japanese and Korean. Japanese and Korean society frequently use loanwords when new trends or issues happened. Loanwords seem to work as signals alarming the cultural trend in Japanese and Korean. Thus, we propose the hypothesis that loanwords have a role as an indicator of the cultural trend change. This study suggests the tracking method of the contextual change of loanwords through time with the pre-trained contextual embedding model (BERT) for verifying this hypothesis.

This word embedding-based method can detect the cultural trend change through the contextual change of loanwords.

Throughout these studies, we used our methods in Japanese and Korean data. This shows the possibility for the computational multilingual contrastive linguistic study. These word embedding-based semantic analysis methods will contribute a lot to the development of computational semantics and computational sociolinguistics in various languages.

keywords: Big data, Deep learning, Word embeddings, Loanword, Lexical competition, Semantic change, Sociolinguistics, Cultural trend change detection

student number: 2014-30878

Contents

Abstract **i**

Contents **iv**

List of Tables **viii**

List of Figures **xi**

1 Introduction **1**

1.1 Overview of Loanword Study 1

1.2 Research Topics in this Dissertation 6

1.2.1 Lexical Competition between Loanword and Native Synonym 6

1.2.2 Semantic Adaptation of Loanwords 8

1.2.3 Social Semantic Function and the Cultural Trend Change . . . 11

1.3 Methodological Background 14

1.3.1 The Vector Space Model 14

1.3.2 The Bag of Words Model 15

1.3.3 Neural Network and Neural Probabilistic Language Model . . 15

1.3.4 Distributional Model and Word2vec 18

1.3.5 The Contextual Word Embedding and BERT 21

1.4 Summary of this Chapter 23

2 Word Embeddings for Lexical Changes Caused by Lexical Competition between Loanwords and Native Words **25**

2.1 Overview 25

2.2 Related Works 28

2.2.1 Lexical Competition in Loanword 28

2.2.2 Word Embedding Model and Semantic Change 30

2.3 Selection of Loanword and Korean Synonym Pairs 31

2.3.1 Viable Loanwords 31

2.3.2 Previous Approach: The Relative Frequency 31

2.3.3 New Approach: The Proportion Test 32

2.3.4 Technical Challenges for Performing the Proportion Test . . . 32

2.3.5 Filtering Procedures 34

2.3.6 Handling Errors 35

2.3.7 Proportion Test and Questionnaire Survey 36

2.4 Analysis of Lexical Competition 38

2.4.1 The Geometrical Model for Analyzing the Lexical Competition 39

2.4.2 Word Embedding Model for Analyzing Lexical Competition . 44

2.4.3 Result and Discussion 44

2.5 Conclusion and Future Work 48

3 Applying Word Embeddings to Measure the Semantic Adaptation of English Loanwords in Japanese and Korean **51**

3.1 Overview 51

3.2 Methodology 54

3.3 Data and Experiment 55

3.4 Result and Discussion 58

3.4.1 Japanese 59

3.4.2 Korean 63

3.4.3 Comparison of Cosine Similarities of English Loanwords in Japanese and Korean 68
3.4.4 The Relationship Between the Number of Meanings and Cosine Similarities 75
3.5 Conclusion and Future Works 77

4 Detection of the Contextual Change of Loanwords and the Cultural Trend Change in Japanese and Korean through Pre-trained BERT Language Models 78

4.1 Overview 78
4.2 Related Work 81
4.2.1 Loanwords and Cultural Trend Change 81
4.2.2 Word Embeddings and Semantic Change 81
4.2.3 Contextualized Embedding and Diachronic Semantic Representation 82
4.3 The Framework 82
4.3.1 Sense Representation 82
4.3.2 Tracking the Contextual Changes 85
4.3.3 Evaluation of Frame Work 86
4.3.4 Discussion for Framework 89
4.4 The Cultural Trend Change Analysis through Loanword Contextual Change Detection 89
4.4.1 Methodology 89
4.4.2 Result and Discussion 91
4.5 Conclusion and Future Work 96

5 Conclusion and Future Works 97

5.1 Summary 97
5.2 Future Works 99

5.2.1 Revealing Statistical Law 99
5.2.2 Computational Contrastive Linguistic Study 100
5.2.3 Application to Other Semantics Tasks 100

A List of Loanword Having One Synset and One Definition in Korean CoreNet in Chapter 2 112

Abstract (In Korean) 118

List of Tables

1.1 The types of semantic change in Korean(Tyson, 1993) 9

1.2 The types of semantic change in Japanese (Kay, 1995) 10

1.3 Pejoration and Amelioration in Korean loanwords (Noh, 2013) 11

2.1 Summary of selection procedures of loanword-Korean synonym pairs. 36

2.2 The result of proportion test for loanword-Korean synonym pairs. . . 37

2.3 The result of the mathematical analysis of the context sharing condition between the loanword and Korean synonym. This table shows the result of m = 50, 100, 200. 42

2.4 The result of the mathematical analysis of the context sharing condition between the loanword and Korean synonym. This table shows the result of m = 500, 1000, 10000. 43

2.5 The nearest neighbors of the high $\mathcal{P}$ loanword-synonym pairs ("UN" and "Orchestra") and the low $\mathcal{P}$ loanword-synonym pairs ("Comedy" and "Bonus"(m=1000). **L** means nearest neighbors of loanword and **S** means nearest neighbors of synonym. The rightmost column shows the top five nearest neighbors. The same nearest neighbors and the target word itself in nearest neighbors are displayed in bold font. 46

3.1 The previous study's examples of the original English word and Korean loanword pairs which have undergone the semantic adaptation. 57

3.2 The top twenty similar and dissimilar loanword-original English pairs in Japanese. 60
3.3 The frequency and the relative frequency (Rel. freq.) of the twenty lowest cosine similarity English words and Japanese loanword pairs in each language corpus. 61
3.4 The nearest neighbors of five most dissimilar loanword pairs in Japanese. 62
3.5 The top twenty similar and dissimilar loanword-original English pairs in Korean. 65
3.6 The frequency and relative frequency (Rel. freq.) of the twenty lowest cosine similarity English words and Korean loanword pairs in each language corpus. 66
3.7 The nearest neighbors of five most dissimilar loanword pairs in Korean. 67
3.8 The top five English loanwords whose cosine similarity with their Korean (or Japanese) loanword is higher than with their Japanese (or Korean) loanword. 69
3.9 The nearest neighbors of top five English loanwords whose cosine similarity with their Korean loanwords is higher than with their Japanese loanword. 70
3.10 The nearest neighbors of top five English loanwords whose cosine similarity with their Japanese loanwords is higher than with their Korean loanword. 73

4.1 The numbers of contexts in the 10 nearest sentences of the mean vector of COVID-related-loanwords in 2012, 2015, 2019, and 2020. "Unknown" means that the context cannot be interpreted from the contextual information. 87
4.2 The three highest contextual changed loanwords in jBERT and mBERT. 90
4.3 The total numbers of loanwords analyzed in this experiment. 90
4.4 The three highest contextual changed Korean loanwords in mBERT. . 93

4.5 The contextual changes of the three highest contextual changed loan-words at the most changed times. 95

List of Figures

1.1 A simple example of document vectorization in Bag of words. d_i means vector of document i. 15
1.2 A simple model of a neural network. 16
1.3 CBOW and Skip-gram (Mikolov et al., 2013a) 18
1.4 The image of cosine similarity in vector space model. 21
1.5 The Bidirectional Encoder Representations from Transformers (BERT) 22

2.1 The experimental procedure of detecting lexical competition between loanword and Korean synonym 26
2.2 Examples "UN" and "Bonus" of Lexical Competition in Korean language represented by the word embedding model. Loanword, synonym, and their nearest neighbors (n=1000) vectors are spotted by using t-SNE algorithm. The two colors distinguish loanword and synonym nearest neighbors. The bar chart means the relative frequency of loanword (red) and the native synonym(blue). 27
2.3 The geometrical model for the context relation between the loanword and the native synonym in this study.This figure represents a set of vectors originally on the surface of a sphere. 40

2.4 This figure shows loanword-pairs from right to left in descending order of the value $\mathcal{P}$ (m=1000). The 2-D projected vector spaces of the loanword-Korean synonym pairs–"UN", "Orchestra", "Uniform", "Member", "Bonus", and "Comedy"–are shown as an example. 45

3.1 The experimental procedure of training data extraction and transformation. 54

3.2 The result of the statistical test between the number of sense in the English dictionary and the cosine similarity. 76

4.1 Two-dimensional visualization of semantic shift in COVID-19 related Japanese loanwords (クラスター *kurasuta* “cluster", ロックダウン *rokkudaun* “lockdown", オーバーシュート *obashuto* “overshoot", ディスタンス *deisutansu* “distance") using mean vectors of contextualized vectors in Japanese BERT. The semantic usage of these loanwords has dramatically shifted between 2019 and 2020. 79

4.2 The experimental procedure of detecting the contextual change. . . . 83

4.3 The distance of the mean vector in very other years in COVID-19 related Japanese loanwords.The left graph shows the result of jBERT and the right graph shows the result of mBERT. 86

4.4 The distance of the mean vector in every other years in the top three highest contextual changed Japanese loanwords. Left is the loanwords in jBERT and right is the loanwords in mBERT. 91

4.5 The distance of the mean vector in very other years in the top three highest contextual changed Korean loanwords in mBERT. 94

Chapter 1

Introduction

In this chapter, the first section provides the historical flow of loanword research and summarizes the three topics of loanword study covered in this dissertation. The next section summarizes the research topics that this dissertation focuses on. The final section summarizes the word embeddings, which is an essential technique in this dissertation.

1.1 Overview of Loanword Study

For cultural, economic, and political interaction between countries or between social communities, one language contacts with several languages. In language contacts, foreign words often enter another language without being translated. This phenomenon is called linguistic borrowing, and those foreign words are called **Loanwords**. Every human interaction inevitably causes linguistic contacts and the great influx of loanwords must occur in almost every language (Sapir, 1921).

Linguistic borrowing has happened from ancient times to the present day. For example, the influx of wine culture has spread many Latin words, and the spread of Christianity has led to the influx of loanwords into England. Chinese introduced many words into Japanese and Korean, and English imported many words and affixes from French (Sapir, 1921). Currently, loanwords have a large proportion in a lot of languages (Poplack and Sankoff, 1984).

Mainly historical linguists studied linguistic borrowing and loanwords because linguistic borrowing externally causes language change. Before the 18th century, linguists mainly focus on the internal linguistic factors of language change and paid no much attention to linguistic borrowing and loanword (Pedersen and Spargo, 1965).

Since the 19th century, many linguists have studied loanwords. Jespersen (1905) focused on English loanwords and especially on pronunciation changes. Sapir (1921) selected many language examples and mentioned the importance of interinfluences between languages. He introduced the psychological factors related to the degrees of linguistic borrowing in each language. He also focused on the phonological and morphological aspects of loanwords.

Bloomfield (1933) mentioned linguistic borrowing and loanword through three chapters, *Cultural Borrowing*, *Intimate Borrowing*, and *Dialect Borrowing*. He explained the phonological and morphological restrictions which loanwords received in the process of integration and adaptation with a lot of examples of loanwords in various languages. He also organized the current English loanwords at that time and the loanwords that had spread all over the world. The chapter on *Intimate Borrowing* explained the one-sided linguistic borrowing. In this case, an upper or dominant language gives their words to the lower language due to the effects of aggression–for example, England gave a lot of English words to America in the colonization era–. He also described the linguistic borrowing that occurred between social communities.

Haugen (1950) defined the terminology of the loanword study to solve the difference between the behavior of the bilingual speakers and the results of the borrowing research by the linguist. He clearly distinguished linguistic borrowing from the *mixed* language. He defined linguistic borrowing as "the attempted reproduction in one language of patterns previously found in another" and analyzed linguistic borrowing and loanwords.

In his paper, *Importation* means a foreign word came in another language with almost no change and *Substitution* means a foreign word came in with a large change due to the linguistic restrictions of the recipient language. His classification of loanwords have three types like below;

- *Loanwords* show morphemic importation without substitution.
- *Loanblends* show morphemic substitution as well as importation. Loanblends include "hybrid".
- *Loanshifts* show morphemic substitution without importation. Loanshifts include "loan translations" and "semantic loans".

The remaining part of his paper analyzed in detail the loanword phonology, the grammar of loanwords, the structural resistance to borrowing, and the structural effect of borrowing.

Weinreich (1954) explained that a language is in *contact state* when the same person uses more than one language. He defined the state of using two languages as *Bilingualism* and defined such persons as *Bilingual*. Those instances of derivation resulted from the Bilingualism was called *the Interference phenomena*. *Interference* meant a *rearrangement* of language systems–phonetics, morphemes and syntax–caused by the foreign word influx. He argued that understanding bilingualism requires an understanding of the extra-linguistic factors: psychological and socio-cultural settings. The extra-linguistic factors include stereotyped attitudes (prestige) towards each other's languages and differences in tolerance for receiving foreign languages. The interplay of linguistic structural factors and non-structural factors is essential in the study of linguistic interference. In the final chapter, he described multiple linguistic contacts, the language in the Balkan peninsula and Yiddish, and inferred the prospects for research on multiple linguistic contacts in India, Israel, and America.

Lehmann et al. (1962) and Labov (1966) focus on borrowing between social-groups, dialects, and developed the sociolinguistic study on borrowing. Labov (1966) studied English social variations in New York City and discovered language variations such as pronunciation feature correlates with social class and ethnicity in a regular pattern. This borrowing from the upper class or prestige class is an example of *hypercorrection*. His results indicated the similarity between borrowing from a prestigious class and borrowing from a prestigious language. Anttila (1989) also emphasized that language variation is essential to understanding language change and discussed linguistic borrowing in terms of language change.

In addition to the structural and social aspects of linguistic borrowing, some researchers have started to study the pragmatic borrowings, such as discourse markers: *and*, *but*, and *of course*. Hasselmo (1970) studied American-Swedish Bilingualism, and Clyne (1972) studied German-English bilingualism in Australia. Clyne (1972) interviewed 330 Australian German-English bilinguals and showed that "and" and "but" were prominent borrowings in the German-English speech condition. He gave various examples of discourse markers: "well" and "anyway". He also revealed regional differences in the use of "yes and ja" and "no and nein" from his interviewed data.

Hoffer (2002) introduced more recent linguistic borrowing research: the practical loanwords appearing in daily life such as TV, Newspapers, Multimedia, Movies, Advertisements. He gave examples of *Dual Language Neologisms*: "nacho average convenience store" (Spanish + English) in an advertisement, "Bon Voyager" (French + English) in a caption of TV Guide, and "Hairigami" (English + Japanese) of a hair-styling company. Japanese provides several interesting examples. One of them is 美サイレント *bi sairento* "beautiful and silent". 美 *bi* "beautiful" has almost the same pronunciation as English "be" in Japanese. Replacing English "be" of "be silent" with the same pronunciation word 美 *bi* "beautiful" created a new meaning: "silence is beautiful". Japanese provides these complicated and interesting linguistic borrowing phenomenon because of the unique writing system called *Katakana*. The influx of loanwords also dramatically occurs in Japanese. Hoffer suggested that studying Japanese linguistic borrowing will bring significant results in the study of language contact and loanword. Additionally, he suggested that comparing Japanese and other languages will also advance the study of language contact and loanword.

Haspelmath and Tadmor (2009) proceeded *The Loanword Typology project* and built *the World Loanword Database*. While almost traditional loanword researches only found and described several examples of loanwords within a single language, this project has enabled comparative studies of loanwords across many languages. They investigated 41 languages based on a meaning list composed of 1460 meanings of 24 semantic fields. They revealed the differences in borrowing between languages and particularly conduct an experimental study on the borrowability: a degree to which type of words are more easily imported as loanwords.

Recently, many researchers continually have much more attention to sociolinguistics, pragmatic, and comparative study of loanword (Andersen et al., 2017; Peterson and Beers Fägersten, 2018).

Looking back on previous studies of linguistic borrowing, as Haspelmath and Tadmor (2009) mentioned, most previous studies only have described special examples of linguistic borrowing and loanwords in several countries. Analyzing more linguistic data will deeply reveal the social function and semantic phenomena of loanwords. Especially, *big data* has brought great advances in people's life and academic fields (Chen et al., 2014) recently. *Big data* also greatly have impacted the linguistic study (Hirschberg and Manning, 2015) but linguists experience the novelty of big-data-driven research and face difficulties (Lu, 2020). In this situation, big-data-driven loanword research remains unexplored nowadays (Serigos et al., 2017). Particularly, loanword research meets difficulties in language resource-poor languages like Japanese and Korean.

In these situations, presenting a pioneering study of big-data-driven loanwords will contribute to the development of loanword research. Especially, big-data-driven loanword study for Japanese and Korean will break through the difficulties in resource-poor languages study. Motivated by these expectations, this dissertation sheds light on big-data-driven loanword research in Japanese and Korean. We suggest several methods, mainly machine learning methods and deep learning methods, for using big data. The next section summarizes background knowledge on the research topics.

1.2 Research Topics in this Dissertation

When a loanword enters the recipient language, the lexical competition between the loanword and native word can possibly happen if a native word whose meaning is similar to the loanword (the native synonyms) has already existed (Bolinger, 1977; Winter-Froemel et al., 2014). As a result, through the process of language changes, loanwords finally settle in the recipient language. These loanwords are used in a variety of contexts, depending on the social or cultural needs of the recipient language. Many researches have investigated the process of loanword influx, adaptation, and settlement of loanwords and the social or cultural semantic role of loanwords, but quantitative research has remained almost unexplored. Especially, big data and deep learning-based study has been sparse. This dissertation sheds light on the quantitative study on the linguistic phenomena of loanwords by applying the Big-data and Deep learning methods. The following parts explain the linguistic phenomena of loanword in detail in three parts: **Lexical competition**, **Semantic adaptation**, and **Social semantic function and the cultural trend change**.

1.2.1 Lexical Competition between Loanword and Native Synonym

Two main purposes mainly cause the influx of loanwords. Firstly, importing a completely new foreign concept inevitably causes the influx of loanwords because the recipient language has no corresponding word for the concept. Secondly, the loanword influx happens even though corresponding words have already existed in the recipient language. Mostly social linguistics or pragmatic factors, such as prestige, trigger the second influx (Zenner et al., 2019). As mentioned in Weinreich (1954), especially in the second case, loanwords affect the existing vocabularies in the recipient language.

Weinreich (1954) states that new loanwords affect existing vocabulary in three ways unless new loanwords have completely new concept words like below.

1. Confusion between the content of the new and old word: Confusion in usage, or

full identity of content, between the old and the new word is probably restricted to the early stages of language contact.

2. Disappearance of the old word: Old words may be discarded as their content becomes fully covered by the loanword

3. Survival of both the new and old word, with a specialization in content: the content of the clashing old and borrowed words may become specialized.

The concept of the economy will explain the change of existing words (Native words) due to the influx of loanwords. The economy principle or the principle of least effort means the tendency of producing maximum results with minimal effort. This tendency ubiquitously exists everywhere in human activity and many researchers have been investigated in various fields (Bregasi, 2016). In linguistics, Martinet (1955) investigated this principle. Martinet (1955) stated the language economy principle as the eternal demand for linguistic communication, the needs of infinite linguistic units for clear and precise expression, and the inertia of humans, less numerous and less specific, cause optimization of the linguistic system (Bregasi, 2016; Vicentini, 2003).

Zipf (1949) is also a well-known study of this language economy principle. He found Zipf's law: the frequency of the kth frequent word equals $1/k$ of the most frequent word in a text. He explained the principle of least effort induces this frequency tendency in language (Zipf, 1949). Since Zipf's law solves problems not only in linguistics but also in various human activities, many researchers have used this law in various fields (Gabaix, 1999; Adamic and Huberman, 2002).

Bolinger (1977) stated that the existence of two or more words with the same meaning is economically inefficient and eventually causes a meaning change. Loanwords also show the same behavior. If the same meaning native word with a new loanword already exists in the recipient language, two major changes can occur to solve this inefficiency between loanword and native word (Winter-Froemel et al., 2014). Firstly, a loanword and a native word will coexist through undergoing meaning changes: broadening or narrowing. Secondly, through competition between a loanword and a native word, one term will win over the competitor. Word replacement also can occur. Lexical competition is an insightful topic in language contact, but little research has investigated this topic (Winter-Froemel et al., 2014). Particularly, big-data-driven research and research in non-European languages like Japanese or Korean remains unexplored. This dissertation will discuss this lexical competition between loanwords and native words in chapter 3. The first type of change is *Word Replacement* and the second type of change is *Semantic Differentiation* in this dissertation.

1.2.2 Semantic Adaptation of Loanwords

Many loanword studies have researched the semantic change or the semantic adaptation of loanwords in various languages (Tyson, 1993; Kay, 1995; Weinreich, 1954; Pavlou, 1994; Hall-Lew, 2002; Al-Athwary, 2016). Basically, loanwords maintained its original meaning in the recipient language. Names or concrete things, such as *Bus* or *Radio*, have little changed. However, many loanwords have undergone semantic adaptation to the cultural and linguistic constraints of the new language. Kay (1995) mentioned the difficulty to find loanwords that have exactly the same original meaning.

As a pattern of semantic change, Tyson (1993) gave some examples in Korean and Table 1.1 shows the examples. Kay (1995) gave the following Japanese examples and Table 1.2 shows the examples.

Noh (2013) shows other types of semantic adaptations: *Degeneration* or *Pejoration* and *Elevation* or *Amelioration* in Korean loanwords. *Degeneration* or *Pejoration* degrade the original meaning to negative meaning. Euphemisms or taboos usage of loanwords frequently causes this negative change. On the contrary, *Elevation* or *Amelioration* upgrade the original meaning to the positive meaning. The prestige image of loanwords frequently causes this positive change. Table1.3 summarizes the examples of Korean loanwords.

Studying semantic adaptation will reveal the linguistic constraints and cultural

a.Semantic Narrowing	
Loanword	Restricted Meaning
pants	underwear
meeting	blind date
out	baseball term only
tape	recording tape
b. Semantic Widening	
Loanword	Extended Meaning
ice cream	any frozen dessert or snack
service	anything offered free of charge
wine	any alcoholic beverage
c. Semantic Transfer	
Loanword	Shifted Meaning
blues	slow dance
talent	TV actor
cunning	cheating on an examination
Kentucky	fried chicken
mansion	apartment
manicure	fingernail polish

Table 1.1: The types of semantic change in Korean(Tyson, 1993)

a. Semantic Narrowing	
Loanword	Restricted Meaning
film	a roll of film
extra	a film extra
machine	only a sewing machine
tuna	tinned, not fresh tuna
pudding	only to caramel custard pudding
b. Denote Western style	
Loanword	
restaurant	
cup	
table	
apple pie	
c. Semantic Tranfer	
Loanword	Shifted Meaning
mansion	high-class block of flats
front [desk]	reception desk
trump	playing cards
Viking	buffet meal
pot	thermos flask
echo	acoustics
seal	sticky label

Table 1.2: The types of semantic change in Japanese (Kay, 1995)

Type	Loanwords
Pejoration	madam, hostess, glamour, boss,commission connection, broker, premium, claim, rebate, gate
Amelioration	restaurant, cookie, tissue, mood, trend, maker, chef vision, visual, hairshop, hairdresser, syndrome

Table 1.3: Pejoration and Amelioration in Korean loanwords (Noh, 2013)

influence in the loanword adaptation. These semantic differences will also make some trouble in language learning and international communication. Chapter three investigates the semantic adaptation of loanwords in Japanese and Korean using big data and deep learning methods.

1.2.3 Social Semantic Function and the Cultural Trend Change

As mentioned above, loanwords can enter the recipient language even though the same concept or meaning native words have already existed. Social or pragmatic reasons mainly will cause this type of inflow of loanwords. Onysko and Winter-Froemel (2011) states that many linguists distinguished loanwords: *Necessary loans* and *Luxurious loans*. *Necessary loans* introduces a completely new concept, and *Luxury loans* have the same meaning as a native word that has already existed in a recipient language. For example, *Computer* represents an example of necessary loans, and *people* in French represent an example of luxury loans: *people* means *famous persons* like *celebrities* in French. Onysko and Winter-Froemel (2011) defined new terminologies, *Catachrestic innovations* and *Non-catachrestic innovations*, for loanwords distinction from the perspective of whether a word expresses a completely new concept that has never existed in another word.

Recently, in particular, this social linguistics role (Social semantic function) of loanwords–*luxury loans*, and *non-catachrestic innovations*–have growing interest (Zenner et al., 2019). *Prestige* image mainly can trigger the social linguistic use of loanwords. Zenner et al. (2019) categorizes this prestige role into four in more detail: *Social meaning*, *Indexicality*, *(Social) Identity*, and *Language regard*.

