Building a Korean Twitter Corpus using Python

Wonbin Kim

University of Georgia – wk32657@uga.edu

Abstract

A corpus is a collection of texts, which is an important language resource where we can observe how people actually use language. It has been widely used in various fields such as lexicography and natural language processing (e.g., Hanks, 2012; Pustejovsky & Stubbs, 2012) as well as linguistics. However, despite its importance, the Korean national corpora have not been updated since 2007. Also, the Yonsei Twitter Corpus, which is a large-scale Korean Twitter corpus, consists of old data. Thus, this paper aims to build a new Korean Twitter corpus on the basis of up-to-date data and present how to create a Korean Twitter corpus by means of Python.

1 Introduction

A corpus is a collection of texts sampled from produced speech and writings. With the development of computer technology, the definition of a corpus has been changed a little. In modern linguistics, it is defined as a large set of authentic written or spoken texts saved in a computer readable format which can be representative of a particular language (e.g., Atkins et al., 1992; Baker, 1995; Francis, 1992, for the definition of a corpus). A corpus has made a significant contribution to linguistic research, which includes allowing linguists, for example, to identify lexical relations (e.g., Partington, 1998; Stubbs, 2001, for collocational analysis) or to keep track of language change by providing the snapshots of languages over certain time periods (e.g., Säily, 2014). Moreover, in addition to the field of linguistics, it has been usefully used in a number of different fields such as lexicography and natural language processing (e.g., Hanks, 2012; Pustejovsky & Stubbs, 2012).

As language continuously changes, a corpus has to keep expanding in accordance with such change. It has to include up-to-date data regularly so that researchers can study the current language phenomenon or diachronic change of language. However, the update of the Korean national corpora, i.e., the Sejong Corpora stopped in 2007 when the related project finished so they do not contain recent data now. Also, the Yonsei Twitter Corpus, which is one of the

large-scale Korean Twitter corpora in South Korea, consists of tweets written in October in 2011. These outdated corpora cannot be employed for research on the recent language phenomena in the Korean language. To address such issue, this study aims to build a new Korean Twitter corpus with up-to-date data.

For the creation of the new Korean Twitter corpus, the programming language of Python is used because it can more efficiently deal with several processes required for the construction of a corpus from data collection to preprocessing to annotation as well as it allows researchers to collect a large set of data for free. The specific way of creating a Korean Twitter corpus by means of Python will be able to serve as a guide for those who want to make their own corpus but do not have any knowledge in the construction of a corpus. The new Korean Twitter corpus created in this paper will be an important language resource which enables researchers to have access to and explore naturally-occurring authentic language in real life. Also, it is expected to be helpful in investigating, in particular, the characteristics of spoken Korean in the late 2010s.

1.1 Types of Corpora

With regard to types of corpora, there are many different types and it is possible that one type of corpus shares some characteristics with different types of corpora. Table 1 (Appendix A) shows ten types of corpora, which are commonly classified corpora in corpus linguistics, with the characteristics and example English corpora for each type. The new Korean Twitter corpus belongs to the type of specialized corpus in that it contains language used in a social networking service over a specific time period.

Besides the described corpora above, there are a number of different types of corpora depending on specific purposes. They include spoken and written corpora depending on the mode of data, monolingual and multilingual corpora depending on the number of covered languages, raw and tagged corpora depending on the presence or absence of annotation, reference and target corpora depending on the purpose of comparison, paralinguistic and sign language corpora depending on the target of research, and so on.

2 Korean Corpora

This section introduces Korean corpora with focus on three noticeable Korean corpora, i.e., the Yonsei corpora, the Trends 21 Corpus, and the Sejong Corpora. Korean corpora consist of not words but eojuls, unlike English corpora. An eojul[1] refers to a content word itself or the morphosyntactic combination of a

[1] e.g.) 비가 온다. NM: Nominative case particle
pi-ka o-n-ta. IN: Indicative mood suffix
rain-NM come-IN-DC DC: Declarative sentence-type suffix
'It is raining.'

content word and thing(s) in charge of grammatical function (e.g., particles and endings). Eojuls compose a Korean sentence and they are separated by spacing. Whereas spacing occurs between words in English, it occurs between eojuls in the Korean language. This is why eojuls are employed to parse sentences in Korean corpora. Eojuls are used to mention the sizes of Korean corpora. They are also used to mention the sizes of Korean corpora introduced in the following subsections and the new Korean Twitter corpus created in this paper (exactly, a raw Korean Twitter corpus).

