Specifications and Analysis of the Korean Sentiment Analysis Corpus*

Hyopil Shin and Munhyong Kim
(Seoul National University)

Shin, Hyopil & Munhyong Kim. (2013). Specifications and Analysis of the Korean Sentiment Analysis Corpus. ***Language Research*** **49.2, 227-250.**

This paper describes the two year endeavor of constructing the Korean Sentiment Analysis Corpus (KOSAC), focusing on the theoretical background and the analysis of the corpus itself. Our aim is to provide a solid theoretical background for the corpus which reflects the characteristics of the Korean language and includes approximately 7,744 sentences taken from news articles. The corpus annotation scheme, based on the MPQA, is described along with the statistics of features specified in the corpus. The analysis of the corpus can be a starting point for how to utilize the corpus not only for sentiment analysis but also for semantic or pragmatic work in terms of speaker's attitude and emotional expressions.

Keywords: Sentiment Analysis, Korean Sentiment Analysis Corpus, Opinion Analysis, MPQA

1. Introduction

There has been much research on the automatic identification and extraction of sentiments and opinions in text. Researchers have been working on these issues by focusing mainly on subjectivity and sentiment classification either at the document or sentence level. Classifying editorials or movie reviews as positive or negative are both examples of document classification tasks while classifying individual sentences as subjective or objective would be an example of a sentence-level task (Wiebe et al. 2005).

Along these lines of research, a need for corpora annotated with

* This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-327-A00322). This work is an extension of Shin et al. (2012) and adopts basic descriptions from the work.

rich information about opinions and emotions has also emerged. This would allow for the development of statistical and machine learning approaches for various practical NLP applications. As such a resource, the Multiperspective Question Answering (MPQA) Opinion Corpus, developed by Wiebe (2002), Wiebe et al. (2005), and Wilson et al. (2008), plays an important role in sentiment and opinion analysis. It contains the manual annotation of a 10,000 sentence-corpus of articles from the world press. Since this corpus provides a fine-grained annotation scheme, it is widely used as a source for training data in machine learning approaches and serves as the gold standard in sentiment classification tests.

We started constructing a language sentiment corpus, called the Korean Sentiment Analysis Corpus (KOSAC).[1] We received two years of support in this project from the Korean Research Foundation (KRF) from May of 2011 to April of 2013. We aimed to provide both a solid theoretical background for the Corpus, reflecting the characteristics of the Korean language, as well as fine-grained annotations for the 7,744 sentence-corpus of news articles. The total number of annotated sentences is less than that of the MPQA, but since our annotation is morpheme-based due to the agglutinative nature of Korean, the number of annotation units is much greater. We have also adopted the basic annotation scheme of the MPQA for comparative research purposes.

The remainder of this paper is organized as follows. Section 2 gives a brief overview of the MPQA corpus as a starting point. Section 3 elaborates on the annotation scheme for the Korean sentiment corpus, providing examples of annotations with attributes. Section 4 shows observations on the corpus. Section 5 presents future work and conclusions.

2. The MPQA Corpus

As a fundamental resource for sentiment corpus construction in Korean, this work takes advantage of the Multiperspective Question Answering (MPQA) Opinion Corpus which began with the conceptual structure for private states in Wiebe (2002) and developed manual an-

1) http://word.snu.ac.kr/kosac.

notation instructions. The MPQA Corpus version 1.0 was released in 2003, and now version 2.0 is available with more detailed attitude annotations. In this section we briefly review the annotation scheme and structures of the corpus with a view to providing a theoretical background.

2.1. Private States

According to Quirk et al. (1985), a private state refers to mental and emotional states such as the opinions, beliefs, and intentions of a writer. Wiebe et al. (2005) focused on identifying private state expressions in contexts and presented numerous examples annotated with schemes that cover a broad range of linguistic expressions and phenomena.

Private states and speech events are the core of the MPQA corpus. Private states cover *opinions*, *beliefs*, *thoughts*, *feelings*, *emotions*, *goals evaluations*, and *judgments* (Wiebe et al. 2005). Private state frames cover expressive subjective element frames, which are used to represent expressive subjective elements, as well as direct subjective element frames, which are used to represent subjective speech events. In order to distinguish opinion-oriented material from fact, objective speech event frames are also defined in terms of speech events. Private state frames have the following attributes directly excerpted from Wiebe et al. (2005).

Direct subjective frame:

- text anchor: a pointer to the span of text that represents the speech event or explicit mention of a private state
- source: the person or entity that is expressing the private state, possibly the writer
- target: what the speech event or private state is about
- properties
  - intensity: the intensity of the private state (low, medium, high, or extreme)
  - expression intensity: the contribution of the speech event or private state expression itself to the overall intensity of the private state (neutral, low, medium, high, or extreme)
  - insubstantial: true, if the private state is not substantial in the discourse

- attitude type: represents the polarity of the private state. The possible values are positive, negative, other, or none

Expressive subjective element frame:
- text anchor
- source
- properties
  - intensity
  - attitude type

Unlike the MPQA, we do not distinguish direct subjective frames from expressive subjective elements. Rather, those two frames are merged into SEED subjective expressions in our approach.

2.2. Objective Speech Event

Objective speech event in the MPQA is used to distinguish opinion-oriented material from material presented as factual and has the following attributes.