Social meaning represents all social attributes of the language or the language speaker: Italian (Italian people) has an image of people having a warm family, thus some company uses Italian word or phrase in an advertisement. *Indexicality* represents a self-inference to clearly indicate the boundaries of the social circle: Female students who use Spanish *chica* at an English school. *Social identity* indicates the qualities belonging to a social group. *language regard* means a speaker's cultural knowledge and belief systems concerning the social meaning of the language variants and varieties in their repertoire (Preston, 2013).

The social semantic function of loanword has exited in Asian languages like Japanese and Korean. Stanlaw (1992) states that social situations, symbolic and cognitive things intricately affect the use of loanwords in Japanese. For example, the traditional Japanese cuisine and Japanese-style dining room use Japanese native word 米 *kome* "rice", whereas foreign cuisine and the western-style dining room uses ライス *raisu* "rice". In this way, English loanwords and Japanese native words can convey different feelings, connotations, and rhetorical styles even if they have a similar meaning. Stanlaw also stated that the meaning of loanwords differs among people, and the loanword ambiguity has created connotation in Japanese communication. In Japanese society, these social semantic functions greatly promote the use of loanwords in advertising, broadcasting, TV shows, newspapers, books, comics, magazines, and everyday conversation.

Haarmann (1984) emphasized the relationship between language and ethnicity (ethnic identity) for understanding the stereotypes and prestige of loanwords. Loveday (1986) states that young people use loanwords as markers of youth identity because loanwords much frequently appear in pop culture, mass media, and fashion trends of young people. Japanese people have a social internal desire for the image and standard of *sophistication* for English, and advertisements and mass media accelerate the desire. Loveday (1986) states copy-writers, journalists, media personnel, translators and academics mainly distribute the loanwords in Japanese society.

Rebuck (2002) also summarized several sociolinguistic roles of Japanese loanwords. First, loanwords can effectively bestow social attention to social issues. For example, セクシュルハラスメント *sekusharuharasumento* "sexual harassment", ドメスティックバイオレンス *domesuteikkubaiorensu* "domestic violence", and ストーカー *sutoka* "stalker" win over social attention to these traditional social problems by using loanwords. Second, an arising of social need and a changing of social attitude toward authority will trigger the loanword use. Third, loanwords also can convey a *scientific reliability* image, and the company frequently uses loanwords in advertisements such as drugs and medical products. Fourth, loanwords can internationalize the message and effectively present sympathy and resolution for international events. For example, the Japanese Prime Minister used many English loanwords and English phrases in his Japanese speech about the World Trade Center terrorist attack. Fifth, loanwords can express a trendy and modern image. For example, the titles of the latest popular pop music or the name of occupations frequently use loanwords: プロデューサー *purodeyusa* "producer" and ファッションモデル *fuasshommoderu* "fashion model". Sixth, loanwords can give a more expert and high-performance impression: コーチングスタッフ *kochingusutaffu* "coaching staff" and スイムスーツ *suimusutsu* "swimsuits". Finally, the loanword has a euphemistic function. Loanword can replace the shocking or upsetting Japanese expressions with more polite expressions: マイホーム *maihomu* "my home" and カードローン *kadoron* "card loan".

Based on these sociolinguistic functions of English loanwords and the excessive use of loanwords in the mass media, the topic or the context in which loanwords are used can change according to the occurrence of social problems and the change of cultural trend. Namely, loanwords can be an indicator of social and cultural trend changes. However, no previous research has tested this function of loanwords and no previous research detects and analyzes the cultural trend change through the context change of loanwords. This dissertation focuses on this social function in chapter four. This study investigates the relationship between the contextual change of loanwords and the cultural trend change. The result sheds light on the possibility of loanwords as an indicator of social and cultural trend changes.

1.3 Methodological Background

This section reviews the theoretical background of the core technology, **Word Embedding**, in the experiment of this dissertation.

1.3.1 The Vector Space Model

Many researchers have explored the technology that enables computers to understand natural language like humans. However, no matter how fast the computer processing speed is, the computer itself has no ability to understand the natural language like human beings. Thus, many researchers have developed the technology to extract semantic information from language data and convert it into a format that a computer can process. For processing natural language with a computer, it is necessary to represent linguistic elements, such as document, sentence, and word, with numeric values. Researchers have verified that converting these linguistic elements into a vector format is a useful method (Almeida and Xexéo, 2019). As a pioneering study, Salton et al. (1975) converted a Document into a t-dimensional vector based on terms contained in the Document. This model, *Vector Space Model*, has greatly improved the performance in document classification and information retrieval. This vectorization of linguistic elements is called *Embedding* in the Natural language processing (NLP) field.

	Word 1	Word 2	Word 3	Word 4
Document 1	0	0	11	12
Document 2	10	12	1	0
Document 3	12	15	2	1
Document 4	0	1	10	20

vectorization

$$d_1 = [0,0,11,12]$$
$$d_2 = [10,12,1,0]$$
$$d_3 = [12,15,2,1]$$
$$d_4 = [0,1,10,20]$$

Document Vectors

Figure 1.1: A simple example of document vectorization in Bag of words. d_i means vector of document i.

1.3.2 The Bag of Words Model

One of the simplest embedding models is *the Bag of words model*. The bag of words model vectorizes sentences based on the frequency of the component word without considering the word order. Figure1.1 shows a simple example of document vectorization in *Bag of words*.

1.3.3 Neural Network and Neural Probabilistic Language Model

Neural Network is a very popular method for calculating vectors from a large amount of linguistic data these days. The neural network mathematically imitates the neural cells and their connections in a human brain. Figure1.2 shows a simple model of neural network. Basically, an input layer, one or more hidden layers, and an output layer compose the neural network. In the human brain, neurons use electrical signals for information transmission. The amount of information depends on the connection

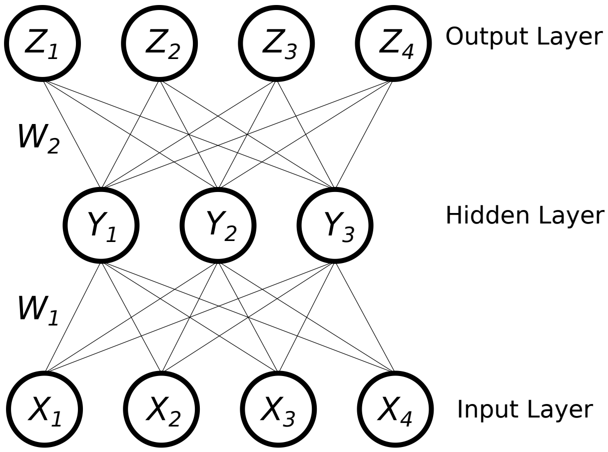

Figure 1.2: A simple model of a neural network.

strength between the neurons. In the neural network model, the weight W means the connection strength between artificial neurons from one layer to another layer. When a data set enters the input layer X, this network multiplies the data set X_n by the weight W_1 and sends the resulting value Y_n to the hidden layer. Next, this model multiplies Y_n by the value of the weight W_2, and send the resulting value Z to the output layer. The formulation of this calculation process in the neural network can be expressed as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_H \end{bmatrix} = \begin{bmatrix} (W_1)_{11} & (W_1)_{12} & \cdots & (W_1)_{1N} \\ (W_1)_{21} & (W_1)_{22} & \cdots & (W_1)_{2N} \\ \vdots & & \ddots & \vdots \\ (W_1)_{H1} & (W_1)_{H2} & \cdots & (W_1)_{HN} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \tag{1.1}$$

$$\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_D \end{bmatrix} = \begin{bmatrix} (W_2)_{11} & (W_2)_{12} & \cdots & (W_2)_{1H} \\ (W_2)_{21} & (W_2)_{22} & \cdots & (W_2)_{2H} \\ \vdots & & \ddots & \vdots \\ (W_2)_{D1} & (W_2)_{D2} & \cdots & (W_2)_{DH} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_H \end{bmatrix}. \tag{1.2}$$

In this equation, the value $X = (X_1,\ X_2,\ \cdots, X_N)$ denotes the input data, $Y = (Y_1,\ Y_2,\ \cdots, Y_H)$ denotes the values in hidden layer, $Z = (Z_1,\ Z_2,\ \cdots, Z_D)$ denotes the output data, N denotes the dimension of input data, H denotes the number of neurons in hidden layer, and D denotes the dimension of output data. The values $(W_1)_{ij}$ and $(W_2)_{jk}$ are weight parameters, and this model must train these weight parameters.

Neural network model compares the value of Z_n with the actual value in the data set called the answer. Referring to the gap between the value of Z_n and the answer value, the neural network tunes the weights W and undergoes the same process again. This process repeatedly continues and converges the weight W until minimizing the gap between the final output value Z_n and the answer value. Mostly, this approximation uses cross-entropy loss, which is calculated as

$$\mathcal{L}(X) = -\log \frac{e^{Z_{y_X}}}{\sum_j e^{Z_j}}. \tag{1.3}$$

where y_X gives the actual label of data X. If the neural model is trained towards the direction of decreasing the loss (1.3), weight parameter W_1 and W_2 increases Z_{y_X}, which makes the model can detect the label of data well.

Bengio et al. (2003) suggested *Neural Probabilistic Language Model* (NPLM) which trains a statistical language model to predict a target word from previous words by using neural network. This model learns to predict the tth (target) word with the word order from the $t-n$th (n is a range of word searching) to the $t-1$th words appearing before the tth word. The equation (1.4) represents the probability of predicting the tth word with the words appearing before tth word.

$$P(w_t|w_{t-1}, w_{t-n+1}) = \frac{\exp{(y_{w_t})}}{\sum_i \exp{(y_i)}} \tag{1.4}$$

In this equation, w_t is target word. w_{t-1} is the word appearing directly before w_t and w_{t-n+1} is the first word in a range of word searching. The y represents the final value calculated through the neural network.

Large language data trained the model toward increasing this probability P. As a result of training, the vectors of words used in similar contexts point in similar directions.

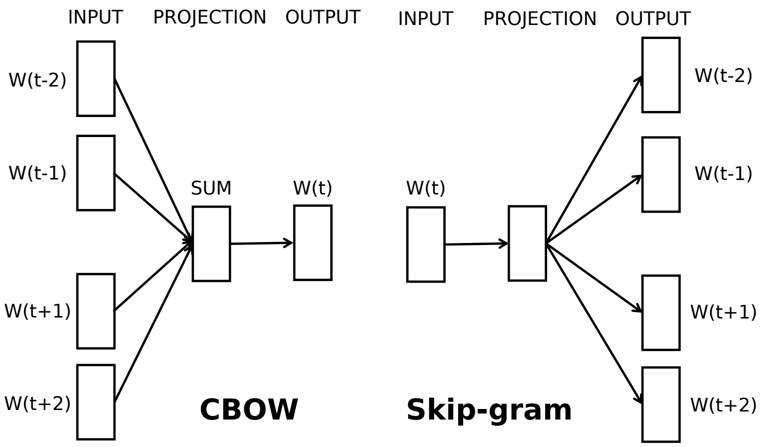

Figure 1.3: CBOW and Skip-gram (Mikolov et al., 2013a)

In this way, the contextual information was represented by a vector.

1.3.4 Distributional Model and Word2vec

Recently, many researchers use distributional models to calculate a vector value of a word meaning. Basically, the distributional models stand on the distributional hypothesis (Harris, 1954):

> "*. . . words that occur in the same contexts tend to have similar meaning.*" (Pantel, 2005)

Distributional models usually contain the high-dimensional vector space and represent a word as a spot on the surface of high-dimensional space.

Many computational linguistic researchers have developed several distributional models (Mikolov et al., 2013a; Turian et al., 2010; Mikolov et al., 2013b; Pennington et al., 2014; Mikolov et al., 2018). One of the most famous model is *word2vec*. Mikolov et al. (2013a,b) developed a new calculation method–*Continuous bag-of-words*(CBOW) and *Skip-gram*–and applied it to the neural network model. As a result, a word can be vectorized more effectively and more accurately. Figure1.3 shows the model of CBOW and Skip-gram (Mikolov et al., 2013a,b).

CBOW predicts the target word wt by using the surrounding words of the target word (context words) $(w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2})$ as input. Skip-gram predicts one of the context words $(w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2})$ by using the target word wt as input. In figure1.3, Skip-gram has four opportunities–$(w_t, w_{t+2}), (w_t, w_{t+1}), (w_t, w_{t-1}), (w_t, w_{t-2})$–to learn the probability for target word and context word, whereas CBOW has only one opportunity, $(w_t, (w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}))$. For the different numbers of learning opportunity between CBOW and Skip-gram in the same data, skip-gram is frequently used these days. In the training process, CBOW and Skip-gram continually adjust the value of the vector of words toward increasing the probability of target word in CBOW and context words in skip-gram. This model must learn to increase the probability of only one target word or only one context word to 1 and to decrease the probability of all other words in the database to 0. This learning process is inefficient, thus Mikolov et al. (2013a) proposes *positive sampling* and *negative sampling*. The positive sampling labels the words that actually appear around the target word in the database with $+$ (positive). The negative sampling labels the words that actually do not appear around the target word in the database with $-$ (negative). This technique changes the inefficient complicated learning process into a very simple binary decision learning process. The learning process of this model is calculated as

$$P(+|t,c) = \frac{1}{1+\exp\left(-u_t v_c\right)} \tag{1.5}$$

$$P(-|t,c) = 1 - P(+|t,c) = \frac{\exp(-u_t v_c)}{1+\exp\left(-u_t v_c\right)} \tag{1.6}$$

$$\mathcal{L}(\theta) = \log P(+|t_p, c_p) + \sum_{i=1}^{k} \log P(-|t_{ni}, c_{ni}) \tag{1.7}$$

where t is target word and c is context word. t_p and c_p are pair in the positive sampling and t_{ni} and c_{ni} are pair in the negative sampling. If c actually appears around t, the probability(1.5) must approach to 1. If c actually does not appear around t, the probability(1.6) must approach to 1. The object function(1.7)(θ is a model parameter) combines the (1.5) and the (1.6). The model proceeds with learning while renewing the vector value and parameters so that the model maximizes this object function. As a result of this learning process, the word vectors reflect the context information.

Word2vec has several advantages. First, word2vec can rapidly process a large amount of data. Second, word2vec requires no labeled data. Most machine learning methods require a data set labeled by humans. Preparing labeled data needs much time, much money, and much labor. Word2vec can extract contextual information from a large amount of non-labeled sentences. This unsupervised feature of word2vec contributes to lower the cost of computational semantic research

Finally, word2vec can calculate the differences in meaning between words. Word2vec represent the meaning of the word as a vector, thus we can calculate the difference of meanings using a simple vector calculation(1.8). Cosine similarity equals the cosine value of the angle between two-word vectors. The angle of two words having similar meanings tends to be nearly 0 and the cosine similarity becomes nearly 1. Conversely, the angle of two words having dissimilar meanings tends to be large and the cosine similarity becomes nearly -1. Figure1.4 shows the image of cosine similarity in vector space.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}, \tag{1.8}$$

Additionally, word2vec has an interesting feature of adding and subtracting the word meanings. The vector value reflects the relations of linguistic meanings as a linear translation in a vector space (Mikolov et al., 2013b). This allows for the addition and subtraction of meaning: $vec(\textit{Tokyo}) - vec(\textit{Japan}) + vec(\textit{Korea}) \approx vec(\textit{Seoul})$. These linguistic features have advanced using a neural network model in broad areas of computational linguistics. Because of these advantages, this dissertation also uses word2vec.

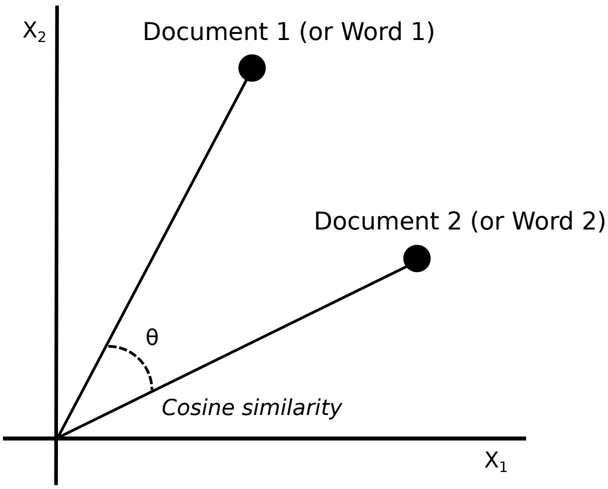

Figure 1.4: The image of cosine similarity in vector space model.

1.3.5 The Contextual Word Embedding and BERT

Word embeddings such as word2vec have provided high performance for various tasks in NLP. However, this word embedding method has a critical problem: this model only can assign one fixed vector value to a word. Namely, this model assigns one fixed vector to a polysemous word which meaning can change depending on the context. For example, the following two example sentences,

1. He withdrew money from his *bank* account.
2. He sat on the *bank* of the river.

the meaning of *bank* is different. The *bank* in the first sentence means a financial institute, whereas the *bank* in the second sentence means a raised area. The traditional word embedding assigns both *bank* the same vector value. The word embedding that assigns one fixed vector value to one word regardless of changes in context is called *Static Word Embeddings*.

A recent study has developed a new language model outputting the vector value of the target word according to the context of the input sentence. This model is called *Contextualized word embeddings* or *Dynamic embeddings*. Typical models include

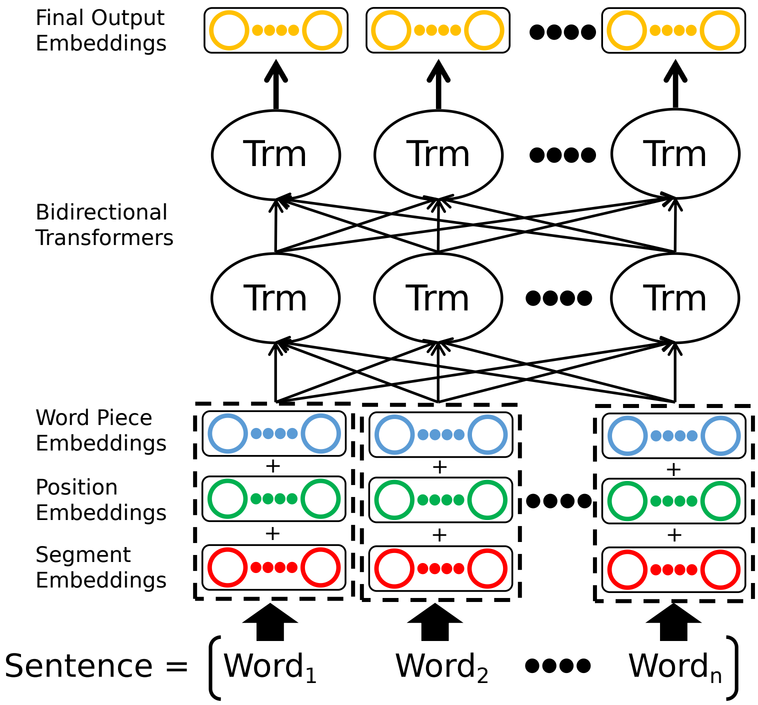

Figure 1.5: The Bidirectional Encoder Representations from Transformers (BERT)

ELMo (Peters et al., 2018), GPT (Radford et al., 2018), and BERT (Devlin et al., 2018). This dissertation uses the BERT model, which is one of the most popular models. Figure1.5 shows the basic mechanism of BERT.

BERT is an abbreviation for *Bidirectional Encoder Representations from Transformers*. Traditional language models have used the only one-way (left-to-right) sentence sequence information for training. BERT can use the bidirectional sentence sequence information for training. As a result, BERT has advanced the state-of-the-art on a variety of NLP tasks, such as sentence-level sentiment analysis.

For learning bidirectional word sequential information, BERT used two training methods. One is *Masked Language Model* (MLM). This method was inspired by the problem of filling in the blanks. The training process randomly removes some words from the training sentences. BERT uses these blanked sentences as training data. BERT tries to predict the appropriate word in the blank from the words sequence information before and after the blank. This training process continues to maximize the probability that BERT answers the correct word. Through this learning process, BERT can successfully acquire the bidirectional word sequence information from training sentences.

Another task is *Next sentence prediction*. BERT tries to predict the next sentence from one sentence and continuously maximize the probability of choosing the correct next sentence. Through these two tasks, BERT efficiently learns the language information.

A BERT model trained with a large amount of linguistic data is called a pre-trained model. Traditional models have required a large amount of data and lengthy training because traditional models must be trained from the beginning. However, pre-trained BERT trained with suitable data for the task to solve can achieve high performance in several language tasks even with a small amount of data and short training time. This training method is called *fine-tuning*. Fine-tuning is a great advantage of BERT and is the reason why BERT is widely used in a variety of language tasks.

Figure 1.5 briefly describes the calculation process of BERT. First, the input sentence becomes the final input vector through the process of adding the three embedding information: word piece embedding, position embedding, and segment embedding. Second, the bidirectional transformers layer calculate the bidirectional word sequence information. Finally, BERT outputs the vector value reflecting the bidirectional contextual information.

1.4 Summary of this Chapter

This chapter overviews the history and challenges of loanword research and the development of the word embedding language models. The following chapters will introduce each detailed researches applying the word embedding to the linguistic phenomena of loanwords explained in this chapter over three chapters: *Lexical competition*, *Semantic* *adaptation*, and *Social semantic function and the cultural trend change*.

In the following chapters, this dissertation conducted each experiment using word2vec or BERT. As explained above, these language models learn the meanings based on the Distribution theory (Harris, 1954):

> "... *words that occur in the same contexts tend to have similar meaning.*" (Pantel, 2005)

Therefore, in the following chapters, the similar or different *meaning* represents the word relation sharing the same context or not.

This dissertation targets only loanwords directly from English or loanwords introduced through English: English loanwords usually written in *katakana* letters in Japanese.[1] Most of the words written in *kanji* letters (*Chinese letters*) have entered from ancient Chinese, but this dissertation only covers English loanwords.

[1]Strictly speaking, some *katakana* words may come from other languages, but in this dissertation, if the word is used in English, we regard it as English loanwords.

Chapter 2

Word Embeddings for Lexical Changes Caused by Lexical Competition between Loanwords and Native Words

2.1 Overview

From ancient times, foreign words have flowed in alongside various cultures and cultural products from abroad. A loanword is a foreign word that is used without being translated into the recipient language. Loanwords have been extensively studied by historical linguists who study linguistic change because they cause changes in the linguistic system of the recipient language (Sapir, 1921; Bloomfield, 1933). Especially in recent years, with the development of the Internet and inexpensive system of global travel, the problems of excessive loanword overflows and of language changes for the worse attributed to loanwords have become social issues. Therefore, the importance of research on loanwords and language changes is increasing.

Despite the long history of loanword research in the field of linguistics, the majority of loanword research has centered around merely introducing myriad examples of loanwords and summarizing them systematically. Therefore, no research has been done on how loanwords settle in the recipient language and cause a change with the native language. This study refers to this phenomenon as *Lexical Competition*.

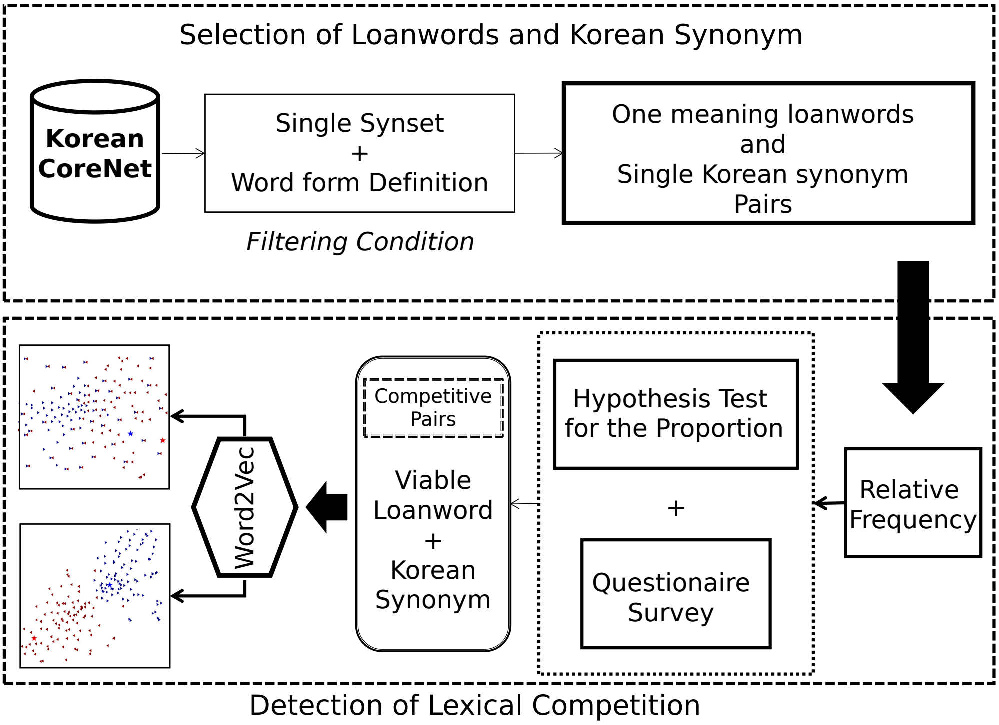

Figure 2.1: The experimental procedure of detecting lexical competition between loanword and Korean synonym

With the recent development of big data and the machine learning models, it has been possible to research the dynamic aspect of semantic change (Tahmasebi et al., 2018). These models, especially the word embedding model, could not only reveal the semantic changes of words in detail but also provide experimental evidence for the linguistic laws that have been under discussion for a long time (Xu and Kemp, 2015). Additionally, these models have clarified the changes in society that accompany changes in words (Garg et al., 2018).