2.1 Construction of Korean Corpora

The history of Korean corpora is short, compared to that of English corpora; it has only been about 30 years since the first electric Korean corpus was constructed (cf. the Brown Corpus of Standard American English, the first modern computerized corpus, which was constructed by Henry Kučera and W. Nelson Francis at Brown University in the United States in the 1960s). In South Korea, interest in corpora began in the late 1980s, from which corpora started to be constructed by a number of institutes and universities revolving around Yonsei University, Korea University, and the Korea Advanced Institute of Science and Technology.

The first electric Korean corpus was built by the Yonsei Institute of Language and Information Studies in 1988, which is called the Yonsei Corpus 1. The Yonsei Corpus 1 was constructed based on how people read, i.e., their actual reading habits, with the assumption that their vocabulary is formed and reinforced by what categories of reading materials they read and which category they read more. To find out in what proportion they read different kinds of written texts, a survey was carried out with a thousand adult Koreans. The result from the survey of asking how much time to spend on average reading different kinds of written text showed that the participants spend 36.6 % of time on daily newspapers, 22.8% on magazines, 20.4% on books of fiction, 11.2% books on general culture and information, and 9.0% on biographies. On the basis of these text category proportions, the Yonsei Corpus 1 was constructed. It consists of 2.9 million eojuls and its contents were collected from materials published from 1980 to 1987.

Since the Yonsei Corpus 1, the Yonsei Institute of Language and Information Studies has worked on expanding their corpus. As a result, many different kinds of corpora following the first one have been constructed depending on specific purposes and topics. Among them, twenty-six Yonsei Corpora are considered as main corpora of the institute. At first, the Yonsei Corpora were not open to the public but since 2016, anyone can have access to some of them online (https://ilis.yonsei.ac.kr/corpus for access to the corpora). Table 2 (Appendix

The example sentence consists of the following two eojuls: i) *pi-ka* and ii) *o-n-ta*. To be specific, the first eojul is the morphosyntactic combination of the noun *pi* and the nominative case *ka* and the second eojul is the morphosyntactic combination of the verb stem *o-*, the indicative mood suffix *-n*, and the declarative sentence-type suffix *-ta*.

A) shows the size and brief description for each of the twenty-six main corpora (https://ilis.yonsei.ac.kr/ for the information in Korean).

Another remarkable corpus is the Trends 21 Corpus, which was built by Research Institute of Korean Studies at Korea University. It consists of 600 million eojuls and was collected from the texts from four main daily newspapers in South Korea, i.e., The Dong-a Ilbo, The Chosun Ilbo, JoongAng Ilbo, and The Hankyoreh, which are materials for 14 years from 2000 to 2013. In addition to the Yonsei Corpora and the Trends 21 Corpus, many different kinds of corpora have been constructed by a large number of research institutes and universities depending on various research objectives. For example, Newspaper Corpus, Chinese-English-Korean Multilingual Corpus, Korean Tree-Tagging Corpus, Automatically Analyzed Large Scale KAIST Corpus, Terminology Corpus, and so on were built by the Korea Advanced Institute of Science and Technology (http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus for access to those corpora). Also, Korean Sentiment Analysis Corpus was constructed by Seoul National University, the Korean-German parallel Corpus by Hankuk University of Foreign Studies, the Corpus of Historical Materials by Kyung Hee University, and the Corpus of Korean Resident in Japan by Jeju National University. However, the sizes of all these corpora are much smaller than those of the Yonsei corpora, the Trends 21 Corpus, and the Korean national corpora, i.e., the Sejong Corpora.