Objective speech event frame:
- text anchor
- source
- target

2.3. Nested Sources

In sentiment analysis, it is very useful to recognize the person whose opinion or emotion is being expressed. Thus 'source' is introduced in the MPQA.

The source of a speech event is implicitly the speaker or the writer while the source of a private state is the experiencer. However, there are situations where speech events and private states are assessed by more than one source. In this case, an additional explicit source was introduced. This source generally corresponded to the subject of the embedded predicate. This is a so-called nested source, as adopted by Wiebe et al. (2005), Wilson (2008), and Sauri (2008). Nested sources include other people's speech events and private states as well as the

speaker's. Please look the following examples adopted from Wiebe et al. (2005: 9):

(1) a. Sue said, "The election was fair."
b. Sue thinks that the election was fair.
c. Sue is afraid to go outside.

In the above sentences, Sue is the source of speech event (1a) and of private states (1b, 1c). However, we do not know what Sue says, thinks, or feels directly. We only know Sue's speech event according to the writer. In the MPQA Corpus, such a nested source would be represented as <*writer*, *Sue*>. Private states can be directed toward the private states of others. Consider Wiebe et al. (2005)'s example:

(2) "The U.S. fears a spill-over," said Xirao-Nima.

In (2), it is not *the U.S.* that directly states its fear. Rather, according to the writer, the *Xirao-Nima* states that the U.S. fears a spill-over. Thus the nested source of the fear can be expressed as <*writer, Xirao-Nima, U.S*>.

3. Outline of Annotation Scheme for Korean Sentiment Analysis Corpus

Our work essentially follows the MPQA, but we have also modified the existing MPQA attributes as well as introduced new attributes to address the characteristics of Korean.

The annotation scheme starts with distinguishing a SEED from a whole sentence in terms of subjectivity. In a SEED, each individual unit expresses a private state. By contrast, the subjectivity of the whole sentence is about whether we feel the sentence is objectively true or not in terms of the speech event. Even though a sentence bears many subjective expressions in it, the sentence can carry objective facts. Thus our annotation principle separates basic subjective expressions from subjectivity of a whole sentence. That is, unlike the MPQA, we explicitly annotate subjectivity or objectivity of the sentence. This principle can be illustrated as follows.

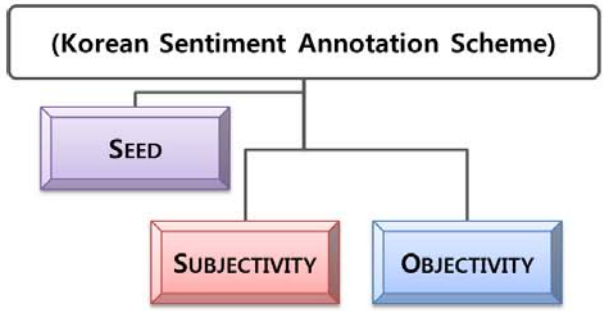

Figure 1. Korean Sentiment Annotation Scheme.

As a basic annotation unit, we chose a morpheme rather than a word. Korean is an agglutinative language and many meaning-bearing particles and sentence endings can carry private states, therefore we need to be able to pinpoint these precise segments as a basic unit. Although such morpheme-based annotation helps to produce a fine-grained corpus, the trade-off is that it also requires a great deal of time and effort spent on annotating.

3.1. SEED

The elements of SEED are as follows:

- anchor: morpheme id(s)
- id: tag id
- expressive type: direct-explicit, direct-speech, direct-action, indirect, writing-device
- subjectivity type: emotion-pos, emotion-neg, emotion-neutral, emotion-complex, judgment-pos, judgment-neg, judgment-neutral, agreement-pos, agreement-neg, agreement-neutral, argument-pos, argument-neg, argument-neutral, intention-pos, intention-neg, speculation-pos, speculation-neg, others
- nested-source: w-sources
- target: target id(s)
- polarity: positive, negative, neutral, complex
- intensity: low, medium, high

According to Wiebe et al. (2005: 4) private states are states of *experiencers* holding *attitudes*, optionally toward *targets*. For example, in the sentence *John hates Mary*, the experiencer is *John*, the attitude is *hate*,

and the target is *Mary*. Thus, in order to annotate subjective expressions, all three attributes of the private state should be properly represented. In the MPQA, the following three main types of private state expressions were included: explicit mentions of private states, speech events expressing private states, and expressive subjective elements. MPQA's expressive subjective elements, speech events, and private state attitudes roughly correspond to SEED, expressive type, and subjectivity type in our scheme, respectively.

3.1.1. Expressive Types

Express types specify either speech events (acts) that express private states (or other subjective elements) or non-speech events. These fit into five subtypes: *direct-explicit*, *direct-speech*, *direct-action*, *indirect*, and *writing-device*. While the former three types are related to speech events and usually originate from subject-predicate relations, *indirect* and *writing-device* are used for a writer to show his/her own subjectivity through non-predicate expressions. These include using a nominal as an argument, adverbials, conjunctive endings, or some particles in Korean. *Indirect* and *writing device* are common in that subjectivity is not carried through speech event. In the case of *indirect*, the source of the expression is not clear compared to *direct* or *writing device*. The following shows examples of each expression type.