Motivated by these technological developments in machine learning and its contribution to revealing semantic change, this study proposes the model of Lexical Competition caused by loanwords and explain the real situation of lexical competition using the word embedding model. The contributions of this research are as follows:

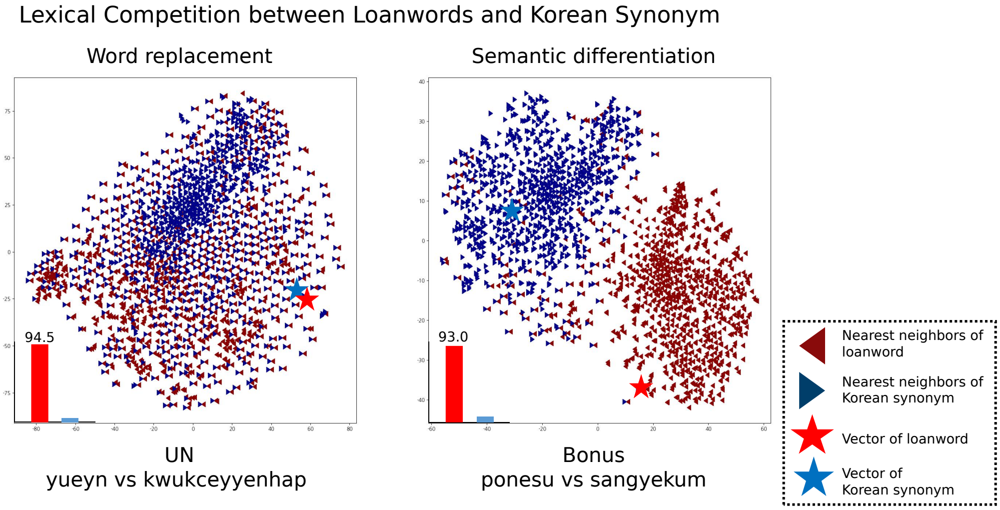

Figure 2.2: Examples "UN" and "Bonus" of Lexical Competition in Korean language represented by the word embedding model. Loanword, synonym, and their nearest neighbors (n=1000) vectors are spotted by using t-SNE algorithm. The two colors distinguish loanword and synonym nearest neighbors. The bar chart means the relative frequency of loanword (red) and the native synonym(blue).

1. This study constructs a method to find viable loanwords which settle successfully in a recipient language by combining a relative frequency, a statistical test, and a human rating test. This method enables us to detect viable loanwords that correspond to human intuition.

2. The word embedding model reveals details of the lexical competition between loanwords and native synonyms. In particular, this study shows the word embedding model is also a powerful model for finding and observing two phenomena: word replacement and semantic differentiation.

2.2 Related Works

This section summarizes the previous studies on loanwords and on word embedding model-based semantic changes. It also introduces the theoretical background of this study.

2.2.1 Lexical Competition in Loanword

As mentioned, most previous loanword research has only systematically summarized several interesting loanword examples in each language (Haugen, 1950; Hoffer, 2002). However, recently, research on loanword use and the pragmatic meaning of loanwords has advanced (Peterson and Beers Fägersten, 2018; Zenner et al., 2019; Onysko and Winter-Froemel, 2011).

There are two major cases where loanwords are used. First, when there are no native words to express new foreign concepts and products in the recipient language. Another case is when a loanword is used even though the recipient language already has a native word for the same concept. An example is the English word *restaurant*. Although there are already native words such as 식당 *siktang* "restaurant" in Korean and 食堂 *shokudo* "restaurant" in Japanese as a word corresponding to English *restaurant* "restaurant" is frequently used as a loanword in both Korean and Japanese. This second motivation for using a loanword is not about filling in the gap of concepts that are not in the recipient language, but also expressing a sense of high class, social identity and prestigiousness (Winter-Froemel, 2017). In particular, regarding this second pragmatic or social use of loanwords, some researchers began to pay attention to what kind of loanwords successfully settle in the recipient language even if there are already extant native synonyms (Winter-Froemel et al., 2014; Zenner et al., 2012; Shin, 2010).

This loanword settlement causes language to change. The language economy principle is one of the fundamental linguistic principles that explain this change (Martinet, 1955). According to this principle, the existence of two or more words that represent the same concept is economically inefficient because people must remember and understand more words (Bolinger, 1977). To solve this inefficiency, there are two possible changes:

- A loanword is replaced with the native word
- Loanword and native synonyms come to dominate different semantic fields from each other

This study calls the first change pattern **Word replacement** and the second **Semantic differentiation**.

Previous studies used the relative frequency between loanwords and synonyms as the index of how much the loanwords are winning on the lexical competition in the recipient language (Winter-Froemel et al., 2014; Zenner et al., 2012). Relative frequency is calculated by dividing the frequency of loanwords by the total frequency of loanwords and synonyms. This relative frequency represents how often loanwords were used compared to the recipient language synonyms and whether a loanword that survives successfully in the recipient language (Calude et al., 2017). The equation of the relative frequency is below.

$$Freq_{relative}(\%) = \frac{Freq_{loanword}}{Freq_{loanword} + Freq_{synonym}} \times 100 \tag{2.1}$$

where $Freq_{relative}$ is the relative frequency of the loanword, $Freq_{loanword}$ is the frequency of the loanword, $Freq_{synonym}$ is the frequency of the native synonym in a corpus.

Although this method works as an indicator of how well loanwords establishing in the recipient language, there is a technical limitation as an indicator of the lexical competitions. The limitation is that the relative frequency alone can not determine whether the loanword-synonymous relationship is a word replacement or a semantic differentiation. The difference between word replacement and semantic differentiation depends on whether the loanword and the synonym share the same context or not.

Relative frequency cannot reveal the sharing of context. Therefore, the relative frequency has a limit in accurately capturing the lexical competition.

Solving this limitation, this study uses word embedding to provide contextual information on loanwords and synonyms. Contextual information can determine whether loanwords and native synonyms are in word replacement or semantic differentiation. Next section explains the detailed methods of this improvement. The rest part of this section briefly overviews previous studies in which word embedding has been applied to the linguistic study.

2.2.2 Word Embedding Model and Semantic Change

The word embedding model has been developed and used as a technology for natural language processing since early on, but with the advent of word2vec using the Skip-gram model (Mikolov et al., 2013a), it is possible to vectorize words more accurately. After that, various models such as fastText (Bojanowski et al., 2017) and GloVe (Pennington et al., 2014) were developed. Especially, recently, many contextualized word embedding models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) have been developed, which has brought significant progress to natural language processing.

This word embedding model has been used not only for basic tasks such as sentiment analysis (Barnes et al., 2018) and document classification but also for linguistic tasks such as diachronic semantic change over time. Using this word embedding model, Xu and Kemp (2015) conducted a data-based experiment on two laws related to language change, that is, the law of differentiation and the law of parallel change, and verified them. Hamilton et al. (2016a) tested some models and examined statistical laws that relate frequency and polysemy to semantic change. Hamilton et al. (2016b) investigated the cultural factors associated with semantic changes, and Garg et al. (2018) contributed to sociolinguistics by investigating the semantic changes associated with gender. Recently, Hu et al. (2019) succeeded in capturing more detailed meaning changes by using BERT. In this way, it can be said that the word embedding model captures the meaning of natural language well and has contributed to great progress in linguistic research.

These previous researches provide the insight that this word embedding model can also quantitatively examine lexical competition and contribute to explaining the phenomenon more clearly. Therefore, this study applies the word embedding model to investigate the lexical competition that accompanies the influx of loanwords.

2.3 Selection of Loanword and Korean Synonym Pairs

Our experimental procedure is summarized in Figure 2.1. The experimental procedure is divided into two: *Selection of Loanwords and Korean Synonym* and *Detection of Lexical Competition*. The following sections will discuss these two parts separately. First, this section describes the procedure of a loanword and Korean synonym pairs selection.

2.3.1 Viable Loanwords

As mentioned in the introduction, the lexical competition that occurs between loanwords and the native synonym when loanwords enter the recipient language (Korean in this study) has possibly two types: *word replacement* and *semantic differentiation*. It is thought that this lexical competition will occur when loanwords are well established in the recipient language and are actively living: *Viable loanwords* in this study. This study purposes to observe the pattern of lexical competition of loanwords and this purpose must require the selection of viable loanwords. The next two subsections describe the previous approach and our new approach on how to select viable loanwords.

2.3.2 Previous Approach: The Relative Frequency

As explained in Section 2.2, previous studies used relative frequencies to study loanwords that have successfully entered the recipient language. However, the problem of this relative frequency index is the difficulty of giving an absolute threshold of a viable loanword. In other words, it is not possible to say at what percentage a loanword is viable. Therefore, it is difficult to select viable loanwords that are in the lexical competition only based on this relative frequency information. The next subsection explains our new approach to selecting viable loanwords.

2.3.3 New Approach: The Proportion Test

To overcome this weak point of the relative frequency method, this study used the hypothesis test for proportions (the one-proportion z-test) to detect viable loanwords. The biggest advantage of the Proportional test is the absolute verification that the relative frequency of loanwords p_l is statistically higher than that of the native synonym p_s. Specifically, we calculate

$$z = \frac{\hat{p}_l - 0.5}{\sqrt{\frac{\hat{p}_l(1-\hat{p}_l)}{n}}}$$

to guarantee $p_l > p_s$ in proportional test, where $\hat{p}_l$ is a relative frequency obtained from the experiment, n denotes the total frequency of the loanword and the native synonym. The p-value result is obtained by the fact that z has a standard normal distribution $\mathcal{N}(0, 1)$ approximately. If the frequency of the loanword is statistically higher than the native synonym, it can be said that such loanword is well-adapted and has successfully survived the challenge of the corresponding the native synonym.

2.3.4 Technical Challenges for Performing the Proportion Test

This proportional test uses the frequency of loanwords and the frequency of the native synonym. Thus, we must calculate the frequency of the loanword and the native synonym, but three challenges will happen in this process.

The first challenge is the polysemous loanwords. If a loanword is polysemous, it is almost impossible to find all native synonyms for the loanword: all competitors of the loanword. Even if it might be possible to find all native synonyms for each meaning of a polysemous loanword, we will meet the difficulty in the process of calculating the frequency of the native synonyms. For performing the proportional test between the loanword and the native synonyms, we must sum up the frequencies of all the synonyms and compare that with the frequency of a loanword. In most cases, the sum of the frequency of all Korean synonyms will win the frequency of loanwords. This would indicate the competitive relationship between a loanword and a synonym group rather than the competitive relationship between a loanword and a synonymous word. This situation does not correctly represent the lexical competition between a loanword and the native synonym, which is the purpose of this study. For this reason, this study must target loanwords with a single meaning.

The second challenge is to select only loanwords that can compete with Korean synonyms. If no word in Korean expresses the meaning of the loanword, the loanword can not form a competitive relationship with the Korean word. Therefore, to investigate the competitive relationship between loanwords and synonyms, it is necessary to remove these loanwords and select loanwords that have synonyms in Korean.

The third challenge is single meaning loanwords having multiple native synonyms. As with the polysemous loanword mentioned earlier, these loanwords cause problems in the process of calculating the frequency of the native synonym. In order to accurately compare the frequency of the loanword and the native synonym, it is necessary to compare the total frequency of the native synonyms with the frequency of the loanword. In most cases, as the sum of the frequency of the native synonyms will be larger than the frequency of a loanword, comparing the frequencies will be difficult, and finding viable loanwords will be mostly impossible. To handle these problems and observe an accurate lexical competition, this study focused on only the loanwords having not only a single meaning but also single Korean synonyms.

2.3.5 Filtering Procedures

To overcome these challenges and create a list of viable loanwords-Korean synonyms pairs, this study used Korean CoreNet (Choi et al., 2004)[1]. Korean CoreNet is a Korean word database that classifies each word by a synset like WordNet (Miller, 1998). This study regards the number of this synset as the number of meanings of the word.

First, to overcome the challenge of polysemous loanwords, we used the number of synsets and the number of definitions of loanwords in the dictionary built-in CoreNet. We assumed the number of synsets and definitions are the number of meanings of the loanword. With this assumption, we regarded the loanword that has only one synset and only one definition as the single meaning loanword. For retrieving the loanwords as much as possible from CoreNet, we used 외래어 표기 용례 정보[2] *oylaye phyoki yonglyey cengpo* "Loanword notation usage information" distributed by the National Institute of Korean Language as a loanword list. This loanword list contains 26965 general loanword terminologies. We passed the loanword list through the one synset filtering process and one definition filtering of CoreNet and we finally got 554 loanwords. These filtering process removed polysemous loanwords and resolve the problem of polysemous loanwords. This 554 loanword list is in Appendix.

Second, to overcome the challenge of selecting loanwords that actually compete with Korean synonym, we used the definition type of the dictionary built-in CoreNet. This dictionary has two types of definitions: the sentence form definition and the word form definition. After observing the definition in detail, we assumed that the reason for explaining the meaning of a loanword in a sentence form definition is that no Korean synonym corresponding to the loanword exists. Based on this assumption of definition type, this study used these definition types as a filtering standard for removing loanwords having no competitive Korean synonyms and overcame the challenge of

[1] `http://semanticweb.kaist.ac.kr/org/bora/CoreNet_Project/`

[2] `https://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=208&etc_seq=636`

selecting loanwords that actually compete with the synonym. We call this filtering process *Definition type filtering* in this study.

Finally, to overcome the challenge of the single meaning loanwords having multiple native synonyms, we used the number of words in the word form definition. Namely, if the word form definition has only one word, we assumed the loanword has only one competitive Korean synonym. We call this filtering process *Definition word number filtering*.

Additionally, as the next section will explain deeply, this study analyzes the relationship between loanwords and Korean synonyms using the Word Embedding Model. Therefore, the word embedding model must have the vector values of loanwords and the Korean synonym. Unregistered loanwords and Korean synonyms must be removed from the loanword-Korean synonym pairs. After we passed the list of loanwords having one synset and one definition through *Definition type filtering* and *Definition word number filtering*, we got 95 loanword-Korean synonym pairs.

2.3.6 Handling Errors

Additionally, we removed incorrect loanword-synonym pairs from the result with referencing Standard Korean Dictionary[3] and got 63 loanword-Korean synonym pairs. In the process of final checking the last candidate pairs for the proportional test, we found some errors related to Wikipedia data. This experiment used Wikipedia Data as training data for the Word Embedding Model. Therefore, we found that the loanword and the Korean synonym itself or the homonym is used in Wikipedia for a proper nouns such as place names, person names, group names, and movie titles: for example, 본드 *pontu* "bond" is used in person name like 제임스 본드 *Ceyimsu Pontu* "James Bond" and 키스 *khisu* "kiss" is used in person name like 키스 미첼 *khisu micheyl* "Keith Claudius Mitchell" and a Korean musical group name 키스 *khisu* "kiss" in Wikipedia article. We judged these noise data to give influence on the frequency and the vector

[3] `https://ko.dict.naver.com/#/main`

	Selection Procedures of Loanword-Korean synonym pairs	Word number
1.	One synset and one definition loanwords in CoreNet	554
⇓	*Filtering with Definition type and Word number in definition*	
2.	Loanword-Korean synonym pairs in Word2Vec	95
⇓	*Handling errors*	
3.	Loanword-synonym pairs for proportional test	33
⇓	***Proportional Test***	
4.	Final viable loanword-Korean synonym pairs	**14**

Table 2.1: Summary of selection procedures of loanword-Korean synonym pairs.

value of loanwords and the Korean synonyms. To eliminate this error, we removed a loanword-Korean synonym pair from the experimental target if a loanword or Korean synonym is used for a proper noun regardless of its original meaning and an independent article about the proper noun is registered on Wikipedia. Finally, we got 33 loanword pairs for the proportional test.

2.3.7 Proportion Test and Questionnaire Survey

This study used the hypothesis test for proportions to detect viable loanwords. By performing the proportion test between the relative frequency of the 33 loanword-Korean synonym pairs obtained above, this study tries to find which loanwords have a relative frequency greater than the Korean native synonym. If the relative frequency of the loanword is statistically higher than the synonym, it can be said that such loanword is well-adapted and has successfully survived the challenge of the corresponding Korean synonym. With the p-value value obtained through the proportional test, we obtained 14 loanword-Korean synonym pairs having a higher relative frequency of loanwords with 95% confidence. Table 2.2 shows the result of proportional test.

Additionally, a questionnaire survey was conducted to confirm how well the detected

Loanword:Synonym	p-value	Loanword:Synonym	p-value
meympe:kwusengwen	<0.001	paksu:sangca	0.997
yueyn:kwukceyyenhap	<0.001	thawel:swuken	1.000
ponesu:sangyekum	<0.001	phapsong:taycwungkayo	1.000
khonsethu:umakhoy	<0.001	lwul:kyuchik	1.000
suphochu:wuntongkyengki	<0.001	theyma:cwucey	1.000
khemphyuthe:cencakyeysanki	<0.001	mithing:moim	1.000
yuniphom:ceypok	<0.001	phokheys:cwumeni	1.000
phulinthe:inswayki	<0.001	sukhophu:pemwi	1.000
khomiti:huykuk	<0.001	eythikheys:yeyuy	1.000
khomitien:huykukpaywu	<0.001	sutholi:iyaki	1.000
okheysuthula:kwanhyenaktan	<0.001	suthayntetu:phyocwun	1.000
lawunci:hyukeysil	<0.001	kulangphuli:taysang	1.000
yuthophia:isanghyang	<0.001	paktheylia:seykyun	1.000
sulloken:phyoe	<0.001	phaysuphothu:yekwen	1.000
theynthu:chenmak	0.163	simphociem:tholonhoy	1.000
pheysuthu:huksapyeng	0.368	phulaipesi:sasaynghwal	1.000
insuthenthu:cuksek	0.604		

Table 2.2: The result of proportion test for loanword-Korean synonym pairs.

viable loanwords matched the intuition of language-speaking people in reality. In the questionnaire survey, each survey subject was asked to express his or her preference to the list of pairs of loanwords and synonyms (mentioned above section) using 5-point Likert scale. (1: loanword is used much more than the Korean synonym; 3: loanwords and synonyms are similarly used; 5: Korean synonym is used much more than the loanword; 2 and 4: intermediate between 1 and 3 and between 3 and 5, respectively). 58 Korean native speakers answered the survey in total.

Analyzing the average scores of each loanword from the survey finds that the survey subjects also judged the selected loanwords as highly used loanwords. This result implies that our finding is well-matched with the intuition of language-speaking people in reality. These 14 pairs strictly screened through these procedures are considered to be a pair of viable loanwords causing lexical competition and Korean synonym in competition. This study focused on these 14 pairs to analyze the lexical competition more accurately. Table 2.1 summarizes the selection procedures of loanword-Korean synonym pairs. Next section will show the result of analyzing the lexical competition of these 14 loanword-Korean synonym pairs.

2.4 Analysis of Lexical Competition

The previous section showed how to select a viable loanword that is more likely to cause lexical competition. This section describes our model analyzing the lexical competition and the setup procedure of the model. To analyze what type of lexical competition that viable loanwords have undergone, this study focused on the usage context sharing condition between loanwords and Korean synonyms. As mentioned in the introduction, lexical competition mainly has two main types: word replacement and semantic differentiation. Because word replacement means that a viable loanword has won over and replaced with the Korean synonym, it can be implied that the loanword has a similar semantic field as the Korean synonym. Whereas, because semantic differentiation means that loanwords and Korean synonyms have differences in meaning, it can be implied that the viable loanword has a different semantic field from the Korean synonym. This study assumed this semantic field as the context where loanwords and Korean synonyms are used (the usage context). This assumption can allow describing that loanwords and Korean synonyms share the same usage context in word replacement and share the different usage context in semantic differentiation. To model this usage context quantitatively, we suggest using the vector space of Word Embedding Model. Additionally, we propose the geometrical model to represent the sharing condition of the usage context in vector space. The next section describes this geometrical model.

2.4.1 The Geometrical Model for Analyzing the Lexical Competition

As mentioned above, judging whether the lexical competition is the word replacement or the semantic differentiation requires the usage context sharing condition, namely the degree of intersection of the usage context between the loanword and the Korean synonym. The large intersection of the usage context will represent the word replacement because the loanword and the synonym compete in the same usage context and the loanword wins at that usage context. While the small intersection will represent the word differentiation because the loanword and the native word live in another usage context. To investigate these contextual relations quantitatively, this study applies the simple mathematical concept of geometry: Overlapping circles model like in Figure 2.3. In Figure 2.3, the two circles are a diagram of the usage context vector space of loanword and Korean synonym. In fact, the vector space of the Word Embedding Model used in this experiment is 200 dimensions and it is impossible to explain in this 2D figure. Therefore, it is suitable to understand that Figure 2.3 shows the vector space sharing condition existing on the surface of a 200-dimensional sphere. This figure shows that the more to the left, the sharing condition of two vector spaces will become smaller, and the more to the right, the sharing condition of two vector spaces will become larger. From a lexical competition perspective, this figure shows that the relationship between

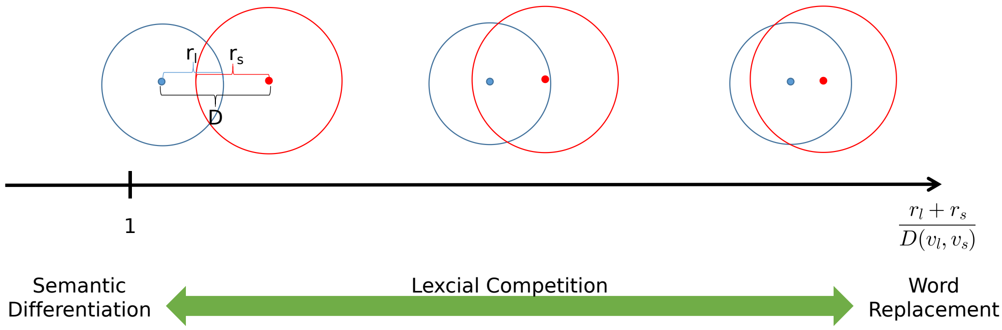

Figure 2.3: The geometrical model for the context relation between the loanword and the native synonym in this study.This figure represents a set of vectors originally on the surface of a sphere.

loanword and Korean synonym is closer to Semantic differentiation as it goes to the left and closer to Word replacement as it goes to the right.

In the lexical competition study, the center of the circle is the vector of the target word (a loanword or a native synonym in this study) and the radius of the circle (r) is defined as the average distance between the vector of the target word and the vectors of the nearest neighbors(one thousand nearest neighbors in this study). The distance between the centers of the circle (D) is defined as the distance between the vectors of target words. All distance is calculated in radian measure. The calculation of the radius and the distance between the centers is given by the equation

$$r_t = \frac{1}{m}\sum_{i=1}^{m} \arccos \frac{v_t \cdot v_{n_i}}{\|v_t\| \cdot \|v_{n_i}\|} \tag{2.2}$$

$$D(v_l, v_s) = \arccos \frac{v_l \cdot v_s}{\|v_l\| \cdot \|v_s\|} \tag{2.3}$$

$$\mathcal{P} = \frac{r_l + r_s}{D(v_l, v_s)} \tag{2.4}$$

where r_t is the radius of the context vector space (the average distance between the target and the nearest neighbors) of target word, v_t is the vector of target word, and v_{n_i} is the vector of the ith nearest neighbors. The v_l is the vector of the loanword, v_s is the vector of the synonym, and $D(v_l, v_s)$ is the distance between the loanword vector and the synonym vector: the distance between the centers of the circle in Figure 2.3. We set the number of nearest neighbors $m = 50, 100, 250, 500, 1000, 10000$ and compared each results in this experiment. Table 2.3 and Table 2.4 summarize the change in the result with changing the value of m. The change in the result with changing the value of m indicates that the ranking of pairs does not largely change and especially the upper and lower pairs are constantly stable.