The most important landmark in Korean corpora is the Sejong Corpora of 200 million eojuls. They were constructed under a 10-year national project, i.e., the 21st Century Sejong Project (more detailed information about the Sejong Corpora is covered in the next section). Since the 21st Century Sejong Project finished in 2007, the Sejong Corpora have not been updated with up-to-date data so they are considered out of date now. Fortunately, the National Institute of the Korean Language is currently constructing the Web Corpus by collecting language data on the Web such as social media platforms and blogs. They aim to collect two million posts from social media platforms, ten thousand posts from blogs, ten thousand posts from bulletin boards, and a hundred thousand posts from reviews like comments on products with the investment of $60 million in the data collection. The Web Corpus will be able to make up for the Sejong Corpora and further be widely applied to the field of artificial intelligence in South Korea.

2.2 Sejong Corpora

The Sejong Corpora are Korean national corpora, which are open to the public. They were built by Korea University and Yonsei University under a $12 million government-sponsored national project named the 21st Century Sejong Project, which was performed from 1998 to 2007. Constructing a large-scale national corpus comparable to the British National Corpus in the UK was one of the goals that the project pursued in order to promote the development of language research and technology in South Korea. When the project finished, the Sejong Corpora used to be distributed in DVD in the beginning, with each version

updated over 4 times, 2007, 2009, 2010, and 2011, respectively. They are not distributed in DVD any longer now but they can be downloaded from the website of the National Institute of the Korean Language (http://ithub.korean.go.kr/ for access to the Sejong Corpora).

Specifically, the size of the Sejong Corpora is about 200 million eojuls. They consist of seven corpora: written modern Korean, spoken modern Korean, North Korean/Korean used abroad, old Korean, Korean-English parallel corpora, Korean-Japanese parallel corpora, and Korean terminology. Table 3 shows the format and size for each corpus in the Sejong Corpora (Kang, 2008, for detailed information about the Sejong Corpora). In the table, a raw corpus refers to a corpus of the original. A tagged corpus means a corpus with morphological information added to the raw corpus. A word-disambiguated corpus is a semantically tagged corpus with disambiguated sense information added to the tagged corpus. A treebank is a corpus with syntactic structural information added to the word-disambiguated corpus.

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

Table 3 The format and size for each corpus in the Sejong Corpora

For the content of each category,

- **Written and spoken Korean corpora**: They consist of materials after 1910s. The corpus of written modern Korean was collected from various types of texts such as newspapers and magazines. The corpus of spoken

modern Korean was sampled from monologues and dialogues, both of which are divided into public and private subcategories.

- **North Korean/Korean abroad corpora**: They were constructed from Korean texts used in North Korea, Chinese, and Commonwealth of Independent States.
- **Old Korean corpora**: They were collected from Korean texts from the 15th century to the beginning of 20th century.
- **Korean-English and Korean-Japanese parallel corpora**: They contain both source language and their translated language texts.
- **A Korean terminology corpus**: It was built upon professional texts in various fields.

When the Sejong Corpora were released, the size was large enough to match corpora in the United States, the UK, Japan, and so on (their sizes were 200 million to 500 million words back then). However, they have got left behind since 2007 when the project stopped. Although a number of new words have been created and commonly used for about thirteen years after the termination of the project, the Sejong Corpora do not even contain them, let alone distinguish what parts of speech they are. A corpus has to continue to grow to reflect the dynamics of language change over time because it cannot be considered as a representative corpus unless it is regularly updated (Hunston, 2002). Representativeness is one of the important conditions that a corpus must fulfill and a representative corpus is necessary for more accurate and precise linguistic research.

3 Building a Korean Twitter Corpus

The reason for choosing a Korean Twitter corpus is that the Korean language used in Twitter is close to spoken Korean rather than written Korean (Shi, 2020). In spoken language, the length of sentences is relatively short and less refined sentences are used because people instantly speak without having enough time to elaborate what they say, compared to written language. Moreover, spoken language has more ungrammatical expressions and slang than formal writings such as news articles and editorials. Korean tweets containing these characteristics of spoken language will be able to clearly show the aspects of spoken Korean, i.e., how people actually use the Korean language in everyday life, in particular, in informal settings.