- explicit: *cikyepta* 'boring', *inkita* 'be popular'
- direct speech: *cwucanghata* 'insist', *pinanhata* 'blame,'
- direct action: *elkwulsayki pyenhata* 'turn pale', *hwanhohata* 'acclaim'
- indirect: *isanghan salam* 'strange people', *hwullyunghi* 'greatly'
- writing-device: *-man* 'only', *isanghakeyto* 'strangely'

3.1.2. Subjectivity Types

The attribute subjectivity type is used to classify subjective expressions according to their sources' attitudes; lexically determined as the core meaning of subjective expression. It consists of the following subtypes: *emotion*, *judgment, agreement, argument, intention*, and *speculation*. These types can be further combined with other polarity attributes such as *positive, negative, neutral* and *complex* according to their semantic orientations which may lead to complex attributes such as *emotion-positive*, *emotion-negative*, and so on. Generally, a *complex* attribute

is due to a combination of positive and negative words, such as in the Chinese character expression '幸不幸,' 'happiness and unhappiness'. The MPQA does not provide this kind of detailed classification. Considering our previous sentiment research, we think that classifying subjectivity into more refined types provides benefits not just when determining whether a document is subjective but also when determining what kind of attitude the document contains. The subjectivity types are exemplified as follows:

Table 1. Subjectivity Types

Type	Values	Examples
Emotion	Emotion-positive	*kipputa* 'glad', *miso-lul cista* 'make a smile'
	Emotion-negative	*mwusepta* 'afraid', *kothongsulepta* 'feel pain'
	Emotion-neutral	*kamtong-i epsta* 'not touching'
	Emotion-complex	*hayngpwulhayng* 'happiness and unhappiness'
Judgment	Judgment-positive	*yongkamhata* 'be brave', *cangcem* 'merit'
	Judgment-negative	*napputa* 'bad', *kepcayngi* 'a coward'
	Judgment-neutral	*aymayhata* 'vague', *cal molukessta* 'don't know well'
Agreement	Agreement-positive	*tonguyhata* 'agree', *yongnaphata* 'accept'
	Agreement-negative	*pantayhata* 'do not agree', *kikak* 'rejection'
	Agreement-neutral	*kikwenhata* 'give up', *cwunglip* 'be in the middle'
Argument	Argument-positive	*cungmyenghata* 'verify', *seltukhata* 'persuade'
	Argument-negative	*panpakhata* 'refute', *kecisita* 'not true'
	Argument-neutral	*cham kecis-ul kwupwunhal swu epsta* 'can't know if it is true or not
Intention	Intention-positive	*uytohata* 'intend', *kyelsimhata* 'make one's mind'
	Intention-negative	*~hal maum-i epsta* '~not willing to', *wuyenhi* 'accidentally'
Speculation	Speculation-positive	*chwuchukhata* 'speculate', *somang* 'wish'
	Speculation-negative	*epsta* 'there is not'

3.1.3. Targets

Attribute targets are used to specify objects or themes to which the subjective expressions are directed. In many cases targets can be clearly specified but in some cases pinpointing source and target is not that

simple. The following is a complicated example of target which requires an embedded clause as target.

(3) Mary-nun ku-wa hamkkey issnun
Mary-subj he-with together be-adnom
kes-i koylowessta
that-sub feel uncomfortable (past)
"That he was with Mary made her feel uncomfortable

The target of *koylowessta* 'be hard' is not *ku* 'he' but an embedded clause which has a meaning of 'the fact that he was with Mary'. Next, due to the possibility of double subjects in Korean, some expressions can have more than two targets.

(4) Sakwa-ka phwumcil-i cohta.
apple-subj quality-subj good
"The apple has a good quality"

3.1.4. Nested Sources

Since source information is crucial to sentiment analysis, the MPQA elaborates on sources and nested sources in annotations. As described in 2.3, nested sources include other people's speech events or private states as well as those of the speaker or writer. Table 2 shows some examples of nested sources. Here, underlining means a subjective expression and bold face means a nested source.

Table 2. Example of Nested Sources

Types	Example	Values
a. Source = writer	*Kwail-un* *sakwa-ka* *ceyilita* 'fruit'-topic apple-subj best-be As for fruit, apple is best	w
b. Source = writer According to = subject Subject = writer	***Na**-to* *sakwa-lul* *cohahanta* I–too apple-obj like I like an apple too.	w w-I
c. Source = writer According to = subject	***Tom**-un* *sakwa-lul* *cohahanta* Tom-subj apple-obj like Tom likes an apple	w-Tom

Types	Example	Values
d. Source = writer According to = A According to = B	***Tom**-un* ***Mary**-ka* *sakwa-lul* Tom-subj Mary-subj apple-obj *cohahanta-ko* *malhayssta* like-comp say-past Tom said that Mary likes an apple	w-Tom-Mary
e. Source = unclear, or general population	*Cohun* *khameyla-nun* *pissata* 'good' camera-subj expensive Good cameras are expensive	w-out
f. Source = not explicitly specified source in a sentence	*Yocum* *inkki-iss-nun* Now popular-be-adnom *khameyla-nun* *gf-1 ita* camera-subj gf-1 be Now popular camera is gf-1	w-imp

Following the MPQA, we specify nested sources from left to right. That is, <*w-Tom-Mary*> means that writer states Mary's speech event through Tom's eye. w-out and w-imp represent generic sources and implicitly specified sources, respectively. In (f), we can guess the source of 'be popular' from the context. Meanwhile, general population is the source of the belief 'good' in (e).