From this mathematical basis, this experiment calculated the ration ($\mathcal{P}$) between the sum of the loanword radius r_l and the synonym radius r_s, and the distance between the vector of the loanword and the synonym: the ration ($\mathcal{P}$) between r_l+r_s and $D_(v_l, v_s)$. Table 2.3 and Table 2.4 summarizes the result of this mathematical context information in the loanword-synonym pairs. For a visual understanding of the difference of the lexical competition, the figure of a vector projected in 2D-dimensions using t-SNE (Maaten and Hinton, 2008) is displayed for several loanword-native synonym pairs. The following section analyzes two lexical competitive relationships related to loanwords, namely word replacement and semantic differentiation, from the viewpoint of the word embedding model.

m=50	$\mathcal{P}$	m=100	$\mathcal{P}$	m=250	$\mathcal{P}$
yueyn:kwukceyyenhap	3.04	yueyn:kwukceyyenhap	3.20	yueyn:kwukceyyenhap	3.42
okheysuthula:kwanhyenaktan	2.61	okheysuthula:kwanhyenaktan	2.80	okheysuthula:kwanhyenaktan	3.05
sulloken:phyoe	2.44	sulloken:phyoe	2.51	sulloken:phyoe	2.59
yuthophia:isanghyang	2.25	khonsethu:umakhoy	2.33	khonsethu:umakhoy	2.50
khonsethu:umakhoy	2.19	yuthophia:isanghyang	2.32	yuthophia:isanghyang	2.42
lawunci:hyukeysil	2.14	lawunci:hyukeysil	2.24	lawunci:hyukeysil	2.38
suphochu:wuntongkyengki	2.12	suphochu:wuntongkyengki	2.21	suphochu:wuntongkyengki	2.34
yuniphom:ceypok	2.04	yuniphom:ceypok	2.14	yuniphom:ceypok	2.30
khemphyuthe:cencakyeysanki	1.95	khomitien:huykukpaywu	2.02	khomitien:huykukpaywu	2.12
khomitien:huykukpaywu	1.94	khemphyuthe:cencakyeysanki	2.02	phulinthe:inswayki	2.12
phulinthe:inswayki	1.94	phulinthe:inswayki	2.01	khemphyuthe:cencakyeysanki	2.11
meympe:kwusengwen	1.90	meympe:kwusengwen	1.98	meympe:kwusengwen	2.09
ponesu:sangyekum	1.68	khomiti:huykuk	1.75	khomiti:huykuk	1.85
khomiti:huykuk	1.67	ponesu:sangyekum	1.75	ponesu:sangyekum	1.85

Table 2.3: The result of the mathematical analysis of the context sharing condition between the loanword and Korean synonym. This table shows the result of m = 50, 100, 200.

m=500	$\mathcal{P}$	m=1000	$\mathcal{P}$	m=10000	$\mathcal{P}$
yueyn:kwukceyyenhap	3.59	yueyn:kwukceyyenhap	3.75	yueyn:kwukceyyenhap	4.29
okheysuthula:kwanhyenaktan	3.25	okheysuthula:kwanhyenaktan	3.45	okheysuthula:kwanhyenaktan	3.97
sulloken:phyoe	2.65	khonsethu:umakhoy	2.75	khonsethu:umakhoy	3.19
khonsethu:umakhoy	2.63	sulloken:phyoe	2.72	sulloken:phyoe	2.98
yuthophia:isanghyang	2.50	yuthophia:isanghyang	2.60	yuthophia:isanghyang	2.92
lawunci:hyukeysil	2.48	lawunci:hyukeysil	2.58	lawunci:hyukeysil	2.91
suphochu:wuntongkyengki	2.43	yuniphom:ceypok	2.53	yuniphom:ceypok	2.90
yuniphom:ceypok	2.42	suphochu:wuntongkyengki	2.53	suphochu:wuntongkyengki	2.82
khomitien:huykukpaywu	2.20	phulinthe:inswayki	2.29	phulinthe:inswayki	2.66
phulinthe:inswayki	2.20	khomitien:huykukpaywu	2.29	khomitien:huykukpaywu	2.60
khemphyuthe:cencakyeysanki	2.19	khemphyuthe:cencakyeysanki	2.27	khemphyuthe:cencakyeysanki	2.59
meympe:kwusengwen	2.18	meympe:kwusengwen	2.26	meympe:kwusengwen	2.58
ponesu:sangyekum	1.94	ponesu:sangyekum	2.02	ponesu:sangyekum	2.30
khomiti:huykuk	1.93	khomiti:huykuk	2.01	khomiti:huykuk	2.29

Table 2.4: The result of the mathematical analysis of the context sharing condition between the loanword and Korean synonym. This table shows the result of m = 500, 1000, 10000.

2.4.2 Word Embedding Model for Analyzing Lexical Competition

This study used the word2vec model to observe the lexical competition. The reason is that this model has been widely selected in semantic research for a long time, thus it is thought that this model is a reliable language model that accurately represents the meaning of language and the lexical competition. The training process for this model is described here.

The data set used for training word2vec was obtained from a May 2017 Korean Wikipedia dump data[4]. The text data was extracted by a Wikipedia extractor[5] from each Wikipedia dump data set. The Korean Wikipedia data are 606 MB. We used the open-source Korean text tokenizer Twitter[6] for segmentation. These preprocessed data are used for training word2vec (dimensions = 200, min count = 20, window size = 15) in the Gensim Python package[7]. The following experiment used this model to reveal how lexical competition occurs between loanwords and Korean synonyms.

2.4.3 Result and Discussion

Table 2.3, Table 2.4, and Figure 2.4 shows the result of this experiment. We will discuss the loanword-synonym pairs that are judged as closer to the word replacement relation and pairs that are judged as closer to the semantic differentiation in the following part of this section.

Table 2.3 and Table 2.4 shows 유엔 *yueyn* "UN", 오케스트라 *okheysuthula* "orchestra" have the high proportion even if m changed. This indicates the loanword-synonym pairs share the large intersection. This implies the possibility that the loanword-synonym relation is closer to the word replacement. Table 2.5 shows the top five nearest neighbors of these pairs. Table 2.5 shows the loanword and the Korean synonym pairs share some same nearest neighbors. Figure 2.4 shows the 2-D projected vectors of the loanword-

[4]`https://dumps.wikimedia.org/kowiki/`

[5]`http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

[6]`https://github.com/twitter/twitter-korean-text`

[7]`https://radimrehurek.com/gensim/`

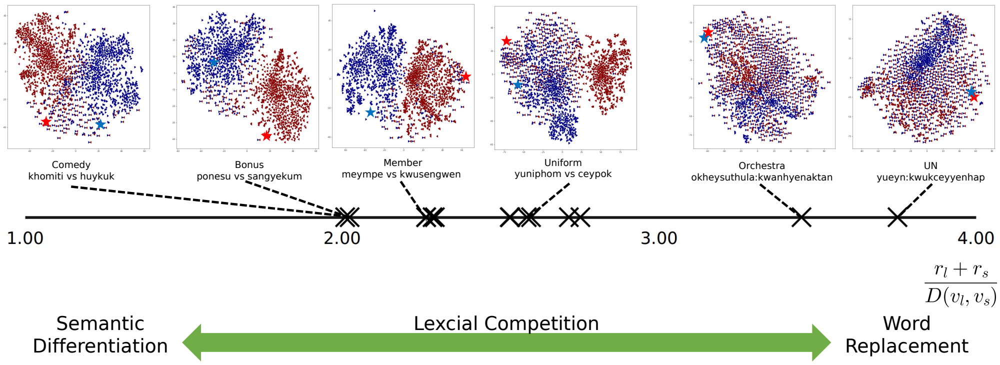

Figure 2.4: This figure shows loanword-pairs from right to left in descending order of the value $\mathcal{P}$ (m=1000). The 2-D projected vector spaces of the loanword-Korean synonym pairs–"UN", "Orchestra", "Uniform", "Member", "Bonus", and "Comedy"–are shown as an example.

synonym pairs and their nearest neighbors vectors. This figure helps to understand visually the context vector spaces share the large intersection.

English	Loanword : Korean	Proportion	The top 5 nearest neighbors
UN	yueyn : kwukceyyenhap	3.75	**L**:kwukceyyenhap, **UN**, anpoli, phyenghwayucikwun, ancenpocang
			S: **yueyn**, **UN**, WTO, Nations, ILO
Orchestra	okheysuthula : kwanhyenaktan	3.45	**L**:**simphoni**, **kwanhyenaktan**, yencwuhoy, hamoni, **kyohyangaktan**
			S:**kyohyangaktan**, aktan, **okheysuthula**, umakcey, **simphoni**
Bonus	ponesu:sangyekum	2.02	**L**:khompo,aitheym,kyenghemchi,peything,khweysuthu
			S:swutang,kupye, cikup,thoycikkum,cikuphanun
Comedy	khomiti:huykuk	2.01	**L**:lomaynsu,lomaynthik,sulille,yenghwa,tulama
			S:huykok,pikuk,secengsi,yenkuk,meyllotulama

Table 2.5: The nearest neighbors of the high $\mathcal{P}$ loanword-synonym pairs ("UN" and "Orchestra") and the low $\mathcal{P}$ loanword-synonym pairs ("Comedy" and "Bonus"(m=1000). **L** means nearest neighbors of loanword and **S** means nearest neighbors of synonym. The rightmost column shows the top five nearest neighbors. The same nearest neighbors and the target word itself in nearest neighbors are displayed in bold font.

Table 2.3 and Table 2.4 shows 코미디 *khomiti* "comedy" and 보너스 *ponesu* "bonus" have constantly the low proportion. This means the context vector spaces of the loanword-synonym pairs share the small intersection. This implies the possibility that the loanword-synonym relation is closer to the semantic differentiation. Table 2.5 shows the top five nearest neighbors of these pairs. Table 2.5 shows the loanword and the Korean synonym pairs does not share the same nearest neighbors. Figure 2.4 shows the 2-D projected vectors of the loanword-synonym pairs and their nearest neighbors vectors. This figure helps to understand visually the context vector spaces share the large intersection.

The nearest neighbors can explain what kind of semantic differentiation "comedy" and "bonus" are causing. In the case of 코미디 *khomiti* "comedy" and 희극 *huykuk* "comedy", while 코미디 *khomiti* "comedy" has nearest neighbors related to foreign art and culture, like 로맨틱 *lomaynsu* "romance", 로맨틱 *lomaynthik* "romantic" and 스릴러 *sulille* "thriller", the nearest neighbors of 희극 *huykuk* "comedy" are about traditional Korean art words such as 희곡 *huykok* "comedy skit", 비극 *pikuk* "tragedy skit", 서정시 *secengsi* "seasonal poem" and 연극 *yenku* "musical". This indicates that 코미디 *khomiti* "comedy" and 희극 *huykuk* "comedy" dominate the different realms of foreign art culture and traditional art culture.

Next, consider the case of 보너스 *ponesu* "bonus" and 상여금 *sangyekum* "bonus". The nearest neighbors of 상여금 *sangyekum* "bonus" mainly mean general salary and economic supply. On the other hand, in the nearest neighbors of 보너스 *ponesu* "bonus", special borrowed terms such as 콤보 *khompo* "combo", 아이템 *aitheym* "item", 경험치 *kyenghemchi* "experience point", 베팅 *peything* "batting", 퀘스트 *khweysuthu* "quest" standout. A closer look reveals that these loanwords are words used in computer games. From this observation, it can be said that 보너스 *ponesu* "bonus" is used for the meaning of supply in a game, although it is the same meaning as the financial supply as 상여금 *sangyekum* "bonus". Accordingly, it is clear that 보너스 *ponesu* "bonus" and 상여금 *sangyekum* "bonus" have the same meaning but dominate different semantic field.

Figure 2.4 visually shows the usage context sharing condition of these loanword-synonym pairs in 2-D projected vector space. As indicated within the figures, each cluster of nearest neighbors are largely separated, thus the result of figures also supports the discussion about semantic differentiation intuitively.

The middle part of our result also shows some difference of usage between a loanword and the Korean synonym. In the case of 유니폼 *yuniphom* "uniform" and 제복 *ceypok* "uniform", the nearest neighbors of 유니폼 *yuniphom* "uniform" such as 축구장 *chwukkwucang* "football stadium", 토트넘 *thothunem* "Tottenham", and 월드컵 *weltukhep* "world cup" imply that 유니폼 *yuniphom* "uniform" have the semantic field of sports clothing. Whreas the nearest neighbors of 제복 *ceypok* "uniform" such as 군복 *kwunpok* "military clothing", 베레모 *peyleymo* " beret", and *centhwupok* "battle dress" imply that 제복 *ceypok* "uniform" have the semantic field of military clothing.

In the case of the case of 멤버 *meympe* "member" and 구성원 *kwusengwen* "member", the nearest neighbors of 멤버 *meympe* "member" is related to the musical band group such as 드러머 *tuleme* "drummer", 베이시스트 *peyisisuthu* "bassist", 기타리스트 *kithalisuthu* guitarist, and 보컬 *pokhel* "vocal". These nearest neighbors indicate 멤버 *meympe* "member" is used in the context of the music group member. While the nearest neighbors of 구성원 *kwusengwen* "member" are the words about an organization or group such as 리더 *lite* "leader", 집단 *ciptan* "group", 개개인 *kaykayin* "individual", 의사결정 *uysakyelceng* "decision-making", 당원 *tangwen* "member". These nearest neighbors indicate 구성원 *kwusengwen* "member" is used in a social organization.

It is thought that these difference in the possessed semantic field has moved these loanword-synonym pairs closer to semantic differentiation.

2.5 Conclusion and Future Work

This study suggested a word vector-based method for investigating the language changes caused by lexical competition between loanwords and native words quantitatively. The vector space and the geometrical concept effectively model the usage context sharing condition between the loanwords and the native synonyms in this method. Although our method has difficulty to find loanword-synonym pairs judged to completely word replacement or completely semantic differentiation, our method succeeded in showing some tendency of the lexical competition. However, we must improve several technical limitations.

First, this model can only show the snapshot of the lexical competition that has been happening in a long time span and can not show the process of the lexical competition. For example, this model can not reveal the process of semantic differentiation: whether loanword-synonym pairs shared the same context at first and moved to different semantic fields over time. For overcoming this limitation of this model, a diachronic language database that reflects the language change through time must be needed, but there is no available diachronic language data in Korean. Developing a new diachronic language database will allow the analysis of the process of the lexical competition more accurately by using the method suggested in this study.

With the collaboration of our method and the methods of diachronic semantic change, the unknown linguistic laws and principles related to language change and lexical competition will be uncovered. Moreover, the word embedding models will solve more complicated semantic problems in the future.

Second, in the process of selecting the loanword-Korean synonym pairs with one to one competition relationship, we lost a lot of candidates. To find out what kind of lexical competition that loanwords have experienced, this model must inevitably select loanwords and Korean synonym pairs having the one-to-one competitive relationship. For this purpose, setting up various filtering procedures caused the number of loanword-Korean synonym pairs to decrease. Through the analysis of these few pairs, our model suggested some trends and the potential for the analytical ability of the lexical competition, but our model can not draw general linguistic conclusions on the lexical competition between loanwords and Korean synonyms. Future research will need to improve the selection process and develop a new method extracting more loanword-Korean synonym pairs.

We believe that these competitive relationships between synonyms can occur not only between loanwords and native synonyms but also between synonyms within the same language. It will be worth investigating what results will be obtained when using our model in the synonym study. Additionally, this study targeted only Korean, but loanwords are a popular phenomenon existing in almost all languages around the world. Thus, targeting various languages and comparing the difference between languages will be interesting in future works.

Chapter 3

Applying Word Embeddings to Measure the Semantic Adaptation of English Loanwords in Japanese and Korean[1]

3.1 Overview

In recent decades, English has become an international language. English is spoken as the native language in several countries and taught as a second language in many more. Over the course of English's rise as an international language, many English words have had an influence on the native languages of countries where English is not the mother tongue. Foreign words are often incorporated into a language in order to express a specific concept that cannot be expressed using the words of the mother tongue alone. For example, consider the word resident. *Resident* means *a person staying in a specific area* and *a person who is training to be a doctor* in English. However, *resident* as a loanword is mostly used with the second meaning in Japanese and Korean, because these two languages each have a native word for the first meaning of *resident*. This example shows that some of the original meanings of a loan word are not used in

[1]The content of this chapter is a correction and supplement of the content of the paper published as Yamada and Shin (2017).

foreign countries (Kay, 1995; Okawa, 2008; Cheon, 2008). Furthermore, loanwords are often used figuratively. In Japanese and Korean, the word *corner* indicates not only *a positional area*, but also *a section provided for a specific purpose*. The word *stand* is also frequently used to mean *a desk lamp* in Japan and Korea. Due to this phenomenon, the same English word is often used differently depending on the language. This semantic difference can pose a challenge in computational tasks such as machine translation and information retrieval. Additionally, the semantic difference of loanwords can also pose a challenge to language learners. For these reasons, the task of investigating the nature of a semantic adaptation when a word enters from a foreign language is an important one.

In order to deal with the challenges posed by loanwords, it is first necessary to develop a methodology for detecting the meaning difference of loanwords. To this end, we review the previous studies of computational models for word meaning change. Kulkarni et al. (2015) propose a new computational approach for tracing change of meaning and usage of words from a historical perspective. They construct a property time series of word usage and apply statistically sound change point detection algorithms to show the semantic change. The result shows interesting patterns of language change. Hamilton et al. (2016a) compare three major computational methods, PPMI, SVD, word2vec, and develop a powerful methodology for quantifying historical semantic change. They also tackle linguistic complications related to historical semantic change—the relationships between semantic change and word frequency and between semantic change and polysemy. As a result, they propose two quantitative laws of semantic change. Takamura et al. (2017) apply a word vector space model for semantic changes in Japanese loanwords. They train a word vector space model with English and Japanese text data and map Japanese loanword vectors onto the English vector space. After that, a Japanese loanword's vector is compared with an original English word vector according to their cosine similarity. This method is evaluated by several tests and is verified as a reliable method for studying semantic change in loanwords.

As demonstrated in these previous studies, the word vector space model is considered one of the most powerful methods for detecting differences in word meaning. Based on these previous studies, it is highly probable that the word vector space model is also powerful for detecting English loanwords as well as their semantic adaptation.

Fenogenova et al. (2017) applies the word vector space model to detect English loanwords in Russian data. Their detection method is based on the idea that the original Latin word is similar to its Cyrillic analogue in terms of scripting, phonetics, and semantics. They also assume that English loanwords and their original English words should be close in their meanings; their vector value is also similar. On this assumption, they develop a filtering system for detecting real loanwords from several loanword candidates in Russian data. As a result, they improve the accuracy of detecting English loanwords. However, their method only manipulates the loanwords that have the same meaning as the original English word. Thus, this study applies the word vector model to the task of detecting English loanwords whose semantic usage is different from the source English word and for measuring the degree of its semantic adaptation.

In addition to this methodological purpose, we verify the relationship between polysemy and meaning adaptation. As mentioned earlier, the main purpose of using loanwords is introducing a new concept. Thus, loanwords will tend to have only a part of the meaning that the word originally had. Given this supposition, it can be predicted that if an original word has several meanings (polysemy), the meaning between loanwords and the original English word will be much different. Hamilton et al. (2016a) study the relationship between polysemy and the meaning change of a word, but they study only from the perspective of historical meaning change and do not investigate the relationship from the point of view of meaning change in loanwords. To verify this prediction, we examine the relationship between the number of original meanings of the English word and the degree of semantic adaptation using the word vector model.

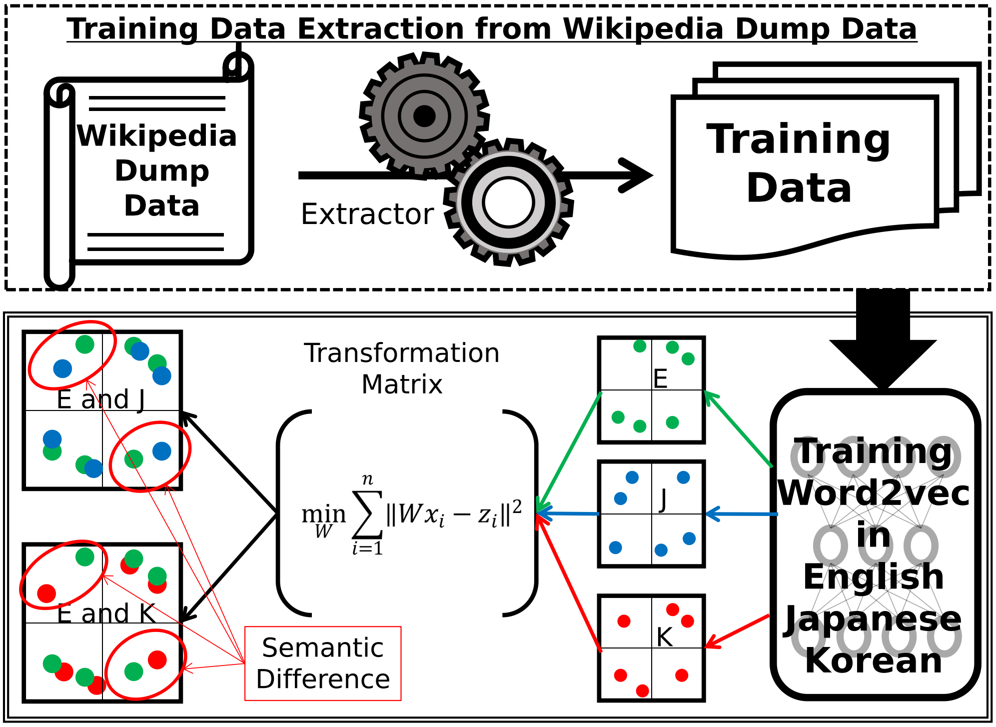

Figure 3.1: The experimental procedure of training data extraction and transformation.

3.2 Methodology

We use the word vector model for detecting English loanwords that have different meanings from their source words and for measuring the degree of their semantic adaptation. For this purpose, the Word2vec model (Mikolov et al., 2013a) is chosen to generate the word vector space model with reference to Hamilton et al. (2016a) and Takamura et al. (2017). We chose English, Japanese, and Korean, because English loanwords that are semantically distinct from their source words are abundant in both Japanese and Korean.

At first, we create word embedding for the three languages: English, Japanese, and Korean. Next, we calculate the cosine similarity and dissimilarity between the original English words and their Japanese or Korean loanword counterparts. For this purpose, the two language's words should be represented in the same vector space. For mapping the embeddings into the same vector space, we choose one of the simplest methods developed by Mikolov et al. (2013b). The method is represented by the equation 3.1. By calculating the equation using seed words, the transformation matrix W is obtained. To make the bilingual seed word pairs, we used the most frequent nouns from monolingual source data sets and translated those words using Google Translate like Mikolov et al. (2013b). By multiplying the value of an English loanword vector in Japanese or Korean by the transformation matrix W, it becomes possible to compare the loanword vectors in the English word vector space.

$$W = \min \sum_{i=1}^{n} \|Wx_i - z_i\|^2 \tag{3.1}$$

After this transformation, we can get the N-nearest neighbors of the English loanword in the English vector space and can calculate the cosine similarity between the English loanwords and the original English words. If the value of cosine similarity is low, it shows that the English loanword meaning is very different from the original word, and thus we can detect the English loanwords that are used with significantly different meanings in Japanese and Korean. In the next section, we present our data set and experiment for English loanword detection in Japanese and Korean.

3.3 Data and Experiment

The data set used for training Word2vec was obtained from Wikipedia dump data, English[2], Japanese[3], Korean[4], in May of 2017 for English, Japanese and Korean. The text data was extracted by a Wikipedia extractor[5] from each Wikipedia dump data set. The English Wikipedia data is 13.6 GB, the Japanese Wikipedia data is 2.5 GB

[2]https://dumps.wikimedia.org/enwiki/

[3]https://dumps.wikimedia.org/jawiki/

[4]https://dumps.wikimedia.org/kowiki/

[5]http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

and the Korean Wikipedia data is 606MB. In the case of English data, non-alphabetic symbols are removed and all alphabetic characters are lowered. For Japanese data, word segmentation is done using the Japanese morphological analyzer MeCab (Kudo et al., 2004). For Korean data, we apply the open-source Korean text tokenizer Twitter[6]. These preprocessed data are used for training Word2vec (dimensions = 200, min count = 20, window size = 15) in the Gensim[7] Python package.