The Yonsei Twitter Corpus mentioned in Section 2.1 is a Korean Twitter corpus which boasts a large size. It was built for the analysis of political inclinations through a large number of tweets during the period of Seoul mayoral election campaign in 2011. The tweets were randomly collected with no specific keyword, in collaboration with a social media analytics company. The

Yonsei Twitter Corpus (945,175,620 eojuls) accounts for about 82% of the entire Yonsei Corpora (1,148,089,842 eojuls). This means that the Yonsei Twitter Corpus makes a great contribution to Korean corpora in terms of size.

However, the Yonsei Twitter Corpus is not open to the public and it is an old corpus now because the tweets in the corpus are tweets written for a single month, October in 2011. With the old and short-term corpus, it is impossible to study on recent language phenomena or the flow of language change. Thus, to overcome these limitations, this paper builds a new Korean Twitter corpus on the basis of up-to-date data over a longer period. The new Twitter corpus is expected to provide much information about new words or expressions which were recently created and language change that happened in the late 2010s. The following two subsections present how to create the new Twitter corpus in detail from data collection to preprocessing to annotation.

3.1 Data Collection

Python is one of the most popular computer programming languages, which was created by Guido van Rossum. Because Python is relatively easy to learn, many non-programmers are learning it and it is widely used by a large number of big companies like Google, Instagram, and IBM. Python has a number of useful libraries which can be used to perform diverse tasks. Here, a library means a collection of packages. A package is a collection of modules. A module is a Python file containing various Python functions and variables. That is, a library is a set of code created to perform a certain task. Thanks to the already made libraries/packages, users do not need to write Python code from scratch for themselves. However, users should modify some of the libraries/packages or write their own code, mixing their code and some of the existing libraries/packages depending on their specific research objectives because the existing libraries/packages only provide so-called general frames.

For the collection of Twitter data, the package of Twitterscraper was used. As we can see from its name, this is a package used to scrape tweets. It was developed by Ahmet Taspinar to improve the disadvantages of scraping Twitter data using Twitter's application programming interfaces (APIs). An API is a software intermediary provided by a particular software program or operating system, where third parties are allowed to access data from them and further, extend the functionality of that software application. With Twitter's APIs, developers or users have to go through a few steps in order to have access to Twitter data. They have to generate Twitter API keys, Access Token and secret keys, and so on. Another disadvantage is that developers or users can only have access to tweets written in the past seven days from the date when collection is started. Thus, Twitterscraper was used to collect tweets because it has no such limitations.

Between 12,000 and 13,000 (inclusive) tweets were evenly sampled per day without any keyword. This is because the number of tweets that can be scraped per day is limited to about 13,000. However, for one day, February 4 in 2019, 11,810 tweets were scraped. Despite several trials, more than 11,810

tweets were not scraped. This might be because that is all tweets Twitter has on that date. For each tweet, the user's name, timestamp, and text were only scraped. The total number of morphemes from tweets scraped per day was about 150,000. Because the target number of words for making a corpus was at least one hundred million, it was estimated that tweets over two years are necessary. According to the estimation, tweets in the last 26 months from January in 2018 to February in 2020 were scraped and the total number of scraped tweets was 10,100,995. It took about 13 hours to collect tweets for one month and about 330 hours, i.e., about fourteen days to collect all the data over the past 26 months in my computing environment. The data was collected with the aim of making two corpora: raw corpus and tagged corpus. The raw corpus consists of original tweet texts without any annotation. That is, it is a corpus of tweets which do not go through the process of data preprocessing. In contrast, the tagged corpus is composed of tweet texts where the process of data preprocessing is applied. It contains the information of parts of speech. More detailed information about the tagged corpus is covered in the following section.

3.2 Data Preprocessing

The Natural Language Toolkit (NLTK) in Python is a suite of libraries and programs for natural language processing but it cannot be used for the Korean language. Thus, for data preprocessing, KoNLPy was used, which is a package for natural language processing of the Korean language. Data preprocessing of the collected Korean tweets comprises the following six processes: tokenization, tagging, normalization, stemming, cleaning, and removal of stop words.