3.1.5. Polarity, Intensity, and Insubstantial

The attribute polarity describes whether the (nested) source has a positive or negative subjectivity toward the target. An example of a positive value would be *coh*-(*ta*) 'good/well' while an example of a negative value would be *nappu*-(*ta*) 'bad'. In addition, there are two more values: neutral and complex. The value of attribute intensity depends on how intensely subjectivity is expressed. For example, (*i chayk-un*) ***kucekuleh**-ta* '(this book is) so-so' shows a neutral intensity while (*i chayk-un*) ***ssuleyki**-ta* '(this book is) trash' shows a highly intense negative subjectivity. Similarly, intensity modifiers, e.g. *maywu* 'very,' *sangtanghi* 'considerably,' or *nemwu* 'too (bad),' can also affect the intensity of an expression. The following illustrates a SEED annotation:

Manh$_0$-un$_1$sayongca$_2$-tul$_3$-i$_4$i$_5$ceyphwum$_6$-ul$_7$cohaha$_8$-ko$_9$iss$_{10}$-ta$_{11.12}$
Many$_0$-ADNOMINAL$_1$user$_2$-PLURAL$_3$-NOM$_4$this$_5$product$_6$-ACC$_7$like$_8$-DURATIVE$_{9,10}$-DECL$_{11.12}$
'Many users like this product'
<SEED> anchor = "8" id = "u1" type = "direct-explicit" subjectivity-

type = "emotion-pos" nested-source = "w-manhun sayongcatul" target = "5-6" polarity = "positive" intensity = "medium" </SEED>

3.2. Sentence Level Subjectivity

Unlike MPQA, we explicitly specify the whole sentence's subjectivity. Although each sentence consists of various numbers of subjective expressions, we feel that a sentence may be an objective fact rather than subjective. Thus we mark the subjectivity of a whole sentence on the basis of the speech event, i.e. from the writer's perspective. We believe that this can help researchers to extract relevant features for subjectivity from those sentences and to train the corpus to see what makes the sentences subjective or objective. Information on the sentence level subjectivity or objectivity differs from SEED tags as they have relatively simple structures, as follows.

- The BNF of SUBJECTIVITY
  anchor: Morpheme id(s)
  id Sentence id
  polarity: positive, negative, neutral, complex
  intensity: low, medium, high

The OBJECTIVITY tag consists of only the attributes *anchor* and *id*.

- The BNF of OBJECTIVITY
  anchor: Morpheme id(s)
  id Sentence id

Examples of SUBJECTIVITY and OBJECTIVITY tags are listed in (5). The subjectivity of objectivity of a sentence can be influenced by SEED tags, but it is not completely dependent on them. In a case of a SEED tag affecting the subjectivity of the whole sentence, usually the original source of the subjectivity indicated by the SEED tag is the writer of sentence. That is, there is no nested-source except the writer: nested-source="w". In (5c), 'was reported as a regrettable event that Yumi bought a house,' the value of nested-source "w-general" represents general population.

(5)

a. $Yumi_{0}$-$ka_{1}$$cip_{2}$-$ey_{3}$$ka_{4}$-$n_{5}$$il_{6}$-$un_{7}$$chamulo_{8}$ $yukamsulep_{9}$-$ta_{10.11}$
$Yumi_{0}$-NOM$_{1}$$home_{2}$-AT$_{3}$$go_{4}$-ADNOMINAL$_{5}$$event_{6}$-TOP$_{7}$$truly_{8}$$regret$-$table_{9}$-DECL$_{10.11}$
'It is truly regrettable that Yumi went home'

<SUBJECTIVITY> anchor="0-11" id="s1" polarity="negative" intensity="high" **</SUBJECTIVITY>**
<SEED> anchor="8-9" id="u1" type="direct-explicit" subjectivity-type="judgment-neg" nested-source="w" target="0-6" polarity="negative" intensity="high" **</SEED>**

b. $Yumi_{12}$-$nun_{13}$$kkoley_{14}$$cip_{15}$-$ul_{16}$$sa_{17}$-$ss_{18}$-$ta_{19.20}$
$Yumi_{12}$-TOP$_{13}$in.a.pathetic.state$_{14}$$home_{15}$-ACC$_{16}$$buy_{17}$-PAST$_{18}$-DECL$_{19.20}$
'Yumi was pathetic but she bought a house'

<SUBJECTIVITY> anchor="12-19" id="s2" polarity="negative" intensity="high" **</SUBJECTIVITY>**
<SEED> anchor="14" id="u1" type="writing-device" subjectivity-type="judgment-neg" nested-source="w" target="12" polarity="negative" intensity="high" **</SEED>**

c. $Yumi_{21}$-$ka_{22}$$cip_{23}$-$ul_{24}$$sa_{25}$-$n_{26}$$il_{27}$-$un_{28}$$yukamsulewu_{29}$-$n_{30}$$saken_{31}$-$ulo_{32}$
$pokotoy_{33}$-ess_{34}-$ta_{35.36}$
$Yumi_{21}$-NOM$_{22}$$home_{23}$-ACC$_{24}$$buy_{25}$-ADNOMINAL$_{26}$$event_{27}$-TOP$_{28}$
$regrettable_{29}$-ADNOMINAL$_{30}$$event_{31}$-$as_{32}$$be.reported_{33}$-PAST$_{34}$-DECL$_{35.36}$
'It was reported as a regrettable event that Yumi bought a house'

<OBJECTIVITY> anchor="21-36" id="o1" **</OBJECTIVITY>**
<SEED> anchor="29" id="u1" type="indirect" subjectivity-type="judgment-neg" nested-source="w," target="31" polarity="negative" intensity ="medium" **</SEED>**

4. Analysis of the Korean Sentiment Analysis Corpus

4.1. Annotation Process

The size of corpus largely depends on the speed of annotation work.