Calculating the transformation matrix requires bilingual word lists: English-Japanese word list and English-Korean word list. This experiment prepared bilingual lists with the method of Mikolov et al. (2013b). Mikolov et al. (2013b) selects the high-frequency words from the English corpus and translates them into another language with Google translator. It may be easy to make a bilingual list in Spanish or French for calculating the transformation matrix but difficult in the case of Japanese and Korean because these languages have intricate inflection systems. This intricate inflection system makes one-to-one mapping of English words to Japanese (or Korean) words difficult and puts difficulty in obtaining an accurate transformation matrix. For example, when Google Translate makes a bilingual list of English and Japanese (or Korean) words, Google Translate translates *eat* to 먹다 *mekta* "eat" and *beautiful* to 아름다운 *alumtawun* "beautiful". If you calculate the transformation matrix using this bilingual list, *eat* is mapped with 먹다 *mokta* "eat" and *beautiful* is mapped with 아름다운 *alumtawun* "beautiful". However, 먹다 *mokta* "eat" is actually used in a different form such as 먹을 *mokulye* "to eat" or 먹겠 *mokess* "will eat" in texts. Similarly, 아름다운 *alumtawun* "beautiful" is used in different forms, such as 아름다웠 *alumtawess* "was beautiful" or 아름답다 *alumtapta* "beautiful". Therefore, if the bilingual list created by Google Translate is used to get a transformation matrix, other forms of 먹다 *mokta* "eat" and 아름다운 *alumtawun* "beautiful" are ignored in the process of calculation. As a result, the transformation matrix based on this bilingual list will not be accurate. Thus we chose

[6] `https://github.com/twitter/twitter-korean-text`

[7] `https://radimrehurek.com/gensim/index.html`

Word pairs	cosine similarity	Nearest Neighbors
consent khonseynthu	0.067	consider, recommend, agree, suggest, proposing chwungcenki, phulleku, suwichi, suphikhe, pulesi
corner khone	0.27	corners, edge, street, entrance, avenue kaykukhonsethu, thokhusyo, chwulyenca, cinhayngca, ayngkhe
stand suthayntu	0.33	sit, hang, hold, standing, walk kwancwungsek, theylasu, pheynsu, philtu, pulisci
date teyithu	0.056	dates, dated, chronology, calendar, birthdate kyocey, yecachinkwu, namcachinkwu, tongke, twulise

Table 3.1: The previous study's examples of the original English word and Korean loanword pairs which have undergone the semantic adaptation.

high-frequency English nouns because noun has little inflection in Japanese and Korean. After translating these English nouns into Japanese and Korean, loanwords are removed for training transformation matrix properly. Finally, we compute the transformation matrix with about 5000 word pairs in the list.

After learning word2vec with the Wikipedia data and doing the transformation, we use the nearest neighbors to check whether the word2vec and the transformation matrix are correctly trained. As an example, several loanwords which meaning is different from the original English word are selected from previous studies (Noh, 2013; Min, 1998). Table 3.1 shows the cosine similarity and the nearest neighbors of the original English word and the loanword pairs which have undergone the semantic adaptation shown in the previous studies.

The target loanwords that we study the semantic differences of in this research are selected from the loanword list distributed by the National Institute of Korean

Language[8]. Calculating cosine similarity requires bilingual loanword pair lists: an English-Korean loanword pair list and an English-Japanese loanword pair list. For obtaining these bilingual loanword lists, we translated the Korean loanwords list into English and Japanese with Google Translate. Checking the translated loanword pairs list found some mistakenly translated loanword pairs, namely the translated word was not the loanword of the original English word. These errors were caused by the mistranslation of Google Translate in the process of making the bilingual loanword lists. We removed these errors from the translated list. Additionally, if a loanword has homonyms or is used for a proper noun regardless of its original meaning and an independent article about the proper noun is registered on Wikipedia, we removed the loanword from this loanword list, because it was observed that loanwords are basically infrequent and their output results are easily affected by such noise data. We calculated the cosine similarities of the bilingual loanword pairs in this translated list. The final English-Korean loanword list has 1267 words and the final English-Japanese loanword list has 1308 words. All experimental processes in this study are summarized in Figure 3.1. The next section presents the result of this experiment.

3.4 Result and Discussion

This section shows how accurately the word vector model finds the differences in semantic usage of loanwords in Japanese and Korean. The value of cosine similarity is calculated based on the bilingual list that was explained in detail in the above section. For verifying the possibility that a learning error of word2vec produced these low cosine similarities, the frequency of the word in each language corpus is also shown in the case of low cosine similarity word pairs. Additionally, the N-nearest neighbors of several words in low cosine similarity word pairs are shown for the purpose of checking the accuracy of the word2vec learning process. In Section 3.4.4, we discuss the relationship

[8] `http://www.korean.go.kr/front/etcData/etcDataView.do;front=8E3DC144E9BBA954E0BE198B8481950A?mn_id=46&etc_seq=322&pageIndex=1`

of cosine similarity with the number of meanings of the original English words as mentioned at the beginning of this study.

3.4.1 Japanese

Table 3.2 shows the cosine similarities between an original English word and the corresponding English loanword in Japanese (the Japanese loanword). For the reference, the frequency of low cosine similarity words in English and Japanese is given in Table 3.3. Table 3.4 contains examples of the nearest neighbors of the lowest cosine similarity pairs in English and Japanese. As can be seen from *synchronize* in Table 3.3 and Table 3.4, we may safely say that word2vec learns the semantic information of the word properly even if the word frequency is only around 20.

Table 3.2 shows the original English words that have high cosine similarities with their corresponding Japanese loanwords in the left column, and the English words having low cosine similarities in the right column. In the left column, almost all words are technical words such as *design*, *tunnel*, *robot*, *data*, *engine*, *model*. This result corresponds to the findings of Takamura et al. (2017).

Table 3.4 shows the nearest neighbors of the five lowest cosine similarity pairs. Table 3.2 shows several interesting tendencies of meaning adaptation. For example, the English *tissue* means not only *soft thin paper* but *the biological component organized with cells*.

The first example is *synchronize*. Table 3.4 shows English *synchronize* means *adjust the two or more events to happen at the same time*. The nearest neighbors of Japanese loanword シンクロナイズ *shinkuronaizu* "synchronize" mainly are the electric-related words. This indicates the Japanese loanword used in the electrical field. One of the Japanese loanword nearest neighbors is a Japanese high school name. Detail research reveals this high school has an event of synchronized swimming. This slightly indicates Japanese loanword シンクロナイズ *shinkuronaizu* "synchronize" used in the context of *synchronized swimming*.

Similar loanwords			Dissimilar loanwords		
English	Loanword	Cosine similarity	English	Loanword	Cosine similarity
design	dezain	0.86	synchronize	shinkuronaizu	0.11
robot	robotto	0.84	checker	chiekka	0.12
data	deta	0.84	pick	pikku	0.18
engine	enjin	0.84	feed	fuido	0.18
text	tekisuto	0.84	foundation	fuandeshon	0.20
model	moderu	0.84	roleplaying	rorupureingu	0.20
infrastructure	infura	0.84	shortening	shotoningu	0.20
knife	naifu	0.83	living	ribingu	0.21
system	shisutemu	0.83	editor	edeita	0.22
inflation	infure	0.82	ground	gurando	0.22
radar	reda	0.82	figure	fuigyua	0.24
laser	reza	0.81	number	namba	0.25
project	purojiekuto	0.81	cabinet	kyabinetto	0.26
algorithm	arugorizumu	0.81	register	rejisuta	0.26
festival	fuesuteibaru	0.81	supporter	sapota	0.26
restaurant	resutoran	0.80	handicap	handeikyappu	0.27
camera	kamera	0.80	label	raberu	0.27
inflation	infureshon	0.80	olympiad	orimpiado	0.27
leader	rida	0.80	demo	demo	0.28
gas	gasu	0.80	handy	handei	0.28

Table 3.2: The top twenty similar and dissimilar loanword-original English pairs in Japanese.

Language	English		Japanese	
Word	Frequency	Rel. freq.	Frequency	Rel. freq.
synchronize	1466	2.89E-06	20	3.95E-08
checker	2171	4.29E-06	627	1.24E-06
pick	64359	1.27E-04	1070	2.11E-06
feed	65222	1.29E-04	563	1.11E-06
foundation	255004	5.03E-04	203	4.01E-07
roleplaying	2267	4.48E-06	71	1.40E-07
shortening	3678	7.26E-06	67	1.32E-07
living	432352	8.54E-04	726	1.43E-06
editor	248728	4.91E-04	1028	2.03E-06
ground	289150	5.71E-04	10239	2.02E-05
figure	155667	3.07E-04	12207	2.41E-05
number	1204683	2.38E-03	10179	2.01E-05
cabinet	83939	1.66E-04	355	7.01E-07
register	158908	3.14E-04	2360	4.66E-06
supporter	32102	6.34E-05	3705	7.31E-06
handicap	13332	2.63E-05	400	7.90E-07
label	173062	3.42E-04	2222	4.39E-06
olympiad	5991	1.18E-05	49	9.67E-08
demo	25974	5.13E-05	11625	2.29E-05
handy	4696	9.27E-06	504	9.95E-07

Table 3.3: The frequency and the relative frequency (Rel. freq.) of the twenty lowest cosine similarity English words and Japanese loanword pairs in each language corpus.

Word	Neighbors of English	Neighbors of Loanword
synchronize	synchronise	puriemputeibu
	synchronizes	inagakuensogo
	configure	Backup
	synchronizing	puroguramabururojikkukontorora
	align	nihondenshisemmongakko
checker	checkers	doraibusurupenarutei
	spellchecker	penarutei
	regex	patoreze
	auto-completion	doraibazupointo
	spell-check	fuomeshonrappu
pick	picks	shoruda
	picking	suteikku
	picked	terekyasuta
	catch	hando
	roughed	pegu
feed	feeds	RSS
	feeding	torakkingu
	forage	kuikku
	consume	wantatchi
	ingest	insaido
foundation	foundations	manikyua
	fund	hiyake
	endowment	pauda
	foundations	kuchibeni
	institute	roshon

Table 3.4: The nearest neighbors of five most dissimilar loanword pairs in Japanese.

In the case of *checker*, the nearest neighbors show the English *checker* means *spell checker in computer software* and the nearest neighbors of loanword–ドライブスルーペナルティ *doraibusurupenarutei* "Drive Through Penalty", ドライバーズポイント *doraibazupointo* "Drivers Point", フォーメーションラップ *fuomeshonrappu* "Formation Lap"–indicate the Japanese loanword チェッカー *chiekka* "checker" used in the context of *moter racing*. In the case of *pick*, the nearest neighbors of English show the English *pick* means *choose a person or thing*. The nearest neighbors of loanword–スティック *suteikku* "stick", ペグ *pegu* "peg", テレキャスター *terekyasuta* "Telecaster" –indicate Japanese loanword ピック *pikku* "pick" is used as *a small flat tool for pulling the strings of a musical instrument*.

In the case of *feed*, the nearest neighbors indicate the English *feed* means *giving food*. In the nearest neighbors of Japanese loanwords, *RSS* and トラッキング *torakkingu* "tracking" indicates the Japanese loanword フィード *fuido* "feed" is used as *documents processed for web distribution (news feed or web feed)*. Other neighbors, ワンタッチ *wantatchi* "one touch" and インサイド *insaido* "inside", indicate the usage in sports as *throw or hit a ball to a teammate*. Wikipedia also provides a lot of sports-related sentences having フィード *fuido* "feed".

Finally, in the case of *foundation*, the nearest neighbors of English *foundation* indicate *foundation* means *the organization providing financial support*. The nearest neighbors of Japanese loanword ファンデーション *fuandeshon* "foundation" indicate the Japanese loanword means *cosmetics*. This meaning difference between English and Japanese loanword affects the large difference in cosine similarity.

From these examples, it is shown that this word vector model detects several patterns of meaning adaptation of English loanwords in Japanese.

3.4.2 Korean

Table 3.5 shows the cosine similarities between an original English word and the corresponding English loanword in Korean (Korean loanword). The frequency of low cosine similarity words in English and Japanese are shown in Table 3.6.

In Table 3.5, the left column shows the English words that have high cosine similarities with their corresponding English loanwords in Korean and the right column shows the English words with low cosine similarities with their corresponding loanwords. In the left column, almost all words are technical terms such as software, energy, producer, network, and algorithm or academic terms such as fascism, and realism. This result is almost the same as in the Japanese data set. From this result we can observe the tendency of technical term meanings to remain constant, which was also observed by Nishiyama (1995); this observation appears to also be applicable in the case of semantic adaptation of English loanwords in Korean.

Table 3.7 shows the nearest neighbors of the top five lowest cosine similarity pairs in Table 3.5. The nearest neighbors of *active* show that the English word means *busy physical or mental condition* and loanword 액티브 *aykthipu* "active" is used in the context of the product names. The English word "professional" and loanword 프로페셔널 *phulopheysyenel* "professional" has the same pattern of semantic adaptation of "active".

The nearest neighbors of *figure* show that the English word means *a person who thinks or explains* and the loanword 피겨 *phikye* "figure" have semantic relation with winter sports. This indicates the loanword 피겨 *phikye* "figure" is mainly used in the meaning of *a figure skating* in Korean.

The nearest neighbors of *total–maximum*, *megatonnes*, and *million*–indicate that the English *total* is used as *mathematical meaning of quantity*. The nearest neighbors of loanword 토털 *thothel* "total" are abstract meaning words: 마인드 *maintu* "mind", 라이프 *liiphu* "life. This indicates that Korean loanword 토털 *thothel* "total" is often used in the sense of conceptual synthesis, not just numerical totals.

The nearest neighbors of *cabinet* show that the English word is used as *a group of high position officials*: *cabinet ministers or secretaries*. Whereas, the nearest neighbors of the loanword 캐비닛 *khaypinis* "cabinet" indicate the loanword means *furniture*

Similar Loanwords			Dissimilar Loanwords		
English	Loanword	Cosine similarity	English	Loanword	Cosine similarity
software	sophuthuweye	0.85	active	aykthipu	0.02
journalist	cenellisuthu	0.83	figure	phikye	0.02
logo	loko	0.82	professional	phulopheysyenel	0.03
fascism	phasicum	0.82	total	thothel	0.04
infrastructure	inphula	0.82	cabinet	khaypinis	0.05
energy	eyneci	0.82	businessman	picunisumayn	0.09
message	meysici	0.81	synchronize	singkhulonaicu	0.12
marketing	makheything	0.81	resident	leycitenthu	0.12
producer	phulotyuse	0.81	speaker	suphikhe	0.13
text	theyksuthu	0.81	caption	khaypsyen	0.13
project	phuloceykthu	0.80	complex	khomphulleyksu	0.14
network	neythuwekhu	0.80	minicar	minikha	0.16
inflation	inphulleyisyen	0.79	trade	thuleyitu	0.18
college	khallici	0.79	french	phuleynchi	0.18
algorithm	alkolicum	0.79	calendar	khayllinte	0.19
genre	canglu	0.79	orientation	olieyntheyisyen	0.19
forum	pholem	0.79	cork	kholukhu	0.20
realism	liellicum	0.79	facsimile	phayksimilli	0.20
tournament	thonementhu	0.79	thrill	sulil	0.21
computer	khemphyuthe	0.79	commission	khemisyen	0.21

Table 3.5: The top twenty similar and dissimilar loanword-original English pairs in Korean.

Language	English		Korean	
Word	Frequency	Rel. freq.	Frequency	Rel. freq.
active	284924	1.41E-04	399	3.91E-06
figure	155667	7.70E-05	1118	1.10E-05
professional	438212	2.17E-04	43	4.22E-07
total	530197	2.62E-04	63	6.18E-07
cabinet	83939	4.15E-05	50	4.90E-07
businessman	51077	2.53E-05	22	2.16E-07
synchronize	1466	7.25E-07	168	1.65E-06
resident	63163	3.13E-05	135	1.32E-06
speaker	66653	3.30E-05	410	4.02E-06
caption	5457	2.70E-06	22	2.16E-07
complex	221913	1.10E-04	279	2.74E-06
minicar	87	4.30E-08	28	2.75E-07
trade	259990	1.29E-04	2632	2.58E-05
french	631056	3.12E-04	321	3.15E-06
calendar	47612	2.36E-05	97	9.51E-07
orientation	28772	1.42E-05	44	4.31E-07
cork	35457	1.75E-05	130	1.27E-06
facsimile	2346	1.16E-06	29	2.84E-07
thrill	3934	1.95E-06	243	2.38E-06
commission	246461	1.22E-04	20	1.96E-07

Table 3.6: The frequency and relative frequency (Rel. freq.) of the twenty lowest cosine similarity English words and Korean loanword pairs in each language corpus.

Word	English word neighbors	Korean loanword neighbors
	inactive	melthi
	important	khenthulolle
active	influential	locik
	involved	haiphe
	engaged	tipaisu
	figures	sukheyithing
	personage	phikyesukheyithing
figure	thinker	sukheyithu
	exponent	sunopotu
	depiction	syothuthulayk
	amateur	hankulkwakhemphyuthe
	semi-professional	aimayk
professional	professionally	locitheyk
	semiprofessional	neyksuthusutheyp
	full-time	khintul
	maximum	maintu
	km2	thothal
total	totaling	lasuthu
	megatonnes	seykhentu
	million	laiphu
	ministerial	thechisukhulin
	parliament	khwetu
cabinet	cabinets	khwulle
	ministers	phayk
	minister	sullos

Table 3.7: The nearest neighbors of five most dissimilar loanword pairs in Korean.

attached with doors and shelves or drawers. Wikipedia sentences of 캐비닛 *khaypinis* "cabinet" also indicate the meaning. The low cosine similarity shows this meaning difference.

The other examples in Table 3.7 show meaning differences between the original English words and the Korean loanwords. For example, the Korean loanwords 스피커 *suphikhe* "speaker" means *Audio equipment* and the Korean loanword 콤플렉스 *khomphulleyksu* "complex" is mainly used as a meaning of *a mental problem of unnecessary anxiety*.

These examples indicate that in Korean data the word vector model detects the several tendencies of meaning adaptations in English loanwords in Korean.

3.4.3 Comparison of Cosine Similarities of English Loanwords in Japanese and Korean

This section presents a contrastive study of the difference in the semantic adaptation of English loanwords between Japanese and Korean. After calculating the cosine similarity between Japanese and English and between Korean and English as in the previous section, the difference of the cosine similarity values finds the highly distinct semantic usage in Korean and Japanese. Table 3.8 shows the five highest (smallest) English words in the difference of the cosine similarity with Korean loanwords and with Japanese loanwords (Korean - Japanese). The above half of Table 3.8 shows the English words having higher cosine similarity with Korean loanwords than with Japanese loanwords. The below half of the Table 3.8 shows the English words having higher cosine similarity with Japanese loanwords than with Korean loanwords. Table 3.9 and Table 3.10 show the nearest neighbors of the English words, the Korean loanwords, and the Japanese loanwords of each cases. Table 3.10 shows the nearest neighbors of the English words, the Korean loanwords, and the Japanese loanwords of the five largest cosine similarity difference. With training the models with different data sets and calculating the transformation matrix independently, comparing the values directly

Word	Korean cosine similarity	Japanese cosine similarity	Difference Korean-Japanese
olympiad	0.64	0.27	0.37
demo	0.54	0.28	0.26
editor	0.46	0.22	0.24
close-up	0.60	0.36	0.23
roleplaying	0.43	0.20	0.23
caption	0.13	0.67	-0.54
scrap	0.22	0.62	-0.40
calendar	0.19	0.58	-0.39
diorama	0.35	0.71	-0.36
microfilm	0.30	0.65	-0.35

Table 3.8: The top five English loanwords whose cosine similarity with their Korean (or Japanese) loanword is higher than with their Japanese (or Korean) loanword.

may prove challenging. But this pioneering contrastive study suggests the possibility of detecting several tendencies of the semantic adaption difference between Japanese and Korean by the word embedding model.

The English Words Having Higher Cosine Similarity with Korean Loanwords than with Japanese Loanwords

The nearest neighbors in Table 3.9 shows the difference in semantic adaptation between Korean loanwords and Japanese loanwords. The English *olympiad* is used not only in the historical term but also in championships such as *the mathematical olympiad* and *the scientific olympiad*. The nearest neighbors of the Korean loanword 올림피어드 *ollimphietu* "olympiad" indicate that Korean loanword also have the same semantic adaptation. The nearest neighbors of Japanese loanword オリンピアード *orimpiado*

Word	English neighbors	Korean neighbors	Japanese neighbors
olympiad	olympiads	simphociwum	oryumpia
	iypt	simphociem	irahabado
	deaflympics	Olympiad	pyuteia
	biennial	senswukwentayhoy	isutomia
	biennale	khongkhwul	maraton
demo	demos	theyiphu	suwarikomi
	demos	laipu	bodo
	recording	theyip	kogi
	ep	nokum	koshin
	4-track	theyiph	gaito
editor	editor-in-chief	sukhulipthu	tekisutoedeita
	editorial	capasukhulipthu	sukuriputo
	columnist	phullekuin	GUI
	contributor	kimphu	uijietto
	publisher	pyue	WYSIWYG
close-up	closeup	khaypche	torizata
	close-ups	phayleti	hodo
	closeups	kulotheysukhu	torizata
	camera	chwalyeng	kenden
	silhouette	kaksayk	sanken
roleplaying	role-playing	MMORPG	rorupureingugemu
	boardgame	FPS	gapusu
	gurps	thencey	RPG
	battletech	akheyitu	FPS
	d20	RPG	uoshimyureshongemu

Table 3.9: The nearest neighbors of top five English loanwords whose cosine similarity with their Korean loanwords is higher than with their Japanese loanword.

"olympiad" indicate that Japanese loanword is used in the historical sense. In Japanese, championships such as *the mathematical olympiad* and *the scientific olympiad* use the different Japanese loanword オリンピック *orimpikku* "Olympic" instead of *olympiad*. This difference has possibly influenced the difference in cosine similarity.

The nearest neighbors of English *demo*, such as *recording* and *4-track*, indicate "demo" is used in acoustic-related meanings such as *recording* and *4-track* in nearest neighbors. The nearest neighbors of the Korean loanword 데모 *teymo* "demo" indicate that the Korean loanword is also used in the acoustic sense. The nearest neighbors of Japanese loanword デモ *demo* "demo" indicate the Japanese loanword means *protesting activity*.

The nearest neighbors of English *close-up* show that the English *close-up* is used in context with *photography* such as cameras, and Korean loanword 클로즈업 *khullocuep* "close-up" has also the same semantic adaptation indicated from the nearest neighbors of Korean loanword. The nearest neighbors of Japanese loanword クローズアップ *kurozuappu* "close-up" indicate the Japanese loanword is used in the context of news report or event. In Japanese, news reports often use クローズアップ *kurozuappu* "close-up" when focusing on important events. This difference has affected the difference in cosine similarity between Korean and Japanese.

The nearest neighbors of English *editor* indicate English word *editor* means *the responsible person in the decision of including which articles in a newspaper or magazine*. The nearest neighbors both Korean loanwords and Japanese loanwords are computer-related words. This result indicates Korean loanword 에디터 *eytithe* "editor" and Japanese loanword エディター *edeita* "editor" means *text editor in computer*.

The nearest neighbors of English *roleplaying* indicate English word *roleplaying* means *general roleplaying game not only video game*. The nearest neighbors both Korean loanwords and Japanese loanwords are video game-related words. This result indicates both Korean loanword 롤플레잉 *lolphulleying* "roleplaying" and Japanese loanword ロールプレイング *rorupureingu* "roleplaying" means especially *roleplaying* *video game*.

From the above results, calculating the difference between cosine similarity of Korean loanword and Japanese loanword can detect the difference of semantic adaptation of loanwords between Korean and Japanese, such as *olympiad*, *demo*, *foundation*, and *close-up*. Although the case of *editor* and *roleplaying* does not show the clear difference of semantic adaptation, the comparison of cosine similarity shows the possibility of comparative research.

The English Words Having Higher Cosine Similarity with Japanese Loanwords than with Korean Loanwords

The nearest neighbors of English *caption* show that the English *caption* means *the description displayed above or below a picture in a book, a newspaper, and a video*. Japanese loanword キャプション *kyapushon* "caption" has also similar meaning indicated from the nearest neighbors: 見出し *midashi* "header", 欄外 *rangai* "margin", and サブタイトル *sabutaitoru* "subtitle". Whereas the nearest neighbors of Korean loanword 캡션 *khaypsyen* "caption" are mainly the English words: page, image, and layout. Checking more nearest neighbors finds more English words: Edit, file, and word. These nearest neighbors can indicate 캡션 *khaypsyen* "caption" is used in computer document editors like *Microsoft Word* or *Hangul Word Processor*. Wikipedia sentences also shows some sentences of 캡션 *khaypsyen* "caption" used in computer context like 클로즈드 캡션 *khullocutu khaypsyen* "closed caption in YouTube". This difference has affected the difference in cosine similarity between Korean and Japanese.