Tokenization is the process of splitting the entire text into tokens (e.g., words, phrases, or sentences). Tagging is the process of tagging each word or morpheme with its part of speech (e.g., noun, verb, adjective, or adverb). Normalization involves converting the entire text into lowercase or uppercase, expanding abbreviations, and so on so that different forms of the same word can be recognized as the same word. The normalization in Korean is different from that in English because, for example, there is no such distinction between lowercase and uppercase in the Korean language. In the normalization in Korean, different variations of the same word created by the use of different vowels or addition of unnecessary consonants are converted into their original forms. Stemming is the process of reducing a word to its stem or root form.

These four processes of tokenization, tagging, normalization, and stemming were performed by Okt[2] in tag Package of KoNLPy, which is an open-source Korean tokenizer developed by Will Hohyon Ryu. For the creation of a tagged corpus, the texts were split into morphemes. The morphemes were tagged with the following twenty-four tags: Noun, Verb, Adjective, Adverb, Determiner, Modifier, Conjunction, Exclamation, Josa (i.e., postposition), PreEomi

[2] https://konlpy.org/en/latest/api/konlpy.tag/ (This website introduces various Korean morphological analyzers including Okt and presents example code for each.)

(i.e., pre-final ending), Eomi (i.e., ending), Prefix, VerbPrefix (i.e., affix located before verb), Suffix, Punctuation, Foreign, Alpha (i.e., alphabet), Number, Unknown, Korean Particle[3] (i.e., unnecessary/extra consonants or vowels which do not compose a syllable, functioning as emoticons), Hashtag, ScreenName (i.e., Twitter username), Email, and URL.

Among the above twenty-four tags, morphemes with fourteen tags were only extracted (i.e., Noun, Verb, Adjective, Adverb, Determiner, Modifier, Conjunction, Exclamation, Josa, PreEomi, Eomi, Prefix, VerbPrefix, and Suffix) because morphemes with the other tags were judged to be unnecessary or improper for the analysis of the Korean language. This kind of task is called data cleaning, which is to correct or remove incomplete/incorrect/inaccurate/irrelevant data.

For removal of stop words, the list of Korean stop words[4] which has already been made and is widely used was employed. The list contains 789 stop words. Stop words are a set of commonly used words which have grammatical function or very little meaning, for example, like *and*, *the*, and *a* in English. The removal of stop words allows users to focus on more important words. For instance, when we typed phrases or sentences into a search engine, the search engine presents more pages about relatively more important words (i.e., content words) than commonly used words (i.e., function words). Because the tagged corpus in this paper was intended to be used for research on lexical relations based on content words, the Korean stop words were removed from the raw data. Through all these six processes, the tagged corpus has been built.

3.3 Results

Both the raw and tagged corpora have been saved in CSV files. Because if a file size is too large, it takes too much time to open the file, they have been saved in the form of day-to-day file (Figure 1). Also, the day-to-day files have been grouped by month for user-friendliness (Figure 2).

Specifically, the raw corpus consists of raw tweet texts including users' names and timestamps. It has no annotation showing grammatical categories. Figure 3 shows some examples of the raw corpus. The tagged corpus is composed of tweet texts which went through the process of data preprocessing. It consists of morphemes with their parts of speech tagged. Figure 4 shows some examples of the tagged corpus. With regard to the size of each corpus, the raw corpus consists of 85,685,671 eojuls and the tagged corpus consists of 118,277,930 morphemes.

4 Discussion

The new Twitter corpora created in this paper will be able to be usefully used for research on the Korean language in the late 2010s in that they contain

[3]This is different from particles mentioned in the definition of eojul in Section 2.
[4]https://www.ranks.nl/stopwords/korean