Without an appropriate annotation tool, it is almost impossible to build a large annotated corpus.

Though the MPQA opinion corpus was built with GATE annotation tool, we developed a morpheme based annotation tool for Korean text as in Figure 2 for three reasons (Cattle et al. 2013). First, none of current annotation tools, such as GATE or brat, supported switching between word and morpheme views. Second, there are non-continuous sentiment expressions that cannot be annotated by current tools. Third, within those tools targets and nested-sources of sentiment expressions need to be annotated in advance to sentiment expressions which is not intuitive and in turn makes process of annotation slow.

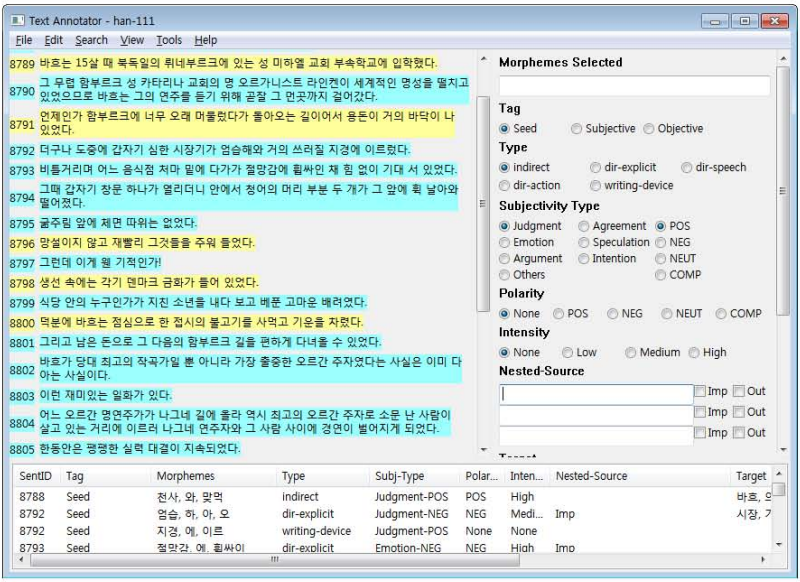

Figure 2. Morpheme Based Annotation Tool.

Moreover, to ensure the quality of annotations, three well-trained linguistic students annotated texts separately, and then double cross-checked the annotations until all annotators agreed on the same annotations. In many cases, two or more people annotat on the same text, and settl on a single annotation. After annotating a corpus, to see reliability of a building process of the corpus, it is recommended that an inter-annotator agreement test be performed. In our work, however, since each person annotated different text and double cross-checked

each other's results, it was not possible to measure the inter-annotator agreement. Thus, the sincerity of an annotated corpus was measured in a different way. Twenty frequently occurring sentiment expressions were chosen from six subjectivity types to see how consistently people annotated those expressions. The ratio of annotated times to the number of occurring times for each of those expressions was used for measurement as in Table 3.

Table 3. Frequency Cross Table of Expressive and Subjectivity Type Needed?

Agreement	ratio	Argument	ratio	Emotion	ratio
hapuyha 'agree'	0.86	*cwucangha* 'insisit'	0.98	*twulyep* 'fear'	1.00
incengha 'admit'	0.90	*cicekha* 'point out'	0.90	*pwunno* 'anger'	0.93
pantayha 'disagree'	1.00	*ceysiha* 'suggest'	0.82	*salangha* 'love'	0.94
kepwuha 'deny'	0.90			*hayngpokha* 'happy'	0.94
Intention	ratio	Judgement	ratio	Speculation	ratio
ko siph 'want'	0.88	*inki* 'popular'	0.87	*nun kes kath* 'might'	0.50
ki wiha 'purpose'	0.63	*caymi* 'fun'	0.59	*ul kes* 'would'	0.20
tolok 'purpose'	0.52	*cwungyoha* 'important'	0.90	*yeysangtoy* 'expected'	1.00
yeyceng 'plan'	0.61	*phwungpwuha* 'plentiful'	0.91		

These sample expressions show that the overall annotation consistency was reasonably high for most expressions except verb inflectional morphemes due to their distinctively high frequencies.

4.2. Corpus Statistics

Among the 7,744 sentences in the corpus, 2,654 were annotated as subjective and 5,090 as objective. 17,582 Seed tags were created, indicating on average 2.3 Seed expressions exist per sentence. In addition, 4,960 types were annotated as positive, 4,373 as negative, and 208 as complex polarity. For 17,582 Seed annotations, the frequencies of expressive types and subjectivity types are given in Table 4. It can be seen that the Judgment subjectivity type is the most predominant type since Judgment type expressions include not just short sentiment words or phrases, but also clauses that show speakers' judgments. Indirect expressions include all sentiment expressions except all main predicates and writing-device expressions; accordingly, indirect ex-

pressive type is also the most frequent type of all. A large portion of Writing-Device expressions are categorized as Others subjectivity type because they do not usually belong to any other subjectivity types.