The nearest neighbors in Table 3.10 shows the difference in semantic adaptation between Korean loanwords and Japanese loanwords. The nearest neighbors of English *scrap*, such as *scrapyards* and *shipbreakers*, indicate that *scrap* means *breaking a machine into pieces*. The nearest neighbors of Japanese loanword スクラップ *sukurappu* "scrap" shows machine and breaking words: 一隻 *isseki* "one ship", ベスレヘム·スチール *besurehemu·suchiru* "Bethlehem Steel Corporation", 解體 *kaitai* "break up". While

	English neighbors	Korean neighbors	Japanese neighbors
caption	captioned	page	midashi
	captions	image	rangai
	placard	yoyak	komidashi
	blurb	layout	bana
	disclaimer	*(astarisk)	sabutaitoru
scrap	scrapping	keysi	joseki
	scrapyards	eplotu	isseki
	shipbreakers	thwuko	baikyaku
	scrapyard	weyppheyici	besurehemu · suchiru
	shipbreaking	keycay	kaitai
calendar	calendars	cwusolok	himekuri
	gregorian	pyue	sutampu
	lunisolar	mopailmi	katarogu
	365-day	culkyechacki	Calendar
	metonic	pwukmakhu	furaiya
diorama	dioramas	phuloceykthe	minichua
	life-sized	malioneythu	mokei
	life-size	hollokulaym	obujie
	taxidermied	cengmwulhwa	purareru
	cyclorama	minieche	jitsubutsu
microfilm	microfiche	electronic	akaibu
	microfilms	Document	akaibu
	microform	phayksu	maikurofuisshu
	microfilmed	cencasacen	insatsubutsu
	microfilming	tisukheys	fukusha

Table 3.10: The nearest neighbors of top five English loanwords whose cosine similarity with their Japanese loanwords is higher than with their Korean loanword.

the nearest neighbors of Korean loanword 스크랩 *sukhulayp* "scrap" shows several internet-related words: 업로드 *eplotu* "upload" and 웹페이지 *weyppheyici* "web page". Additionally, other neighbors–게시 *keysi* "posting", 투고 *thwuko* "submit", 게재 *keycay* "published"–are the publication related-words. These neighbors indicate that Korean loanword 스크랩 *sukhulayp* "scrap" mainly means *cut and collect articles from newspapers and magazines, especially from website*. Checking the Wikipedia data finds several sentences that 스크랩 *sukhulayp* "scrap" means *cut and collect articles*. This meaning difference between the English and the Korean loanword affects the large difference of cosine similarities between the Korean loanword and the Japanese loanword.

The nearest neighbors of English *calendar* show that the English *calendar* means *a set of pages that show the days, weeks, and months of a particular year*. Japanese loanword カレンダー *karenda* "calendar" has also similar meaning indicated from the nearest neighbors: 日めくり *himekuri* "daily pad calendar" or フライヤー *furaiya* "reservation paper for buying a calendar". Whereas the nearest neighbors of Korean loanword 캘린더 *khayllente* "calendar" are the computer software-related words: 주소록 *cwusolok* "address book", 뷰어 *pyue* "viewer", 모바일미 *mopailmi* "Mobile me", 즐겨차기 *culkyechaki* "favorites", 북마크 *pwukmakhu* 'bookmark". Wikipedia sentences of the Korean loanword also shows 캘린더 *khayllente* "calendar" is used as a software name which manages the schedules, such as 구글캘린더 *kwukulkhayllinte* "google calendar". This difference has affected the difference in cosine similarity between Korean and Japanese.

Checking the nearest neighbors and Wikipedia sentence can not find any semantic differences for *diorama* and *microfilm* between Korean loanword and Japanese loanword even their large difference of cosine similarity. One of the reasons for this result can be the drawback of the transformation matrix method comparing the vector value learned using different databases. The cosine similarities between diorama and 디오라마 *tiolama* "diorama" (0.35) and between microfilm and 마이크로필름 *maikhulophillum* "microfilm" (0.30) are not much low compared to 캡션 *khaypsyen* "caption" (0.13) and

캘린더 *khayllente* "calendar" (0.19). These Korean loanwords cosine similarities (0.30 and 0.35) seems to be low, but in reality, it may be close to the meaning of English. In fact, some Korean loanwords having a cosine similarity value of about 0.3 have almost the same meaning as the original English words: 징크스 *cingkhusu* "jinx" (0.29), 신드롬 *sintulom* "syndrome" (0.30), and 룰렛 *lwulleys* "roulette" (0.30). Thus, even if the Korean loan word cosine similarity (0.3 and 0.35) and the Japanese loan word cosine similarity (0.71 and 0.65) seem to be numerically large, it can be quite possible that they have close meaning with the original English word.

Another reason may also be a little bit low frequency of 디오라마 *tiolama* "diorama" and 마이크로필름 *maikhulophillum* "microfilm". The minimum frequency of word2vec this time was set to 20. Previous researches setting the minimum frequency to a lower frequency (Rattinger et al., 2018; Ajees and Idicula, 2018) imply the frequency of 디오라마 *tiolama* "diorama" (28 times) and the frequency of 마이크로필름 *maikhulophillum* "microfilm" (23 times) are not so low. However, the relatively low frequency may have affected the ability of the model. Future research should investigate this point and improve the accuracy of the detection of semantic differences.

3.4.4 The Relationship Between the Number of Meanings and Cosine Similarities

This subsection investigates the relationship between the number of meanings of an English word and the degree of difference in its counterpart loanword's semantic usage. In many cases, loanwords tend to have certain specific meanings that cannot be expressed in a foreign language. Noh (2013) mentions that the universal semantic adaptation of the loanword is that the polysemous original word becomes the narrower sense loanword. Considering this tendency, it can be presumed that the difference of the meaning usage between the loanword and the original word will be large if the original word has a large number of meanings. In order to verify this hypothesis, this study sets the number of index in *Longman Dictionary of Contemporary English Fifth edition*

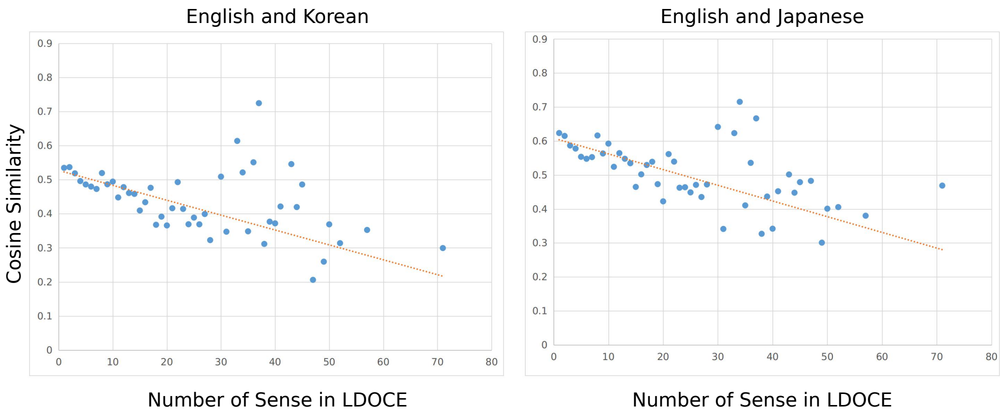

Figure 3.2: The result of the statistical test between the number of sense in the English dictionary and the cosine similarity.

(LDOCE) (Mayor, 2009) as the number of meanings of the English word.

Figure 3.2 shows the result of this experiment. The horizontal line shows the number of senses in LDOCE and the vertical line shows the cosine similarity of the English word and loanword. The trend line is calculated from the original data and the spots mean the average of the number of senses of an English word in LDOCE. The right graph shows the case of an English word and a Korean loanword. The left graph shows the case of an English word and a Japanese loanword.

In the case of Korean data, the slope of the regression line is -0.0044 and the *p-value* is zero (rounded to zero by R). This result shows that the number of senses of the original English word has a significant negative correlation with cosine similarities in Korean data. The case of Japanese data shows the same tendency. The slope of the regression line is -0.0046 and the *p-value* is zero (again rounded to zero by R), which again suggests that the two factors are negatively correlated. Based on these observations, it appears that the semantic difference between an original English word and the loanword largely occurs in cases where an original English word that has many meanings, polysemous, is used as a loanword which has narrow meaning in order to indicate a specific concept in a foreign country.

3.5 Conclusion and Future Works

This chapter analyzed the difference in the semantic adaptation of English loanwords in Japanese and Korean with the word embedding methods: the word2vec and transformation matrix. Word2vec and transformation method successfully detects the semantic adaptation in Japanese and Korean. The contrastive study shows the possibility of the contrastive linguistic study based on the word embedding models. The statistical analysis of the relation between the sense number and the degree of cosine similarity shows the advantage of the word embedding-based quantitative method in verifying the statistical relation of language.

While these good advantages of the word embedding method, there are several problems to solve in the word embedding-based linguistic study. The first problem is the character of the training data. The training data decide almost all the ability of the word embedding model. The unbalanced genre or contents in training data have a critical influence on the output of the word embedding model. The output can not reflect the linguistic meaning of native speakers correctly. Thus, word embedding method-based research must do an effort as much as possible to check the output is overly influenced by the specific data in the training data sets.

Future work should try to reveal the linguistic factors making the difference in the semantic adaptation. This study will reveal the statistical law of the semantic adaptation of loanword. A contrastive study using more languages will also reveal the linguistic mechanisms in semantic adaptation more broadly and more deeply. Hopefully, this pioneering study will solve problems and provide new insights into several academic fields: natural language processing, language education, and contrastive linguistic analysis.

Chapter 4

Detection of the Contextual Change of Loanwords and the Cultural Trend Change in Japanese and Korean through Pre-trained BERT Language Models

4.1 Overview

With the internationalization of English, English words have increasingly flowed into other languages. Most of these English words settle in other languages as English loanwords without being translated. Especially in Japanese, English loanwords frequently appear in the media (newspapers, magazines, TV programs), marketing, and academic fields (Daulton, 2004). Loanwords can strongly emphasize and draw attention to social issues, thus media frequently uses loanwords in broadcasting social events (Rebuck, 2002). Not only in the media but also in the names of new policies and new social systems, a lot of English loanwords are used because of their prestige and refined image (Tomoda, 1999).

From this cultural background of English loanwords in Japanese, we propose a hypothesis that tracking the change in contexts where the loanwords appear (the contextual change of loanwords) will reveal cultural trend changes such as social issues, social systems, policies, and the latest fashion trends. A similar situation of loanwords

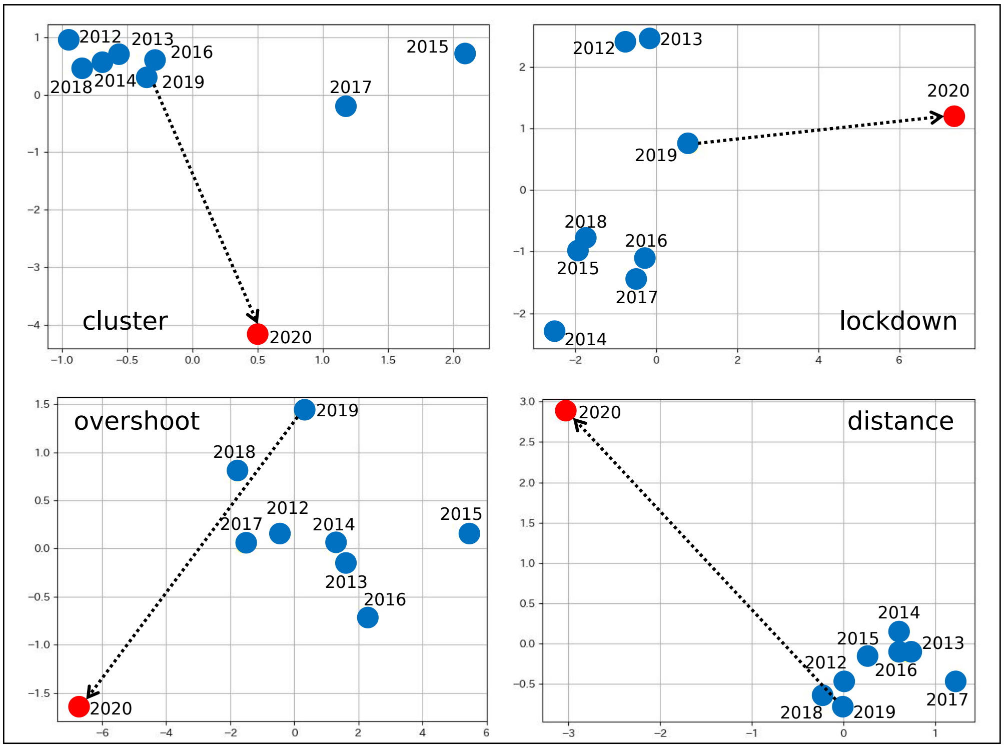

Figure 4.1: Two-dimensional visualization of semantic shift in COVID-19 related Japanese loanwords (クラスター *kurasuta* “cluster", ロックダウン *rokkudaun* “lockdown", オーバーシュート *obashuto* “overshoot", ディスタンス *deisutansu* “distance") using mean vectors of contextualized vectors in Japanese BERT. The semantic usage of these loanwords has dramatically shifted between 2019 and 2020.

is observed in Korean (Rüdiger, 2018; Shim, 1994; Song, 1998; Bruce Lawrence, 2010). In this research, we investigate the relationship between the contextual change of loanwords and cultural trend change in Japanese and Korean. For the purpose of this, we suggest a detection method of the contextual change of loanwords with the contextualized word embedding model.

Semantic change has been actively studied in the field of NLP (Tahmasebi et al., 2018; Kutuzov et al., 2018). However, because a large diachronic database is necessary for training language models with conventional methods, there has been little semantic change research done concerning resource-poor languages such as Japanese and Korean.

Recently, contextualized word embedding models (Peters et al., 2018; Devlin et al., 2018; Lample and Conneau, 2019; Radford et al., 2019) have clarified new characteristics of diachronic semantic change (Hu et al., 2019; Giulianelli et al., 2020). These studies show the usefulness of the pre-trained contextualized models in semantic change research and indicate a large diachronic database is not always necessary. Taking inspiration from the above work, we use a pre-trained contextualized word embedding model for the contextual changes of English loanwords in Japanese and Korean (Japanese loanwords and Korean loanwords). Figure 4.1 shows the contextual changes of COVID-19-related Japanese loanwords detected by our method in 2-D vector space.

For the diachronic language database, we used Twitter data, which is considered to significantly reflect cultural trend changes (Kulkarni et al., 2015; Jawahar and Seddah, 2019). The specific contributions of this study are as follows:

- With the contextualized word embedding model, we suggest a method of detecting the diachronic contextual change of loanwords in Japanese and Korean.
- Through an analysis of the contextual change of Japanese and Korean loanwords, we detect the cultural trend change when the contextual changes happened.
- This research provides a multilingual method for analyzing the contextual change and the culture trend change.
- From the perspective of linguistics, this research proposes the pioneering study of the diachronic contextual change in resource-poor language.

The remaining part of this study is organized as follows. After summarizing the previous studies in the Section 4.2, we describe and evaluate our detection procedure of contextual change of loanwords in Section 4.3. In Section 4.4, we analyze more

Japanese loanwords and Korean loanwords with our detection model. Finally, we summarize our results and suggest future works in Section 4.5.

4.2 Related Work

4.2.1 Loanwords and Cultural Trend Change

Fundamentally, loanwords fill in language-gaps when new concepts and new products inflow into another country. Loanwords also have a social semantic function such as expressing oneself distinctively in an organization or conversation, asserting social identity and giving an impression of prestige (Andersen et al., 2017; Zenner et al., 2019). Kay (1995) argues that loanwords, in Japanese especially, are used flexibly in various contexts because of a low awareness to preserve the original meaning. Rebuck (2002) says loanwords bestow recognition on a social problem. Politics also frequently use loanwords for the name of new policies and official documents (Tomoda, 1999). These features of loanwords promote frequent use of loanwords to suit current trends and needs. Thus, the contextual change of loanwords can be one of the indicators of cultural trend changes.

4.2.2 Word Embeddings and Semantic Change

As diachronic linguistic databases, such as Google N-gram (Lin et al., 2012) and the Corpus of Historical American English (COHA) (Davies, 2010), have been constructed and word embeddings models (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017) have been proposed, a lot of research on diachronic semantic change has been actively conducted in the NLP field. These studies have also greatly contributed to the development of historical linguistics and sociology.

Xu and Kemp (2015) quantitatively studied two opposing linguistic laws related to diachronic semantic change, namely the law of differentiation and the law of parallel change, and presented experimental evidence for this confrontation of two laws.

Hamilton et al. (2016a) used various models such as PPMI, SVD, and word2vec for investigating whether word frequency (the law of conformity) and word polysemy (the law of innovation) have influenced the diachronic semantic change.

As a sociological contribution, Hamilton et al. (2016b) experimentally studied the cultural aspect of a semantic change: "cultural shift". Garg et al. (2018) quantified the changes in social awareness of gender and ethnic stereotypes over the past 100 years by observing changes in words related to gender and ethnic stereotypes.

4.2.3 Contextualized Embedding and Diachronic Semantic Representation

In recent years, contextualized word embedding models have been proposed and brought state-of-the-art results in various tasks of NLP. Contextualized word embeddings models have revealed more detailed properties of diachronic semantic change.

Hu et al. (2019) quantitatively shows how the meanings of polysemous words change according to the times, and experimentally show the competitive relationship between each meaning from an ecological competitive viewpoint. Giulianelli et al. (2020) defines the contextualized vector values of the target word obtained from the pre-trained model as the usage vectors and shows how the cluster's proportion of usage vectors change over time.

We were motivated by the results of these studies and we also applied a contextualized model for analyzing the contextual change of loanwords in resource-poor languages: Japanese and Korean.

4.3 The Framework

4.3.1 Sense Representation

We define the sense of the target word as the contextualized word representation of the target word. For solving the out-of-vocabulary (OOV) problem of Japanese loanwords,

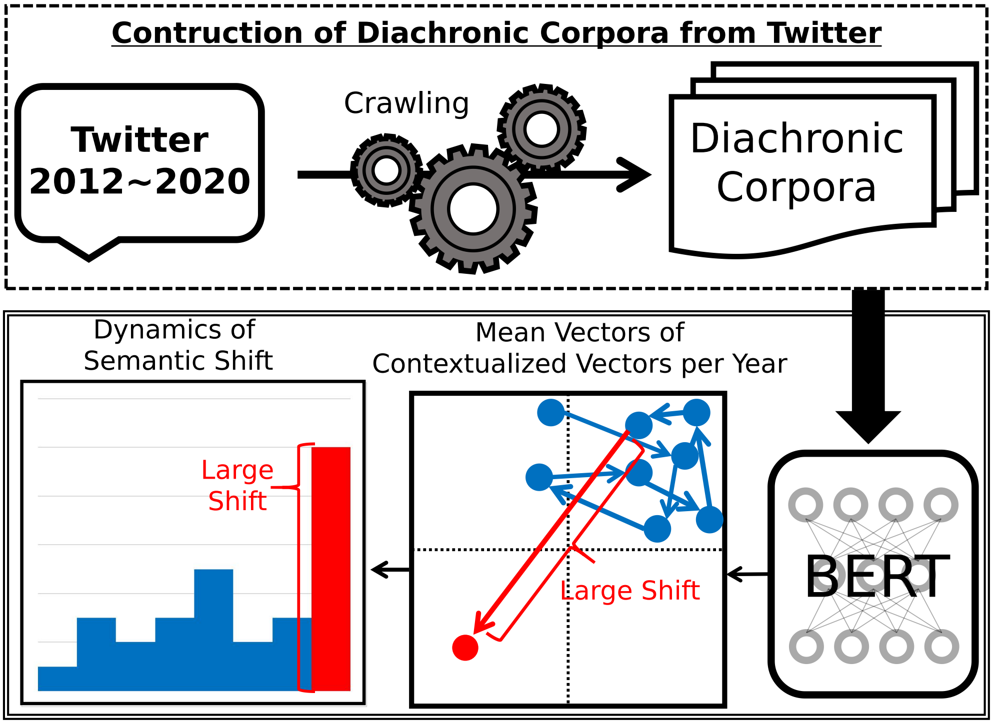

Figure 4.2: The experimental procedure of detecting the contextual change.

we used the *character tokenization-based version* of the pre-trained language model. By feeding sentences $\{Sent_1, Sent_2, \ldots, Sent_n\}$ including the target word w_i to the pre-trained language model, w_i's separate characters $\{c_1, c_2, \ldots, c_n\}$ representations $\{e_{w_i,c_1}, e_{w_i,c_2}, \ldots, e_{w_i,c_n}\}$ can be retrieved from the final hidden layer of the model. We simply summed up w_i's character representations and obtain w_i's token representations $\{e_{w_i,Sent_1}, e_{w_i,Sent_2}, \ldots, e_{w_i,Sent_n}\}$.

By repeating the same procedure for each year's sentences, we can get w_i's token representations for each year *t*. Then, we computed the average of w_i's token representations for each year and we use the mean vector $\overline{e}_t^{w_i}$ as the sense representation of w_i in that year *t* (Hu et al., 2019; Schuster et al., 2019).

$$\overline{e}_t^{w_i} = \frac{1}{m}\sum_{n=1}^{m} e_{t,Sent_n}^{w_i} \tag{4.1}$$

Language Model

We obtain contextualized word representations using two versions of the pre-trained BERT language model (Devlin et al., 2018). First is the Japanese BERT model (jBERT) distributed by Inui Laboratory of Tohoku University in Japan[1]. We chose the Japanese character tokenization based version due to the absence of many loanwords in the built-in dictionary of the tokenizer. This model is trained on Japanese Wikipedia using Whole-Word-Masking and the text is tokenized into characters. This model has 12-layer, 768-hidden, 12-heads, 110M parameters.

Second is the Multilingual BERT model (mBERT): *base-multilingual-cased version*[2]. Because of its excellent zero-shot cross-lingual model transfer capability (Pires et al., 2019), we select this multilingual model for the analysis both Japanese and Korean. This model is trained by Wikipedia in 104 languages and this model has 12-layer, 768-hidden, 12-heads, 110M parameters[3].

Diachronic Data

In this study, we use Twitter data for investigating contextual changes of loanwords according to changes in social trends. Twitter data is frequently used for studying changes in social trends (Atefeh and Khreich, 2015; Benhardus and Kalita, 2013; Mathioudakis and Koudas, 2010). Twitter data is also used in the study of semantic change (Kulkarni et al., 2015; Jawahar and Seddah, 2019). For these reasons, we

[1] `https://github.com/cl-tohoku/bert-japanese`

[2] `https://github.com/google-research/bert/blob/master/multilingual.md`

[3] We rely on Hugging Face's implementation of BERT (available at `https://github.com/huggingface/transformers`).

assumed that Twitter data will reflect the contextual change of loanwords caused by social trends change, thus we decided to use Twitter data in this study.

We crawled tweets from Twitter by using the Twint Python library.[4] Considering the comparison between Japanese and Korean, the target period was from 2012, when the official Twitter distribution service started in South Korea, to 2020. We randomly crawled tweets containing target words for a unit of a year and built a Twitter database for each year for each target word.

4.3.2 Tracking the Contextual Changes

After removing special characters (pictograms, numerical characters, and emoticons) from Twitter data, we randomly fed 200 tweets for each year to BERT and obtain the mean vector $\overline{e}_t^w$ of target words for each year. Although the cosine distance is often used when calculating the distance between vectors, referring to Reif et al. (2019) recommending the euclidean distance for visualization and measuring in the case of BERT, we use the standardized[5] Euclidean distance d between the mean vectors in the original dimension to track the semantic change of the target word like the equation 4.2. This distance represents the degree of contextual change according to the time change in this study. Figure 4.2 summarizes all these experimental procedures.

$$d\left(\overline{e}_{t+1}^w, \overline{e}_t^w\right) = \sqrt{\sum_{i=1}^{n} \left(\frac{\overline{e}_{t+1,i}^w - \overline{e}_{t,i}^w}{\sigma_i}\right)^2} \tag{4.2}$$

where σ_i denotes the standard deviation of ith components of mean vectors.

[4] An advanced twitter scraping tool is written in Python. The detailed information about the scraper is explained at `https://github.com/twintproject/twint`.

[5] Each word has a difference in the rate of distance change. The normal Euclidean distance has difficulty to compare the amount of change between words due to this difference of changing rate. Thus, we used standardized Euclidean distance to compare the real amount of changes.

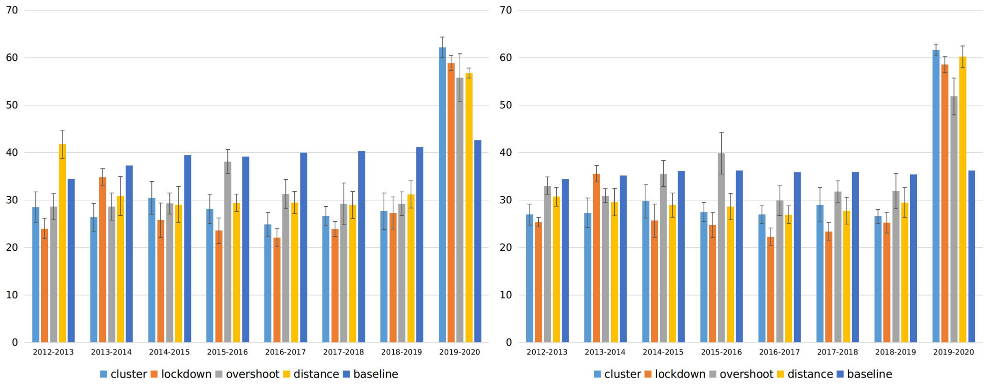

Figure 4.3: The distance of the mean vector in very other years in COVID-19 related Japanese loanwords.The left graph shows the result of jBERT and the right graph shows the result of mBERT.