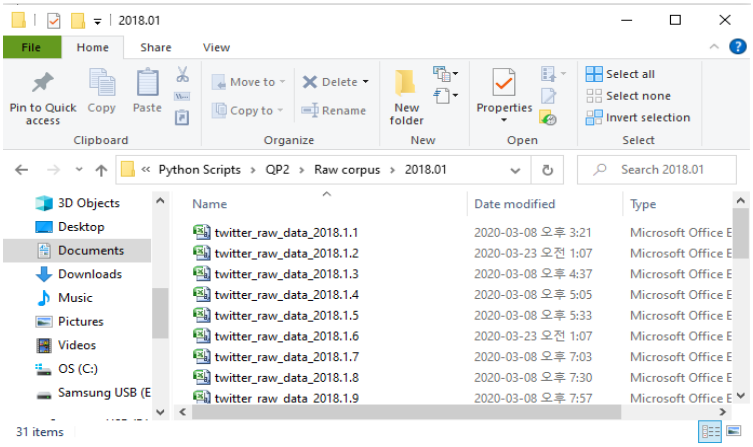

Figure 1 The CSV files of raw corpus saved in the form of day-to-day file

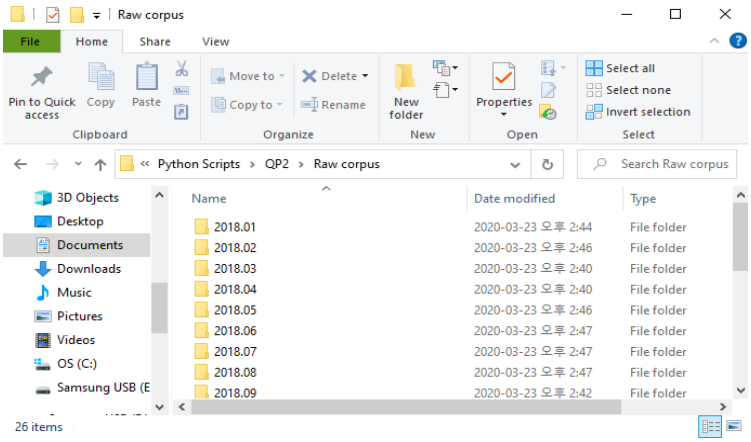

Figure 2 The raw corpus grouped by month

	A	B	C
1	Screen_name	Timestamp	Text
2	tufqpsl0205	2018-01-01 23:59	항상 보아도 들어도 좋아~
3	FdcXqtKJNTm44pr	2018-01-01 23:59	영상전보벌레를 붙잡았습니다!
4	ghdpsdk12	2018-01-01 23:59	띠용 오른쪽에 누구였지 저 장미맨만 기억남 ㅠㅠ
5	Henrya__	2018-01-01 23:59	머리부터 어질어질해짐
6	Eve_VNR	2018-01-01 23:59	(아뮐리 진짜 어떡해요)
7	Am_SooCute	2018-01-01 23:59	낮잠 자구 시퍼 ㅠ(찰푸닥)
8	jjck002	2018-01-01 23:59	@SBS_MTV 컴백 #더쇼 #GOT7 니가하면 멋있어! @SBS_MTV 컴백 #더쇼 #GOT7 니가하면 멋있어!
9	GaeGang_Yada	2018-01-01 23:59	아 자꼬 엘마라하네땎콩~(ㅅㅂ 에세요에세
10	fkfkfnffn	2018-01-01 23:59	@SBS_MTV #더쇼 #GOT7 갓세븐!

Figure 3 The examples of the raw corpus

	A	B
1	Corpus	
2	('항상', 'Noun')	
3	('보아', 'Noun')	
4	('도', 'Josa')	
5	('들다', 'Verb')	
6	('좋다', 'Adjective')	
7	('오페라', 'Noun')	
8	('의', 'Josa')	
9	('유령', 'Noun')	
10	('영상', 'Noun')	

the most up-to-date data for now. For instance, they can be used for research on new words or expressions created in the late 2010s or linguistic phenomena recently happening to the Korean language. The comparison between the new Twitter corpora and the Yonsei Twitter Corpus will show specifically what changed in 2018 and 2019, compared to 2011. Also, the comparison with the spoken Korean corpus of the Sejong Corpora will indicate changes in the characteristics of spoken Korean.

However, the new Twitter corpora have a few limitations. First, their use is restricted to some fields. They cannot be used for discourse analysis or text linguistics where the analysis of context or text structure is required respectively because tweets mostly consist of a few short sentences. Furthermore, they are not proper for sociolinguistic research because the collected tweets do not include the users' personal information such as age and gender.