Table 4. Frequency Cross Table of Expressive and Subjectivity Type

	Agreement	Argument	Emotion	Intention	Judgment	Speculation	Others
Dir-Action	1	8	73	8	41	0	1
Dir-Explicit	156	276	344	276	2740	157	40
Dir-Speech	8	1150	22	28	86	13	7
Indirect	252	321	714	406	6079	61	22
Writing-Device	4	98	9	305	770	171	2935

To help understand which expressions belong to such types above and how they were annotated, Figure 3 shows some examples of some of the types.

<Dir-Explicit & Agreement>
ttusul mou 'agree'
kyeluyha 'resolve'
panpali kangha 'strongly oppose'

<Dir-Action & Emotion>
nwunmwuli hulu 'tear drops'
elssaan 'hug'
khikkhikkeli 'giggle'

<Writing-Device & Judgment>
haci moshamyen 'if do not do (it)'
ceyamwuli 'even if'
ohilye 'rather'

Figure 3. Examples of annotated expressions.

From examples above, it can be seen that annotated expressions are not restricted to a certain syntactic segments, rather they reveal one's subjectivity. Also, it is noticeable that intensifiers are not separated from sentiment expressions.

To more specifically describe the annotation results, SEED tag expressions were sorted depending on the frequencies of Part-of-Speech (POS) patterns.[2] From the sorted list, the most frequently annotated

patterns of subjective expressions could be found. Table 5 shows the top-10 frequent patterns of SEED expressions. It is assumable that these expressions could be used as entries of a sentiment dictionary.

Table 5. The Top 10 Frequent Part-of-Speech Patterns of Seed Expressions

Rank	Part of speech patterns	Examples	Frequencies
1	NNG	*mwuncey* 'problem', *cwucang* 'argument', *salang* 'love', *inki* 'popular'	1924
2	NNG XSV	*kangco-ha* 'point out', *sayngkak-ha* 'think', *yokwu-ha* 'ask', *ihay-ha* 'understand'	1154
3	MAG	*thukhi* 'especially', *ohilye* 'on the contrary', *mwullon* 'of course', *tto* 'also'	716
4	VV	*wiha* 'aim', *palkhi* 'announce', *culki* 'enjoy', *coh.aha* 'like'	659
5	EC	*tolok* 'for(purpose)', *nuntey* 'conjunction', *lamye* 'said that'	657
6	XR XSA	*phwungpwu-ha* 'plentiful', *hwullyung-ha* 'excellent', *kwungkum-ha* 'curious'	538
7	VA	*komap* 'thank', *nollap* 'surprised', *twulyep* 'fear'	491
8	NNG JKO VV	*inki-lul-kkul* 'popular', *cohwa-lul-ilwu* 'be harmonious', *uykyen-ul-mou* 'agree'	479
9	EC VX	*lyeko-ha* 'to do', *a.yaman-ha* 'have to be', *ko-mal* 'eventually do'	396
10	NNG NNG	*naymyen-uysik* 'inward conciousness', *naypwu-kopal* 'whistle blowing'	348

Moreover, depending on expressive types of SEEDs the frequent POS patterns were considered to be shown to see a trend; among the top 50 most frequent POS patterns, twelve of them seem to be predominately dir-explicit type, as seen in Table 6.

2) To understand the meaning of POS tags, the Sejong POS tag set is provided in Appendix A.

Table 6. The Patterns Frequently Occurring as Dir-explicit

POS Patterns	dir-action	dir-explicit	dir-speech	indirect	writing-device
ETM NNG VCP	0	36	7	1	18
MAG NNG XSV	0	41	3	33	1
MAG VA	0	43	0	26	0
NNG JKB VV	1	47	0	36	0
NNG JKO NNG XSV	0	37	4	35	0
NNG JKO VV	26	237	25	190	1
NNG JKO VV EC VX	0	29	2	14	0
NNG JKS VA	0	53	2	45	0
NNG JKS VV	5	69	4	47	0
NNG VCP	0	115	7	17	2
NNG XSV EC VX	0	24	0	21	1
VA EC VX	0	28	0	15	0

These patterns mostly have dir-explicit or indirect patterns. This is because dir-explicit type expressions are the main predicates of sentences, and their patterns, which do not include any sentence ending markers, could also occur as an indirect type in modifying or subordinating clauses. Different from other patterns, the <ETM NNG VCP> sequence is mistakenly annotated as writing-device eighteen times since those expressions seemed to work as strong indicator of writers' subjectivity.

Three of fifty patterns are likely to be dir-speech type as in Table 7. These patterns include particles, such as EC and JKQ, which connect spoken content and the main speech related predicates.

Table 7. The Patterns Frequently Occurring as Dir-speech

POS Patterns	dir-action	dir-explicit	dir-speech	indirect	writing-device
EC NNG XSV	0	8	75	4	0
EC VV	1	10	45	4	18
JKQ NNG XSV	1	4	113	0	0

The majority, twenty one, of the top 50 patterns have indirect type as their most frequent type in Table 8. Here, there is also a tendency that dir-explicit type is as frequent as indirect, indicating these patterns can occur as a main predicate of a sentence. Among these, some pat-

terns that end with NNG, XSN, or XSA are distinctively more frequently used as indirect type rather than dir-explicit type because those patterns are much less likely to be used as main predicates.