4.3.3 Evaluation of Frame Work

In order to evaluate the validity of this tracking method, this section demonstrates a pilot experimental result in some Japanese loanwords, which context has obviously changed. We chose four COVID-19-related loanwords: クラスター *kurasuta* "cluster", ロックダウン *rokkudaun* "lockdown", オーバーシュート *obashuto* "overshoot", ディスタンス *deisutansu* "distance". Originally, each of these loanwords appears in several contexts, but since it was used in the context of COVID-19, these loanwords appear almost only in the limited context of COVID-19 in Japan. We attempt to show the validity of our method by testing whether our method can detect this sudden contextual change correctly.

For tracking this semantic change of loanwords, we calculate the target word mean vectors of each year following the procedure shown above and measure the standardized Euclidean distance between mean vectors. This same procedure was repeated ten times and we get average of standardized euclidean distance for each COVID-19-related loanwords. Figure 4.1 visualizes the shift of the mean vectors in a 2-dimensional vector space and Figure 4.3 summarizes the result.

Loanword	2012	2015	2019	2020
cluster	**goods (2)**	**goods (4)**	**game (6)**	**COVID-19 (9)**
	animation (2)	**game (4)**	goods (2)	education (1)
	game (1)	animation (1)	computer (1)	
	military (1)	internet (1)	leisure (1)	
	unknown (4)			
lockdown	**animation (4)**	**animation (8)**	**game (3)**	**COVID-19 (10)**
	game (2)	game (1)	**animation (2)**	
	unknown (4)	leisure (1)	**music (2)**	
			goods (1)	
			unknown (2)	
overshoot	**economy (8)**	**economy (7)**	**economy (10)**	**COVID-19 (8)**
	engineering (1)	game (2)		economy (2)
	military (1)	engineering (1)		
distance	**music (9)**	**music (9)**	**music (10)**	**COVID-19 (5)**
	unknown (1)	animation (1)		music (3)
				movie (1)
				leisure (1)

Table 4.1: The numbers of contexts in the 10 nearest sentences of the mean vector of COVID-related-loanwords in 2012, 2015, 2019, and 2020. "Unknown" means that the context cannot be interpreted from the contextual information.

Figure 4.3 indicates that all four loanwords have a large mean vector move between 2019 and 2020 in both Japanese BERT and Multilingual BERT. As some of the change is due to sampling and random drift, we additionally plot the average standardized distance changes of several words having more than 500 frequency in twitter against their reference points as a baseline in Figure 4.3. This allows us to detect whether a word's change during a given period is greater (or less) than would be expected from chance. Table 4.1 displays the proportion of the context of sentences which contains the nearest contextualized vectors (the nearest sentences) to the mean vectors in 2012, 2015, 2019, and 2020. In this study, from the content of the example sentences, we manually judged the kind of context in sentences.

The contexts of the nearest sentences to the mean vector of クラスター *kurasuta* "cluster" relate to *goods* (*cluster cristal*) and *game* in 2012, 2015 and 2019, but in 2020, 90% means *the mass infection by COVID-19*. The contexts of the nearest sentences to the mean vector of ロックダウン *rokkudaun* "lockdown" relate to *animation character* in 2012, 2015, 2019, but in 2020, 100% relate to *the shutdown of city buildings due to COVID-19*. The contexts of the nearest sentences to the mean vector of オーバーシュート *obashuto* "overshoot" relate to economical terms (*stock price surge*) from 2012 to 2019, but in 2020, 80% relate to *the outbreaks of COVID-19*. Finally, the contexts of the nearest sentences of ディスタンス *deisutansu* "distance" mainly relate to *a song title* from 2012 to 2019, but in 2020, 50% relate to the preventive measures during COVID-19 (*social distancing*).

This result indicates that our contextual change detection method properly detects the contextual changes of these four loanwords resulting from the recent COVID-19 outbreak. Figure 4.1 also visually shows the movement of the mean vector is large between 2019 and 2020.

4.3.4 Discussion for Framework

From the result of the movement of the mean vector in Figure 4.1, the results of the distance of the mean vector shift in Figure 4.3, and the context shift of the nearest sentences in Table 4.1, our method successfully detected the contextual change of loanwords due to the COVID-19 outbreak.

This result indicates that the contextualized word embeddings (BERT) can detect language change not only English (Hu et al., 2019; Giulianelli et al., 2020) but also Japanese. This result also shows that both Japanese BERT (Monolingual BERT) and Multilingual BERT can accurately capture the contextual change of Japanese. This indicates Multilingual BERT's high ability to analyze various languages accurately (Pires et al., 2019; Karthikeyan et al., 2019).

4.4 The Cultural Trend Change Analysis through Loanword Contextual Change Detection

In Section 4.3, the contextualized embedding model only focuses on the relationship between cultural trend changes and the contextual changes of loanwords after the COVID-19 outbreak. To analyze the cultural trend change and the contextual change of loanwords more broadly, we target more loanwords. This experiment reveals what cultural trend change has occurred at the point in the time a loanword's context changed. Through this experiment, we will verify whether the contextual change of the loanword can detect the social trend change.

4.4.1 Methodology

For the list of Japanese loanwords, we used "Suggestions for Paraphrasing Loanwords",[6] which was published by the National Institute for Japanese Language and Linguistics

[6]This document can be downloaded from `https://www2.ninjal.ac.jp/gairaigo/Teian1_4/iikae_teian1_4.pdf`

The Highest Distance in jBERT			The Highest Distance in mBERT		
Loanwords	Distance	Changed Time	Loanwords	Distance	Changed Time
biomass	52.44	2019-2020	lifeline	46.65	2018-2019
partnership	50.00	2014-2015	screening	45.28	2019-2020
lifeline	48.67	2018-2019	partnership	43.47	2014-2015

Table 4.2: The three highest contextual changed loanwords in jBERT and mBERT.

Language	Model	Loanwords
Japanese	jBERT	106 (words)
	mBERT	84 (words)
Korean	mBERT	69 (words)

Table 4.3: The total numbers of loanwords analyzed in this experiment.

in August 2006.

"Suggestions for Paraphrasing Loanwords" paraphrases some Japanese loanwords into clear Japanese native words. Conferences were held four times from 2003 to 2006 to make this list. It contains 173 pairs of Japanese loanwords and their corresponding native words. We assume that this list will provide Japanese loanwords that are frequently used in recent Japanese society. In the study of the contextual change in Korean loanwords, we translated the Japanese loanwords of this list into Korean. We used jBERT and mBERT to analyze the Japanese loanword and used mBERT to analyze the Korean loanword.

Firstly, we crawled the Twitter data for every loanword in the list. In the process of crawling, we found some loanwords were insufficient for the analysis because of their very low frequency: the crawler could not collect enough data. We removed those loanwords from the list of "Suggestions for Paraphrasing Loanwords." The tokenizing

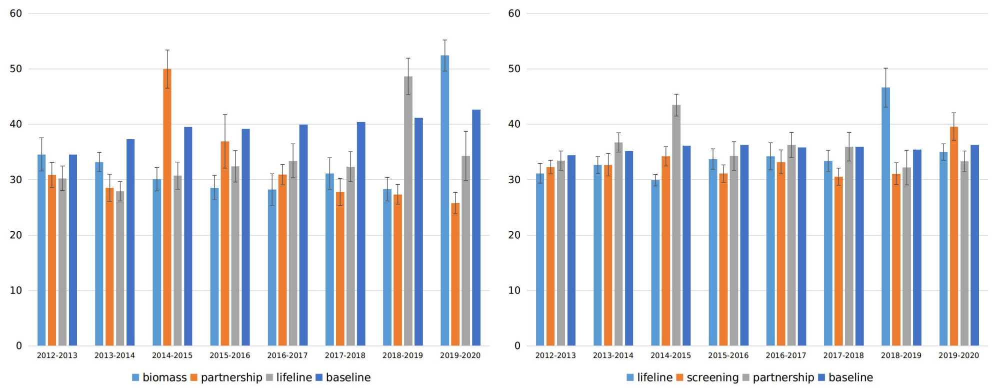

Figure 4.4: The distance of the mean vector in every other years in the top three highest contextual changed Japanese loanwords. Left is the loanwords in jBERT and right is the loanwords in mBERT.

process in BERT also removed some loanwords that were difficult to analyze due to the complexity of the tokenizing pattern. Finally, 106 Japanese loanwords remained in the jBERT analysis, and 84 Japanese loanwords and 68 Korean loanwords remained in the mBERT analysis. Table 4.3 summarizes the details of these experimental settings.

Secondly, we obtained the mean vectors from 2012 to 2020 for each loanword. We then calculated the standardized distance of the mean vector every other year and we checked the nearest sentences. This sentence checking process provides a qualitative analysis of which social trend shift made this contextual change at that time. The same procedure was performed for Japanese loanwords and Korean Loanwords. Tables 4.2 and Table 4.4 summarize the results.

4.4.2 Result and Discussion

The Contextual Change of Japanese Loanwords

For the 106 Japanese loanwords in jBERT and 84 loanwords in mBERT[7], we repeatedly calculated the distance of the mean vector shift 10 times for all loanwords and averaged

[7] As mentioned above, the difference in the number of loanwords that can be analyzed is due to the difference in the pattern of tokenization.

the results. Table 4.2 summarizes the top three Japanese loanwords with the highest distance values in jBERT and mBERT. The "Changed Time" in Table 4.2 means the time when their contextual changes occurred. Table 4.5 shows the contextual change of the nearest sentences of the mean vector. In jBERT, the contextual change of バイオマス *baiomasu* "biomass" between 2019 and 2020 is the largest, followed by is パートナーシップ *patonashippu* "partnership" between 2014 and 2015, and ライフライン *raifurain* "lifeline" from 2018 to 2019.

Checking the nearest sentences reveals that the contextual change of バイオマス *baiomasu* "biomass" was triggered by Japan's new plastic bag charge that started on July 1, 2020. This law requires no charge for the biomass shopping bags, and "biomass" frequently appears in the context of Japan's new plastic bag charge. The distance of mean vector between 2019 and 2020 indicates this cultural trend change.

Checking the nearest sentences of "partnership" revealed the new system about homosexual partnerships triggered the contextual change of パートナーシップ *patonashippu* "partnership" in 2015. In the case of ライフライン *raifurain* "lifeline", a character name in a new computer game triggered a contextual change in 2019. As a result of the Tukey test for distance values, all the highest distance is significantly greater ($p < 0.0001$) than other year's distances in all three loanwords.

In the case of mBERT, パートナーシップ *patonashippu* "partnership" and ライフライン *raifurain* "lifeline" are ranked high, and checking the nearest sentences shows similar contextual change with jBERT. The スクリーニング *sukuriningu* "screening" has the second largest distance in mBERT. The context of スクリーニング *sukuriningu* "screening" has shifted to the context of searching for COVID-19 infected persons like the COVID-19-related loanwords in Section 4.3.3. As a result of the Tukey test for distance values, all the highest distance is significantly greater ($p < 0.0001$) than other year's distances in all three loanwords.

From these results, our contextual change detection model properly detects the contextual changes not only of COVID-19 loanwords but also various others. This

The Highest Korean Loanwords in mBERT		
Loanwords	Distance	Changed Time
partnership	55.19	2018-2019
share	53.29	2013-2014
operation	53.07	2012-2013

Table 4.4: The three highest contextual changed Korean loanwords in mBERT.

result successfully shows that the contextual change of loanwords can be one of the indicators of detecting social trend changes in Japanese society. Table 4.5 summarizes these contextual changes briefly.

The Contextual Change of Korean Loanwords

Table 4.4 shows the results in Korean. The loanword with the largest distance is 파트너십 *phathunesip* "partnership", followed by 셰어 *syeye* "share" and 오퍼레이션 *opheleyisyen* "operation". Checking the nearest sentences of these loanwords revealed that the appearance of a new cartoon, a new game, a new TV program triggered the contextual changes of these loanwords. As a result of the Tukey test for distance values, all the highest distance is significantly greater ($p < 0.0001$) than other year's distances in all three loanwords.

These results indicate the possibility of analyzing cultural trend change by the contextual change detection method even in Korean. In Korean, cultural trend changes, such as animation and games, mainly triggered the contextual changes of loanwords.

These results will support that the contextualized embedding model is useful not only for detecting the contextual changes in loanwords but also for understanding the cultural trend changes.

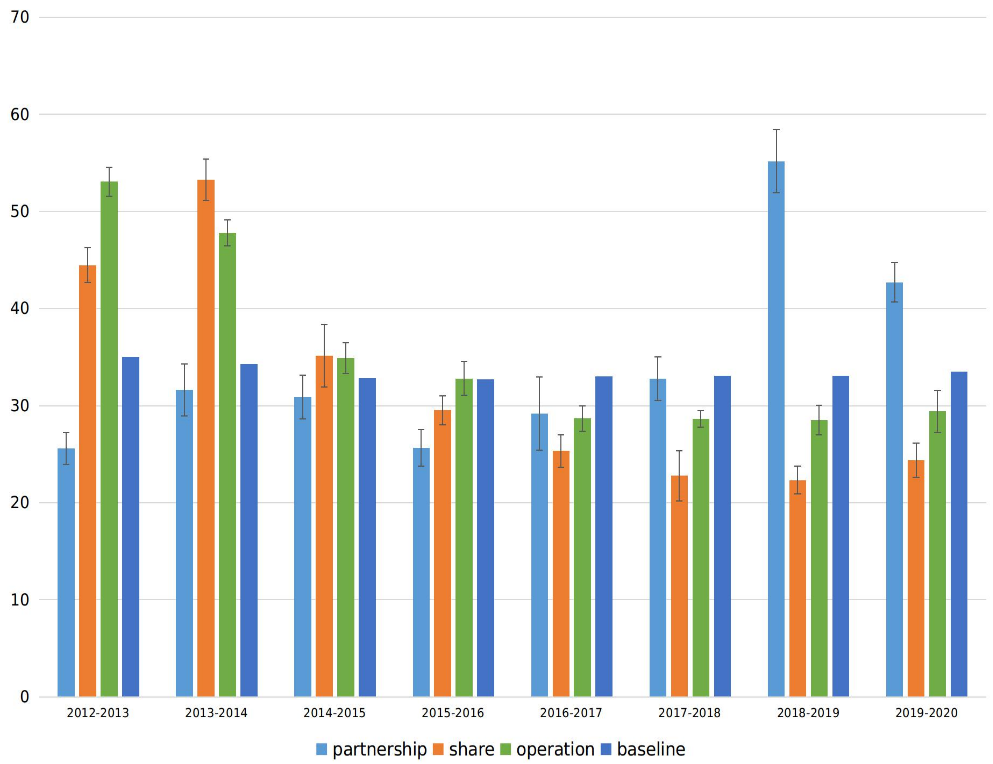

Figure 4.5: The distance of the mean vector in very other years in the top three highest contextual changed Korean loanwords in mBERT.

Language	Model	Loanwords	Old Dominant Context	New Dominant Context
		biomass	Energy	Japan’s new plastic bag charge
Japanese	jBERT	partnership	International relations	System about homosexual relationship
		lifeline	Utilities	Game
		lifeline	Utilities	Game
Japanese	mBERT	screening	Man and female relations	COVID-19
		partnership	Business relations	Man and female relation
		partnership	Business relations	Animation
Korean	mBERT	share	Movie	TV program
		operation	Game	Animation

Table 4.5: The contextual changes of the three highest contextual changed loanwords at the most changed times.

4.5 Conclusion and Future Work

This study indicates that the contextualized embeddings model can detect the diachronic contextual change of loanwords in Japanese and Korean. After evaluating the performance of the model with COVID-19-related Japanese loanwords, we analyzed the social trend change in Japanese and Korean by the contextual change detection of loanwords.

In Japanese, the changes in social systems like a change of law and changes in cultural trends such as games relate to the contextual change of loanwords. In Korean, the changes in cultural trends such as games and animation mainly relate to the contextual change of loanwords. This result indicates the close relationship between the contextual change of loanwords and the change of society and culture. This result also suggests a method of analyzing cultural trend changes through the detection of the contextual changes of the loanword

This method has the advantage of quickly and automatically detecting a cultural trend change from language data. In future research, if we apply this method for more loanwords, we can find the cultural trend changes more quickly and more comprehensively.

Additionally, by tracking changes in loanwords, we can expect to find not only cultural trend changes that have occurred in the past but also cultural trend changes that are occurring now and to predict future cultural trend changes.

This study also indicates the feasibility of multilingual BERT for detecting language changes in Japanese and Korean. In the future, targeting more and varied other languages will greatly contribute to the development of comparative linguistics and comparative sociology.

Chapter 5

Conclusion and Future Works

5.1 Summary

Most previous researches summarize a pattern by showing several examples of loanwords in various languages. Although corpus linguistics has conducted frequency-based loanword researches, quantitative analysis for the complex semantic phenomena of loanword remains undeveloped.

For overcoming these obstacles, this dissertation uses word embedding-based methods in the semantic study of loanwords. We propose computational and quantitative methods to study the semantic phenomena that loanwords undergo in the process of integration and adaptation into the recipient language. Additionally, we propose the methods for detecting the cultural trend change through the contextual change of loanword.

Chapter two investigates the lexical competition between the loanword and the native synonyms in the process of integration and adaptation of loanwords. The conventional frequency-based method can not show what kind of lexical competition is happening. Judging the kind of lexical competition–*Word replacement*, or *Semantic differentiation*–requires the contextual relationship information between the loanword and the native synonyms. The vector space of word embedding and the geometrical concept (the over-lapping circle) enables quantitative modeling of this shared context relationship. This context-sharing relational model quantitatively reveals whether the loanword-synonym pair has a relationship of word replacement or semantic differentiation at that time.

Chapter three investigates the semantic adaptation of English loanwords in Japanese and Korean by comparing the meaning of loanwords and the original English words. Comparing the vector values of the vector space obtained from different databases is impossible directly. Using the transformation matrix enables comparing the vector value of the different vector spaces. This methodology revealed the semantic adaptation of English loanwords in Japanese and Korean. This study also conducts a contrastive study of the difference in the semantic adaptation of English loanwords between Japanese and Korean. Additionally, we statistically verified the very common semantic adaptation pattern of loanwords: polysemous English word has a limited meaning when used as a loanword.

Chapter four focuses on the social semantic role of loanwords which reflects the trend of culture. Analyzing the contextual change of loanwords detects the cultural change which happened at that time. The conventional word embedding methods require a large amount of diachronic corpus as training data, thus studying diachronic meaning changes over time has been difficult in resource-poor languages such as Japanese and Korean. This study uses the pre-trained contextual embeddings model (BERT) to overcome this obstacle and detects the contextual changes of loanwords happening using Twitter data. Analyzing this contextual change find the cultural trend change occurred at that time. These results prove that loanwords work as indicators that reflect cultural trends, and that tracking the contextual changes of loanwords can detect the cultural trend change.

5.2 Future Works

This dissertation proposes several quantitative methods for analyzing the semantic phenomenon of loanwords and assessed the validity of the semantic analysis of loanwords using this method in Japanese and Korean. The following researches are expected in the future using these quantitative methods.

5.2.1 Revealing Statistical Law

This quantitative semantic analyzing method opens the way for statistical analysis of several factors that will have influences on semantic phenomena. As possible factors, referring to what Winter-Froemel et al. (2014) claims, are like below.

1. age of borrowing (when the loanword entered the recipient language)
2. relative word length of loanword compared to the native synonym
3. phonological markedness (whether or not the sound of the loanword cohere the phonological system of the recipient language)
4. graphemic markedness (how well the spelling fits with the recipient language writing system)
5. markedness of phonemic-graphemic correspondence (how well the loanword correspond to the recipient language rules of spelling and pronunciation)
6. lexical field

Although quantifying these factors remains a future challenge, quantifying these factors will statistically verify the effects of these factors on lexical competition and semantic adaptation. These statistical analyses will reveal the statistical laws related to the semantic phenomena of loanwords.

5.2.2 Computational Contrastive Linguistic Study

Although this dissertation only focuses on English loanwords in Japanese and Korean, the proposed methods in this dissertation can analyze any other language. Thus, using various language data will reveal the differences in the semantic phenomena of loanwords between several languages. Particularly contrastive studies between languages that differ in culture and linguistic systems, such as Asian and European languages, will give new insights into the semantic phenomena of loanwords. Additionally, although this dissertation conducted research on English loanwords used commonly all over the world, future works analyzing other language's loanword semantic phenomena will produce interesting results.

5.2.3 Application to Other Semantics Tasks

Although this dissertation focuses on the semantic phenomena of loanwords only, the proposed methods in this dissertation can probably also analyze the semantic phenomena of non-loanwords. A lot of semantic researches has investigated several semantic phenomena in history. However, as with loanword research, computational semantic studies using big language data and deep learning methods remain unexplored. The word embedding based-methods proposed in this dissertation will produce interesting results in the analysis of the several semantic phenomena not only loanwords. Additionally, comparing the result of the semantic phenomena of non-loanwords with the result of loanwords will provide new insight into the contrastive linguistic studies.

As mentioned above, we hope that the word embedding-based semantic analysis method developed in this dissertation will bring great progress to future semantics research.

Bibliography

Rod Tyson. English loanwords in Korean: Patterns of borrowing and semantic change. *Journal of Second Language Acquisition and Teaching*, 1:29–36, 1993.

Gillian Kay. English loanwords in Japanese. *World Englishes*, 14(1):67–76, 1995.

Myung-hee Noh. Aspects of semantic shift in loanwords. *Daedong Institute for Korean Studies*, 82:493–524, 2013.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013a.

Edward Sapir. Language: An introduction to the study of speech, 1921.

Shana Poplack and David Sankoff. Borrowing: the synchrony of integration. 1984.

Holger Pedersen and John Webster Spargo. The discovery of language. 1965.

Otto Jespersen. *Growth and structure of the English language*. BG Teubner, 1905.

Leonard Bloomfield. Language. new york: Henry holt & co, 1933.

Einar Haugen. The analysis of linguistic borrowing. *Language*, 26(2):210–231, 1950.

U Weinreich. Languages in contact: Findings and problems. 1954.

Winfred Philipp Lehmann et al. Historical linguistics. 1962.

William Labov. The social stratification of English in New York City. Washington, DC: Center for applied linguistics.. 1972. sociolinguistic patterns. *Philadelphia: University of Pennsylvania Press.(1972b). Language in the inner city: Studies in the Black English Vernacular. Philadelphia: University of Pennsylvania Press.(1973). The linguistic consequences of being a lame. Language in Society*, 2:81–1, 1966.

Raimo Anttila. Historical and comparative linguistics. 1989.

Nils Hasselmo. Code-switching and modes of speaking. *Texas Studies in Bilingualism*, pages 179–210, 1970.

Michael G. Clyne. Some (German-English) language contact phenomena at the discourse level. *Studies for Einar Haugen*, pages 132–144, 1972.

Bates L Hoffer. Language borrowing and language diffusion: An overview. *Intercultural communication studies*, 11(4):1–37, 2002.

Martin Haspelmath and Uri Tadmor. *Loanwords in the World's Languages*. De Gruyter Mouton, 2009.

Gisle Andersen, Cristiano Gino Furiassi, Biljana Mišić Ilić, et al. The pragmatic turn in studies of linguistic borrowing. 2017.

Elizabeth Peterson and Kristy Beers Fägersten. Introduction to the special issue: Linguistic and pragmatic outcomes of contact with English. *Journal of Pragmatics*, 133:105–108, August 2018. ISSN 0378-2166. doi: 10.1016/j.pragma.2018.06.005.

Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile networks and applications*, 19(2):171–209, 2014.

Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

Tingyu Lu. Analysis on linguistics research directions in the age of big data. *Journal of Physics: Conference Series*, 1606:012008, aug 2020. doi: 10.1088/1742-6596/1606/1/012008. URL `https://doi.org/10.1088%2F1742-6596%2F1606%2F1%2F012008`.

Jacqueline Rae Larsen Serigos et al. *Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish*. PhD thesis, 2017.

Dwight Bolinger. Meaning and form. 1977.

Esme Winter-Froemel, Alexander Onysko, and Andreea Calude. Why some non-catachrestic borrowings are more successful than others: a case study of English loans in German. *Language contact around the globe*, pages 119–142, 2014.

Eline Zenner, Laura Rosseel, and Andreea Simona Calude. The social meaning potential of loanwords: Empirical explorations of lexical borrowing as expression of (social) identity. *Ampersand*, 6:100055, 2019.

Manjola Bregasi. The language economy principle in Albanian syntax. *ANGLISTICUM. Journal of the Association-Institute for English Language and American Studies*, 5 (4):31–37, 2016.

André Martinet. Economie des changements phonétiques. 1955.

Alessandra Vicentini. The economy principle in language. *Notes and Observations from early modern English grammars. Mots, Palabras, Words*, 3:37–57, 2003.

George K Zipf. The principle of least effort: An introduction to human ecology, 1949.