Secondly, the results of tokenization and tagging are not perfect. For example, when the Korean noun *saypyek*, which means 'dawn', is misspelt as sey*pyek*, the wrong word is tokenized into each syllable, leading *sey* to be tagged with 'Modifier' and *pyek* with 'Noun' (because in Korean, *sey* means the number 'three' and *pyek* means the noun 'wall' when they stand alone). However, a bigger problem is that it is impossible to check how many and what kinds of errors are contained in the tagged corpus. Moreover, one person cannot manually check whether the analysis of every morpheme is right and correct every error because the corpus size is so large.

Any other limitations include the following things: (i) due to the removal of stop words, it is impossible to study on function words in detail; (ii) for research on a certain keyword, a large set of data only consisting of sentences including the keyword is needed. But such data might not be available if the new Twitter corpora do not have enough tweets including that keyword (since the new Twitter corpora were collected with no keyword); (iii) software tools such as WordSmith Tools and AntConc, i.e., tools used to analyze linguistic data in corpus linguistics have to be used to analyze texts from the new Twitter corpora.

While the first limitation cannot be improved because of the intrinsic nature of tweets (a tweet can contain up to 280 characters) and no access to personal information, the other limitations can be overcome. With regard to the second limitation, it may be able to be improved by means of khaiii (Kakao Hangul Analyzer III), which is a morphological analyzer developed based on deep learning, specifically, a convolutional neural network. Because the morphological analyzer of khaiii analyzes morphemes based on syllables and corrects the errors of data through enough contexts, the accuracy of morpheme analysis is known to be higher (https://github.com/kakao/khaiii for more details about khaiii). The application of khaiii will be able to improve the accuracy of the new tagged corpus.

Concerning the other limitations, if a researcher or user is familiar with Python, those things are no longer limitations. If they set a keyword when scraping tweets and do not remove stop words in the process of preprocessing, they will be able to collect enough data that they want and study on function words in more detail. Also, if they can write and run Python scripts, software tools such as WordSmith Tools and AntConc are not needed for the analysis of the Twitter corpora because they can analyze the data with their Python code.

5 Conclusion

In consideration of the fact that the Yonsei Twitter Corpus and the Sejong Corpora do not include recent data and a Korean Twitter corpus is useful for research on the Korean language used in everyday life, this paper has built a new Korean Twitter corpus based on up-to-date data. Using some libraries and packages from Python, tweets over the past 26 months from January in 2018 to February in 2020 were collected. About 13,000 tweets were sampled per day (except one day) and a total of 10,100,995 tweets were scraped. From all the scraped tweets, raw and tagged corpora have been constructed. The raw corpus consists of 85,685,671 eojuls with no annotation and the tagged corpus consists of 118,277,930 morphemes with their part of speech information. As large-scale corpora, these new Korean Twitter corpora are expected to be helpful for research on the Korean language in the late 2010s, although they have a few limitations.

Corpora are important language resources that show how words are actually used, how often words are used, what words mean, which words are used

together, and so on. They have been used for language research (e.g., Moon, 1998, for collocational analysis) and widely applied to various fields such as lexicography, natural language processing, and language pedagogy. Regarding language pedagogy, the necessity of corpora in language teaching dates back to 1920s (Thorndike, 1921) and corpora have played a significant role in this field. For example, learner corpora are used to find out what grammatical errors foreign language learners frequently make and improve the learning of foreign language learners on the basis of those errors (e.g., Nesselhauf, 2004). Moreover, corpora can be used to provide foreign language learners with information about grammar rules, collocations, or which words to choose depending on the context so that what they say can sound more natural to native speakers of that foreign language (e.g., Gavioli, 2005).

With the advent of the fourth industrial revolution centered upon technological innovation, corpora have gotten more important. Artificial intelligence (AI), one of the core fields in the fourth industrial revolution, is strongly associated with natural language processing because AI requires speech recognition. In order for machine to recognize spoken words and further, convert them into text, it needs natural language processing, which is founded on tons of language data, i.e., large-scale corpora. Given that corpora have such wide applications to many fields and they can spur the development of AI, the importance of corpora is expected to keep increasingly growing in the future. In accordance with this trend, this paper will be able to promote the use of corpora in a number of various fields as well as linguistic research using corpora by presenting how to create a Twitter corpus more readily and efficiently.