Table 8. The Patterns Frequently Occurring as Indirect

POS Patterns	dir-action	dir-explicit	dir-speech	indirect	writing-device
MAG VV	4	46	1	58	1
MAG XR XSA	0	10	0	28	0
NNG	2	132	18	1709	63
NNG JKG NNG	0	5	0	65	0
NNG NNG	0	16	2	329	1
NNG NNG JKO VV	2	24	1	31	0
NNG NNG NNG	0	3	0	46	0
NNG VA	0	11	0	42	0
NNG VV	3	53	5	103	0
NNG XSA	0	73	0	224	0
NNG XSN	0	6	0	152	9
NNG XSN NNG	0	3	0	42	0
NNG XSN VCP	0	22	0	102	0
NNG XSV	13	402	201	537	1
VA	0	120	0	370	1
VA ETM NNG	0	1	1	63	0
VA ETM NNG JKO VV	1	19	1	20	0
VV	30	192	64	330	43
VV EC VX	2	39	2	39	1
XR XSA	0	89	0	449	0
XR XSA ETM NNG	1	2	0	47	0

Writing-device type patterns seem to be distinguishable from other type patterns as in Table 9. They are usually inflectional morphemes or adverbs revealing a writer's subjectivity.

Table 9. The Patterns Frequently Occurring as Writing-Device

POS Patterns	dir-action	dir-explicit	dir-speech	indirect	writing-device
EC	0	2	87	7	561
EC VX	0	18	2	14	362
EF	0	0	0	1	109
ETM	0	0	29	2	30
ETM NNB	0	1	0	1	63
ETM NNB VA	0	8	0	0	55
ETM NNB VCP	0	7	8	1	138
ETM NNB VV	0	4	0	2	222
ETM NNG	0	7	18	8	19
ETN JX VX	0	0	0	0	54
ETN VV	0	1	0	11	33
JX	0	0	0	0	225
MAG	0	1	0	136	579
MAJ	0	0	0	2	256

In addition to the frequent POS patterns depending on types, modifying intensifiers could be captured in the expressions since they are annotated together with modified sentiment expressions. These expressions determine the degree of intensity of sentiment. Observable intensifier patterns are <MAG>, <NNG JKB>, <VV EC>, <VA EC>, <MM NNG>, <NP JKB>, <NNG XSN>, <XR XSA EC>, and <NNG XSN JKB>. The top 20 frequently occurring patterns are all <MAG>s. A total of 1,112 intensifier types could be extracted with these patterns. Also, it could be found that the same modifying expression is annotated with a different intensity in the annotations, revealing annotators' intuition about the expression. For instance, *kacang* 'the most' is annotated as high 57, medium 34, and low 6 times. This frequency could be used to determine the intensity scale of the expression. Table 10 shows some examples of intensifiers for each pattern.

Table 10. Examples of Intensifier Patterns

Pattern	Word	Intensity	Frequency
MAG	*kakkai* 'almost'	Medium	4
NNG JKB	*kicek chelem* 'like miracle'	Medium	2

Pattern	Word	Intensity	Frequency
VV EC	*takuchi tus* 'press'	High	1
VA EC	*swipkey* 'easily'	Medium	9
MM NNG	*enu cengto* 'in some degree'	Medium	1
NP JKB	*mwues pota* 'above all'	High	6
NNG XSN	*sasil sang* 'actually'	High	1
XR XSA EC	*kanung ha myen* 'if possible'	Medium	1
NNG XSN JKB	*kesi cek ulo* 'macroscopically'	Medium	1

From the fine-grained annotated corpus, the characteristics of a subjective or objective sentence could be described by frequencies of expressive and subjectivity types.

Table 11. Average Frequencies of Types for an Objective or Subjective Sentence

Expressive Types	Objective	Subjective
dir-action	0.015772	0.017097
dir-explicit	0.374925	0.794073
dir-speech	0.225594	0.067629
Indirect	0.678179	1.679711
writing-device	0.354761	0.946809
Subjectivity Type	Objective	Subjective
Agreement	0.041925	0.079787
Argument	0.270313	0.18845
Emotion	0.116191	0.216565
Intention	0.118387	0.162234
Judgment	0.830904	2.087006
Others	0.241366	0.677052
Speculation	0.030146	0.094225
Number of Seeds	1.649231	3.505319

For an objective or subjective sentence, how many expressive types and subjectivity types it has on average is shown in Table 11. A subjective sentence tends to have more dir-explicit, indirect, writing-device expressive type than an objective sentence. The frequency of dir-speech type is higher for an objective sentence due to reporting predicates. For subjectivity type, a subjective sentence has a particularly higher

frequency of judgment, speculation, emotion, and others compared to objective sentences. Also the number of SEEDs in subjective sentences is double that of in objective ones.

5. Future Work and Conclusions

We have recently completed the Korean Sentiment Analysis Corpus. The first step was to investigate theoretical foundations and to make tools for manual annotations. Regarding theoretical background, we followed the annotation scheme and the framework proposed by the MPQA corpus. The framework of the MPQA is similar to that of Appraisal Theory by Martin (2000) and White (2002). The Appraisal framework is composed of concepts including *Affect, Judgment, Appreciation, Engagement*, and *Amplification*. *Affect, Judgment*, and *Appreciation* represent different types of positive and negative attitudes. According to Wiebe et al. (2005) the similarity between these approaches is that they are both concerned with systematically identifying expressions of opinions and emotions in context.