Xavier Gabaix. Zipf's law for cities: an explanation. *The Quarterly journal of economics*, 114(3):739–767, 1999.

Lada A Adamic and Bernardo A Huberman. Zipf's law and the internet. *Glottometrics*, 3(1):143–150, 2002.

Pavlos Pavlou. The semantic adaptation of Turkish loan-words in the Greek cypriot dialect. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 443–443, 1994.

Lauren Asia Hall-Lew. English loanwords in mandarin Chinese. *The University of Arizone*, 2002.

Anwar AH Al-Athwary. The semantics of English borrowings in Arabic media language: The case of Arab gulf states newspapers. *International Journal of Applied Linguistics and English Literature*, 5(4):110–121, 2016.

Alexander Onysko and Esme Winter-Froemel. Necessary loans–luxury loans? exploring the pragmatic dimension of borrowing. *Journal of pragmatics*, 43(6):1550–1567, 2011.

Dennis R Preston. The influence of regard on language variation and change. *Journal of Pragmatics*, 52:93–104, 2013.

James Stanlaw. English in Japanese communicative strategies. *The other tongue: English across cultures*, 2:178–208, 1992.

Harald Haarmann. The role of German in modern Japanese mass media: Aspects of ethnocultural stereotypes and prestige functions of language in Japanese society. *Hitotsubashi journal of social studies*, pages 31–41, 1984.

Leo J Loveday. Japanese sociolinguistics: An introductory survey. *Journal of Pragmatics*, 10(3):287–326, 1986.

Mark Rebuck. The function of English loanwords in Japanese. *NUCB journal of language culture and communication*, 4(1):53–64, 2002.

Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

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

Patrick Pantel. Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 125–132, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219856. URL `https://www.aclweb.org/anthology/P05-1016`.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P10-1040`.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013b.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to lexical semantic change. *arXiv preprint arXiv:1811.06278*, 2018.

Yang Xu and Charles Kemp. A computational evaluation of two laws of semantic change. In *CogSci*, 2015.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

Esme Winter-Froemel. The pragmatic necessity of borrowing. *Taal en Tongval*, 69(1): 17–46, 2017.

Eline Zenner, Dirk Speelman, and Dirk Geeraerts. Cognitive sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics*, 23(4):749–792, 2012.

Naomi Lapidus Shin. Efficiency in lexical borrowing in New York Spanish. *International Journal of the Sociology of Language*, 2010(203):45–59, 2010.

Andreea Simona Calude, Steven Miller, and Mark Pagel. Modelling loanword success– a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*, 1(ahead-of-print), 2017.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493, 2018.

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

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access, 2016b.

Renfen Hu, Shen Li, and Shichen Liang. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, 2019.

Key-Sun Choi, Hee-Sook Bae, Wonseok Kang, Juho Lee, Eunhe Kim, Hekyeong Kim, Donghee Kim, Youngbin Song, and Hyosik Shin. Korean-Chinese-Japanese multilingual wordnet with shared semantic hierarchy. In *LREC*, 2004.

George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Akihiko Yamada and Hyopil Shin. Applying word embeddings to measure the semantic adaptation of English loanwords in Japanese and Korean. *Language Research*, 53(3): 473–500, 2017.

Daisuke Okawa. A study on degree of recognition about Japanese-style English in Korean. *Journal of North-East Asian cultures*, 14:499–523, 2008.

Seung Mi Cheon. A study of English loanwords in Korean. *Korean Studies Information*, 2008.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *WWW*, 2015.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. Analyzing semantic change in Japanese loanwords. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1195–1204, 2017.

AS Fenogenova, IA Karpov, VI Kazorin, and IV Lebedev. Comparative analysis of anglicism distribution in Russian social network texts. *Kompjuternaja Lingvistika i Intellektualnye Tehnologii*, page 65, 2017.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 230–237, 2004.

Hyun-sik Min. A study on the foreign words of Korean language. *Korean Semantics*, 2: 91–132, 1998.

Sen Nishiyama. Speaking English with a Japanese mind. *World Englishes*, 14(1):27–36, 1995.

André Rattinger, Jean-Marie Le Goff, and Christian Guetl. Local word embeddings for query expansion based on co-authorship and citations. 2018.

AP Ajees and Sumam Mary Idicula. A named entity recognition system for malayalam using neural networks. *Procedia computer science*, 143:962–969, 2018.

Michael Mayor. *Longman dictionary of contemporary English*. Pearson Education India, 2009.

Frank E Daulton. The creation and comprehension of English loanwords in the Japanese media. *Journal of Multilingual and Multicultural Development*, 25(4):285–296, 2004.

Takako Tomoda. The impact of loan-words on modern Japanese. In *Japan Forum*, volume 11, pages 231–253. Taylor & Francis, 1999.

Sofia Rüdiger. Mixed feelings: Attitudes towards English loanwords and their use in South Korea. *Open Linguistics*, 4(1):184–198, 2018.

Rosa Jinyoung Shim. Englishized Korean: Structure, status, and attitudes. *World Englishes*, 13(2):225–244, 1994.

Jae Jung Song. English in South Korea revisited via Martin Jonghak Baik (1992, 1994), and Rosa Jinyoung Shim (1994). *World Englishes*, 17(2):263–271, 1998.

C Bruce Lawrence. The verbal art of borrowing: Analysis of English borrowing in Korean pop songs. *Asian Englishes*, 13(2):42–63, 2010.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*, 2018.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings* *of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.365. URL `https://www.aclweb.org/anthology/2020.acl-main.365`.

Ganesh Jawahar and Djamé Seddah. Contextualized diachronic word representations. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 35–47, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4705. URL `https://www.aclweb.org/anthology/W19-4705`.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman, and Slav Petrov. Syntactic annotations for the google books ngram corpus. 2012.

Mark Davies. The corpus of historical American English: 400 million words, 1810-2009, 2010.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, 2019.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, 2019.

Farzindar Atefeh and Wael Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.

James Benhardus and Jugal Kalita. Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9(1):122–139, 2013.

Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158, 2010.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems*, pages 8594–8603, 2019.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*, 2019.

Chapter A

List of Loanword Having One Synset and One Definition in Korean CoreNet in Chapter 2

지퍼	ciphe	요구르트	yokwuluthu	요가	yoka
비타민	pithamin	바이러스	pailesu	바이올리니스트	paiollinisuthu
터키탕	thekhithang	터널	thenel	티셔츠	thisyechu
톨루엔	thollwueyn	티타늄	thithanyum	타이밍	thaiming
테크닉	theykhunik	택시	thayksi	달란트	tallanthu
스페어	supheye	소나타	sonatha	소프트	sophuthu
루머	lwume	룰	lwul	럭비	lekpi
레지	leyci	레코드	leykhotu	리어카	liekha
프린터	phulinthe	프라이드	phulaitu	포스터	phosuthe
파트너	phathune	빨치산	ppalchisan	파카	phakha
온라인	onlain	너트	nethu	뉴	nyu
모르핀	moluphin	모델하우스	moteylhawusu	모델	moteyl
멤버	meympe	메가폰	meykaphon	밀공모선	milkongmosen
매니큐어	maynikhyue	마곡	makok	마르	malu
주장자	cwucangca	지단채	citanchay	예수교	yeyswukyo
힌트	hinthu	힐	hil	하이힐	haihil
에로	eylo	에아	eya	돌리	tolli
샹송	syangsong	샴페인	syampheyin	첼로	cheyllo
버디	peti	베레모	peyleymo	베드	peytu
알레르기	alleyluki	에어쇼	eyesyo	아카데미상	akhateymisang
디자이너	ticaine	데모	teymo	유니폼	yuniphom
모텔	motheyl	호텔	hotheyl	알리바이	allipai
미네랄	mineylal	멜로디	meylloti	말라리아	mallalia
인터페론	inthepheylon	호르몬	holumon	홈런	homlen
마가린	makalin	캥거루	khayngkelwu	인턴	inthen
엔릴	eynlil	실린더	sillinte	칼슘	khalsyum
네온	neyon	메가톤	meykathon	마라톤	malathon
카페인	khapheyin	버튼	pethun	범퍼	pemphe
디지털	ticithel	라운지	lawunci	차임벨	chaimpeyl
에스컬레이터	eysukhelleyithe	유스호스텔	yusuhosutheyl	숄	syol
시뮬레이션	simyulleyisyen	가든골프	katunkolphu	선글라스	senkullasu
드레스	tuleysu	디프테리아	tiphutheylia	카리스마	khalisuma
케이블카	kheyipulkha	애드벌룬	aytupellwun	샐러드	saylletu
브랜디	pulaynti	본드	pontu	밴드	payntu
코너킥	khonekhik	칵테일	khaktheyil	보트	pothu
클로버	khullope	캡슐	khaypsyul	크레인	khuleyin
드링크	tulingkhu	센트	seynthu	시멘트	simeynthu
리셉션	liseypsyen	옵션	opsyen	에스트로겐	eysuthulokeyn
로켓	lokheys	라켓	lakheys	포켓	phokheys
자장면	cacangmyen	도사견	tosakyen	스파게티	suphakeythi

양키	yangkhi	요트	yothu	라이터	laithe
바이올린	paiollin	비닐하우스	pinilhawusu	비닐	pinil
트렁크	thulengkhu	트럼펫	thulempheys	트럭	thulek
티켓	thikheys	테마	theyma	메시아	meysia
탤런트	thayllenthu	스웨터	suweythe	스모	sumo
슬로건	sulloken	슬리퍼	sulliphe	스키장	sukhicang
루비	lwupi	루블	lwupul	라운드	lawuntu
레이저	leyice	레이크	leyikhu	레이더	leyite
폴리스	phollisu	포커	phokhe	플래카드	phullaykhatu
파라솔	phalasol	팬티	phaynthi	페어	pheye
노벨상	nopeylsang	노코멘트	nokhomeynthu	니스	nisu
모빌	mopil	미스터	misuthe	미사일	misail
매트리스	maythulisu	마스터	masuthe	매스미디어	maysumitie
라인	lain	라이거	laike	레닌주의	leynincwuuy
이소효소	isohyoso	아이언	aien	인터뷰	inthepyu
히트	hithu	헤어스타일	heyesuthail	껌	kkem
데뷔	teypwi	카레	khaley	커브	khepu
카타르시스	khathalusisu	카로틴	khalothin	캔버스	khaynpesu
바트	pathu	발코니	palkhoni	백미러	paykmile
아카시아	akhasia	앰프	aymphu	피겨	phikye
유엔	yueyn	유턴	yuthen	토네이도	thoneyito
로열티	loyelthi	릴레이	lilleyi	밀리미터	millimithe
킬로미터	khillomithe	고릴라	kolilla	드릴	tulil
프레임	phuleyim	이어폰	iephon	디자인	ticain
센티미터	seynthimithe	맨홀	maynhol	카운슬러	khawunsulle
알루미늄	allwuminyum	앨범	aylpem	아드레날린	atuleynallin
매너리즘	maynelicum	레슨	leysun	이리듐	ilityum
벤젠	peynceyn	바텐더	patheynte	안테나	antheyna
정글	cengkul	린치	linchi	침팬지	chimphaynci
콜레스테롤	kholleysutheylol	올리브	ollipu	골프	kolphu
패션모델	phaysyenmoteyl	시스템	sisutheym	쇼핑	syophing
아스파라거스	asuphalakesu	코브라	khopula	바리케이드	palikheyitu
길드	kiltu	글러브	kullepu	에메랄드	eymeylaltu
배드민턴	paytuminthen	블라우스	pullawusu	블루스	pullwusu
네트워크	neythuwekhu	수프	swuphu	옵서버	opsepe
카운트	khawunthu	클레임	khulleyim	콘크리트	khonkhulithu
지프	ciphu	초콜릿	chokhollis	점프	cemphu
징크스	cingkhusu	크리스천	khulisuchen	립스틱	lipsuthik
논스톱	nonsuthop	헤마토크릿	heymathokhulis	세라믹	seylamik
메탄가스	meythankasu	프락치	phulakchi	허리케인	helikheyin

레슬링	leysulling	와이퍼	waiphe	윙크	wingkhu
비브리오	pipulio	베란다	peylanta	벨벳	peylpeys
트레이닝	thuleyining	트레일러	thuleyille	트레이드	thuleyitu
매트릭스	maythuliksu	테스트	theysuthu	테러리스트	theylelisuthu
스프	suphu	수드라	swutula	스트라이커	suthulaikhe
스케이팅	sukheyithing	사인펜	sainpheyn	시나리오	sinalio
로프	lophu	론도	lonto	로스	losu
피라미드	philamitu	펄프	phelphu	프로펠러	phulopheylle
피자	phica	핑크	phingkhu	파인애플	phainayphul
페이지	pheyici	패키지	phaykhici	페이스	pheyisu
니코틴	nikhothin	노이로제	noilocey	넥타이	neykthai
미스코리아	misukholia	밍크	mingkhu	미니스커트	minisukhethu
매스컴	maysukhem	매스	maysu	마르크스주의	malukhusucwuuy
키스	khisu	카투사	khathwusa	카르	khalu
인터체인지	inthecheyinci	인스턴트	insuthenthu	인플루엔자	inphullwueynca
그랑프리	kulangphuli	프라이팬	phulaiphayn	프리랜서	phulilaynse
컨디션	khentisyen	콩쿠르	khongkhwulu	코미디	khomiti
캔	khayn	캠페인	khaympheyin	캐디	khayti
바벨	papeyl	오토바이	othopai	아세안	aseyan
토플	thophul	토스트	thosuthu	템포	theympho
투피스	thwuphisu	스토리	sutholi	스피커	suphikhe
밀리리터	millilithe	거들	ketul	포플러	phophulle
콜레라	kholleyla	알로에	alloey	알칼리	alkhalli
컨테이너	khentheyine	코카인	khokhain	코팅	khothing
볼링	polling	바겐세일	pakeynseyil	플루토늄	phullwuthonyum
카메라맨	khameylamayn	심포니	simphoni	사포닌	saphonin
이온	ion	이닝	ining	펜싱	pheynsing
안포폭약	anphophokyak	에이즈	eyicu	주니어	cwunie
잡	cap	러시아워	lesiawe	아이스하키	aisuhakhi
쇼윈도	syowinto	로션	losyen	카네이션	khaneyisyen
렌즈	leyncu	아스피린	asuphilin	치즈	chicu
워드프로세서	wetuphuloseyse	야드	yatu	워드	wetu
그라운드	kulawuntu	브리핑	puliphing	킬로그램	khillokulaym
지그재그	cikucayku	헤드	heytu	브래지어	pulaycie
헥타르	heykthalu	나이트클럽	naithukhullep	라일락	laillak
콘서트	khonsethu	컨트롤	khenthulol	클랙슨	khullayksun
재킷	caykhis	잭	cayk	스코프	sukhophu
크리스트교	khulisuthukyo	크리스마스	khulisumasu	박스	paksu
박테리아	paktheylia	팝송	phapsong	엘리베이터	eyllipeyithe
포졸	phocol	기혼	kihon	샌드백	sayntupayk

위스키	wisukhi	웨딩드레스	weytingtuleysu	와트	wathu
바캉스	pakhangsu	유토피아	yuthophia	우라늄	wulanyum
트랙터	thulaykthe	트랙	thulayk	타월	thawel
테러	theyle	텐트	theynthu	테니스	theynisu
스테로이드	sutheyloitu	스탠더드	suthayntetu	스타디움	suthatiwum
스캔들	sukhayntul	사무라이	samwulai	삼바	sampa
리타	litha	류머티즘	lyumethicum	레스토랑	leysutholang
프로판	phulophan	프로젝트	phuloceykthu	프로그래머	phulokulayme
피아니스트	phianisuthu	필로폰	phillophon	페스트	pheysuthu
오존	ocon	오리지널	olicinel	오케스트라	okheysuthula
나트륨	nathulyum	나단	natan	냅킨	naypkhin
마이크	maikhu	미그	miku	미터	mithe
마리화나	malihwana	매뉴얼	maynyuel	망토	mangtho
카페	khaphey	칼륨	khallyum	점퍼	cemphe
이데올로기	iteyolloki	유머	yume	하키	hakhi
프랑	phulang	포르말린	pholumallin	피신	phisin
코미디언	khomitien	칼럼	khallem	크롬	khulom
카바레	khapaley	버너	pene	바바리	papali
아르바이트	alupaithu	아날로그	analloku	아메바	ameypa
사우나	sawuna	사파이어	saphaie	리비도	lipito
리터	lithe	드라마	tulama	커서	khese
플라타너스	phullathanesu	플랜트	phullaynthu	플랑크톤	phullangkhuthon
다큐멘터리	takhyumeyntheli	보건	poken	심포지엄	simphociem
베이컨	peyikhen	아나운서	anawunse	암모니아	ammonia
리허설	lihesel	메탄올	meythanol	빌리루빈	pillilwupin
라돈	laton	프리즘	phulicum	핑퐁	phingphong
덤핑	temphing	더빙	teping	콘돔	khontom
스포츠카	suphochukha	스포츠	suphochu	소시지	sosici
보너스	ponesu	스위퍼	suwiphe	원피스	wenphisu
애니메이션	aynimeyisyen	아이스크림	aisukhulim	샤머니즘	syamenicum
가스레인지	kasuleyinci	재즈	caycu	체스	cheysu
하드웨어	hatuweye	큐피드	khyuphitu	스피드	suphitu
핸드볼	hayntupol	핸드백	hayntupayk	마그네슘	makuneysyum
위트	withu	튤립	thyullip	톨게이트	tholkeyithu
크레졸	khuleycol	슬립	sullip	올림픽	ollimphik
싱크대	singkhutay	레지던트	leycitenthu	팝콘	phapkhon
프리킥	phulikhik	스커트	sukhethu	리더십	litesip
비스킷	pisukhis	핫도그	hastoku	그룹	kulwup
콤플렉스	khomphulleyksu	인테리어	intheylie	쿠데타	khwuteytha
그린벨트	kulinpeylthu	쇼핑센터	syophingseynthe	해프닝	hayphuning

웨이터	weyithe	비타민제	pithamincey
유네스코	yuneysukho	차르	chalu
터치	thechi	토템	thotheym
텔렉스	theylleyksu	텔레비전	theylleypicen
스파이	suphai	스포트라이트	suphothulaithu
살모넬라균	salmoneyllakyun	샐러리맨	sayllelimayn
렌터카	leynthekha	르네상스	luneysangsu
프로필	phulophil	프라이버시	phulaipesi
파마	phama	패스포트	phaysuphothu
오랑우탄	olangwuthan	오페라	opheyla
마운드	mawuntu	모터보트	mothepothu
메스	meysu	메뉴	meynyu
마네킹	maneykhing	만나	manna
조깅	coking	요오드	yootu
히로뽕	hiloppong	히프	hiphu
팡파르	phangphalu	에티켓	eythikheys
치킨	chikhin	카오스	khaosu
브라운	pulawun	브라운관	pulawunkwan
아미노산	aminosan	앰뷸런스	aympyullensu
호스	hosu	드라이	tulai
라이벌	laipel	플레이보이	phulleyipoi
니켈	nikheyl	머플러	mephulle
네온사인	neyonsain	인터폰	inthephon
컴퓨터	khemphyuthe	미팅	mithing
샘플	saymphul	글리코겐	kullikhokeyn
페니실린	pheynisillin	패턴	phaythen
세슘	seysyum	센터링	seyntheling
저널	cenel	젤리	ceylli
이슈	isyu	버스	pesu
샤먼	syamen	프레온	phuleyon
레저	leyce	에스테르	eysutheylu
시드	situ	리그	liku
그램	kulaym	카드뮴	khatumyum
파일럿	phailles	다이너마이트	tainemaithu
헬리콥터	heyllikhopthe		
잉크	ingkhu		
클라이맥스	khullaimayksu		
블랙홀	pullaykhol		
다이	tai		
팩시밀리	phayksimilli		

초 록

전 세계적으로 활발한 문화 교류가 이루어짐에 따라 외래어가 일반적으로 자주 사용되는데, 외래어의 수용 과정에서 다양한 언어적 현상이 일어난다. 외래어가 수용됨에 따라 원래 차용주에 존재했던 단어가 사라지기도 하고, 차용어의 접미사와 단어가 차용주의 단어와 결합하여 새로운 단어를 생성하기도 하며, 차용어의 전치사가 외래어로서 그대로 사용되기도 한다. 또한, 외래어 자체는 차용주의 언어적 제약으로 인해 외래어의 정착 과정에서 형태, 음운 및 의미 변화를 겪는다. 이와 같이, 외래어의 수용 과정에서 차용주와 차용어의 다양한 변화가 일어나기 때문에 외래어는 역사언어학의 형태론, 음운론, 의미론과 같은 여러 분야에서 중요하게 연구되는 주제 중 하나이다.

외래어는 주로 차용주의 단어로는 표현할 수 없는 완전히 새로운 외국 제품명이나 개념을 나타내는 데 사용된다. 그런데 한편으로는 이미 고유어로 존재하는 단어를 좀 더 고급스럽고 학술적인 이미지로 바꾸기 위해 외래어를 사용하기도 하는데, 이러한 외래어의 사회언어학적 역할은 최근 특히 주목을 받고 있다.

대부분의 외래어 선행연구는 외래어의 많은 예를 수집하고 언어변화 패턴을 정리하는 방법으로 진행되었다. 최근 말뭉치 기반의 정량적 연구에서는 단어 길이와 같은 언어학적인 요인들이 외래어가 차용주에 성공적으로 정착하는 과정에 영향을 미치는지 통계적으로 연구하는 방법이 많이 사용되었다. 그러나 이러한 단어의 빈도기반 연구는 단어의 복잡한 의미 정보를 정량화하는 데에는 어려움이 있어 외래어 의미 현상에 대한 정량적 분석연구는 아직 진행되지 않았다.

본 연구는 외래어와 관련된 의미 현상을 정량적으로 분석하기 위한 단어임베딩

(Word Embedding) 기반의 방법을 제안한다. 단어 임베딩 방법은 딥 러닝 방법과 언어 빅데이터를 사용하여 단어의 의미 문맥 정보를 벡터 값으로 효과적으로 변환할 수 있다. 이 방법을 활용하여 외래어와 관련된 의미 현상의 세 가지 주제, **어휘 경쟁**, **의미적 적응**, **사회적 의미 기능과 문화적 경향 변화**에 초점을 맞추어 연구를 진행하였다.

첫 번째 연구는 외래어와 차용주의 동의어 간의 어휘경쟁에 중점을 둔다. 빈도 기반의 방법으로는 어휘 경쟁의 유형(*단어 대체* 또는 *의미 분화*)을 구별할 수 없다. 어휘 경쟁의 유형을 판단하려면 외래어와 차용주 동의어 간의 문맥 공유 상태를 파악해야 한다. 문맥 공유 상태를 정량적으로 모델링하기 위해 본 연구는 기하학적 개념을 적용한다. 제안된 기하학적 단어 임베딩 기반 모델은 외래어와 수용언어의 동의어 사이에서 발생하는 어휘 경쟁을 정량적으로 판단함을 확인할 수 있었다.

두 번째 연구는 일본어와 한국어에서의 영어 외래어의 의미 적응에 중점을 둔다. 영어 외래어는 차용주에 정착하는 과정을 통해 의미 적응을 겪는다. 본 연구는 외래어와 영어 고유어와의 의미 차이를 비교하기 위해 변환 행렬 방법을 적용하여 영어 외래어의 일본어와 한국어에서의 의미 적응 차이를 분석하였다. 또한, 영어 단어의 다의성이 의미적응에 주는 영향을 통계적으로 분석하였다.

세 번째 연구는 일본과 한국의 최신 문화적 경향을 반영하는 외래어의 사회 의미적 역할에 초점을 맞춘다. 일본과 한국 사회의 미디어에서는 새로운 문화적인 경향이나 이슈가 생겼을 때 외래어를 자주 사용하므로, 외래어가 일본과 한국의 문화적 경향을 반영하는 역할을 가질 것이 예상된다. 본 연구는 이러한 외래어가 문화적 경향의 변화를 반영하는 지표로서의 역할을 한다는 가설을 제안한다. 이 가설을 검증하기 위해 사전 훈련된 문맥 임베딩 모델(BERT)을 사용하고 시간에 따른 외래어의 문맥 변화를 추적하는 방법을 제안한다. 실험 결과, 제안된 방법을 통해 외래어의 문맥 변화 추적을 통해 문화적 경향의 변화를 감지할 수 있었다.

본 연구에서는 기본적으로 일본어와 한국어 데이터를 사용하였다. 이것은 전산 다국어 대조 언어연구의 가능성을 보여준다. 이러한 단어 임베딩 기반의 의미 분석 방법은 다언어 계산의미론 및 계산사회언어학의 발전에 많은 기여를 할 수 있을 것으로 예상된다.

주요어: 빅 데이터, 딥 러닝, 단어 임베딩, 외래어, 어휘 경쟁, 의미 적응, 사회언어학, 문화 경향 변화 감지

학번: 2014-30878