References

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, *7*(1), 1–16.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *International Journal of Translation Studies*, *7*(2), 223–243.

Francis, W. N. (1992). Language corpora b.c. In J. Svartvik (Ed.), *Directions in corpus linguistics. proceedings of the nobel symposium 82* (pp. 17–32). Mouton de Gruyter.

Gavioli, L. (2005). *Exploring corpora for esp learning* (Vol. 21). John Benjamins Publishing.

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography*, *25*(4), 398–436.

Hunston, S. (2002). *Corpora in applied linguistics*. Ernst Klett Sprachen.

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

Moon, R. (1998). *Fixed expressions and idioms in english: A corpus-based approach*. Oxford University Press.

Nesselhauf, N. (2004). How learner corpus analysis can contribute to language teaching: A study of support verb constructions. *Corpora and language learners*, *17*, 109–124.

Partington, A. (1998). *Patterns and meanings: Using corpora for english language research and teaching* (Vol. 2). John Benjamins Publishing.

Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning: A guide to corpus-building for applications*. O'Reilly Media Inc.

Säily, T. (2014). *Sociolinguistic variation in english derivational productivity: Studies and methods in diachronic corpus linguistics* (Master's thesis). Société Néophilologique.

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

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell publishers.

Thorndike, E. L. (1921). *The teacher's word book*. Columbia Teachers College.

A Additional Tables

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

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

Table 1 The characteristics and example English corpora depending on corpus types

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

		education books and was constructed in 1996.
#10. Yonsei Corpus 10	780,000	It is composed with the separate volumes from the first period (1945~1965) corpus supplemented for the compilation of Yonsei contemporary Korean dictionary.
#11. Yonsei Corpus 11	730,000	It is composed with the textbooks from the first period (1945~1965) corpus supplemented for the compilation of Yonsei contemporary Korean dictionary.
#12. Yonsei Corpus of Korean in the 20th Century	150,378,870	It is a raw corpus of written language collected from 20th literature materials depending on publication dates and text types.
#13. Corpus of Korean Textbooks (Complete)	724,856	It was collected from Korean textbooks from Korean language education institutes in 1990s.
#14. Corpus of Korean Textbooks (Conversation)	119,598	It was collected from the dialogues of introductions in Korean textbooks from Korean language education institutes in 1990s.
#15. Yonsei Korean Learner Corpus	278,542	As a Korean learner corpus, it was collected from compositions by students from Yonsei institute of Language Research and Education.
#16. Korean Elementary Textbook Corpus after Independence	1,496,280	It was collected from elementary school Korean language textbooks published after the period from 1945 to 1954.
#17. The 6th and 7th Korean Elementary Textbook Corpus	1,681,769	It was collected from textbooks in the 6th and 7th curriculums. It provides annotations on homonyms.
#18. Yonsei Balanced Corpus of Written Discourse	1,054,362	It is a corpus of written language composed of texts from a range of genres.
#19. Yonsei Balanced Corpus of Spoken Discourse	998,934	It is a corpus of spoken language collected from monologues and public & private dialogues.
#20. Yonsei Corpus of Polysemy	1,165,224	It is a corpus providing annotations on polysemy, which was constructed for a Korean meaning frequency dictionary.
#21. Yonsei Corpus	386,472	It was collected from the doctrines of Buddhism,

of Hangul tripitaka		Buddhist scriptures, Tripitaka, and so on.
#22. Corpus of <Tongnip Sinmun> Newspaper	144,309	It was constructed based on <Tongnip Sinmun>, which was an early Korean newspaper and the first privately managed modern daily newspaper in Korea.
#23. Corpus of Popular Songs in the Modern Era	29,339	It was collected from the lyrics of popular songs in 1930s and 1940s.
#24. Yonsei Corpus of Multimodal Data	18,986	It is composed of videos of shooting utterances, transcription texts, and annotations on non-linguistic actions.
#25. Twitter Corpus	945,175,620	It was collected from Korean tweets which were written for one month, October in 2011.
#26. Political Discourse Corpus	306,681	It was constructed for discourse analysis with the topic limited to politics.
Total	1,148,089,842	

Table 2 The size and brief description for each of the twenty-six Yonsei corpora