Nonetheless, the MPQA corpus does not distinguish different types of private states, such as *Affect* and *Judgment*, which can provide useful information in sentiment analysis. On the other hand, the MPQA corpus distinguished different ways that private states may be expressed, i.e. *directly* or *indirectly*.

Our annotation scheme, however, not only covers many types of attitudes as in Appraisal Theory but also several expressive types as in the MPQA corpus. Subjectivity types correspond to *Attitude* in Appraisal Theory and Expressive types correspond to *direct subjective* or *expressive subjective elements* in the MPQA. We believe that a corpus founded on a comprehensive annotation scheme could be used by researchers as a gold standard for training and testing.

As a preliminary analysis, we showed some statistics of the corpus. Those statistics intrinsically show significantly useful information for sentiment analysis. We believe that researchers will be able to extract variety of linguistic phenomena from the corpus and use the data not only for sentiment or opinion analysis but also for theoretical linguistic work. The main goal behind KOSAC was to support the development and evaluation of NLP systems that exploit opinions and senti-

ments in applications. Furthermore, rich information of opinionated expressions in our corpus annotations will contribute to a new understanding of how sentiments are expressed linguistically in Korean language. The corpus is now open to public for research purpose.

References

Cattle, Andrew, Munhyong Kim, and Hyopil Shin. (2013). Morpheme-based Annotation Tool for Korean Text. The American Association for Corpus Linguistics.

Finergan, Edward. (1995). Subjectivity and Subjectification: an Introduction. In D. Stein & S. Wright (Eds.). Subjectivity and Subjectification: Linguistic Perspectives, 1-15. Cambridge University Pres.

Krippendorff, Klaus. (1978). Reliability of Binary Attribute Data. *Biometrics* 34.1, 142-144.

Krippendorff, Klaus. (2004). *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage.

Langacker, Ronald. (1985). Observations and speculations in subjectivity. In J. Haiman (Ed.), Iconicity in Syntax. Typological Studies in Language 6. John Benjamins.

MPQA. (2005). Multi-Perspective Question Answering. University of Pittsburgh. http://www.cs.pitt.edu/mpqa/.

Pustejovsky, James, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Beth Sundheim, Lisa Ferro, Marcia Lazo, Inderjeet Mani, and Dragomir Radev. (2003). The TimeBank corpus. In Proceedings of Corpus Linguistics *2003*, 647-656.

Quirk, Randolf, Sidney Greenbaum, Geoffry Leech, and Jan Svartvik. (1985). *A Comprehensive Grammar of the English Language*. New York.

Riloff, Ellen Janyce Wiebe, and William Phillips. (2005). Exploiting Subjectivity Classification to Improve Information Extraction. In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-2005), 1106-1111.

Sauri, Roger. (2008). A Factuality Profiler for Eventualities in Text. Ph.D Dissertation, Brandeis University.

Shin, Hyopil, Munhyong Kim, Hayeon Jang, and Andrew Cattle. (2012). Annotation Scheme for Constructing Sentiment Corpus in Korean. In proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, 181-190.

Wiebe, Janyce. (2002). Instructions for Annotating Opinions in Newspaper Articles. Department of Computer Science Technical Report TR-02-101,

University of Pittsburgh.
Wiebe, Janyce, Theresa Wilson, and Claire Cardie. (2005). Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation, 39.2/3, 164-210.
Wiebe, Janyce and Ellen Riloff. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005), 486-497.
Wilson, Theresa Ann. (2008). Fine-Grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States. Ph.D Dissertation, University of Pittsburgh.

Hyopil Shin
Department of Linguistics
Seoul National University
1 Gwanak-ro Gwanak-gu, Seoul 151-742
Email: hpshin@snu.ac.kr

Munhyong Kim
Department of Linguistics
Seoul National University
1 Gwanak-ro Gwanak-gu, Seoul 151-742
Email: likerainsun@snu.ac.kr

Received: June 30, 2013
Revised version received: July 22, 2013
Accepted: July 29, 2013

Appendix.

Sejong Part-of-Speech Tag Set

TAG	Explanation	TAG	Explanation	TAG	Explanation
NNG	일반 명사	VV	동사	MM	관형사
NNP	고유 명사	VA	형용사	MAG	일반 부사
NNB	의존 명사	VX	보조 용언	MAJ	접속 부사
NR	수사	VCP	긍정 지정사	IC	감탄사
NP	대명사	VCN	부정 지정사	JKS	주격 조사
JKC	보격 조사	JKQ	인용격 조사	EC	연결 어미
JKG	관형격 조사	JX	보조사	ETN	명사형 전성 어미
JKO	목적격 조사	JC	접속 조사	ETM	관형형 전성 어미
JKB	부사격 조사	EP	선어말 어미	XPN	체언 접두사
JKV	호격 조사	EF	종결 어미	SP	쉼표, 가운뎃점, 콜론, 빗금
JKC	보격 조사	XSN	명사 파생 접미사	SS	따옴표, 괄호표, 줄표
JKG	관형격 조사	XSV	동사 파생 접미사	SE	줄임표
JKO	목적격 조사	XSA	형용사 파생 접미사	SO	붙임표(물결, 숨김, 빠짐)
JKB	부사격 조사	XR	어근	SW	기타기호(논리수학기호, 화폐기호)
JKV	호격 조사	SF	마침표물음표, 느낌표	SH	한자
NF	명사추정범주	NA	분석불능범주	SN	숫자
NV	용언추정범주	SL	외국어		