

Morpheme-based Annotation Tool for Korean Text

Andrew Cattle, Munhyong Kim, Hyopil Shin
Computational Linguistics Lab
Seoul National University

Table of Contents

- Introduction to Korean Sentiment Corpus
- Overview of Korean Language and Writing
- Overview of Existing Annotation Tools
- Design and Implementation
- Future Directions

Korean Sentiment Corpus

- Annotates subjectivity expressions in Korean text
- Morpheme-based annotations
 - Due to Korean's agglutinative nature
- Consists of 332 newspaper articles taken from the fine-grained morphologically analyzed Sejong Corpus
 - Covers life, society, culture, and politics
- Uses Korean Subjectivity Markup Language (KSML) (Shin et al, 2012a, 2012b)
 - English Sentiment Schema (Wiebe et al 2005, Wiebe 2002) not suitable for Sentiment Analysis of Korean Language

KSML Overview

- Three types of tag: SEED, SUBJECTIVE, OBJECTIVE
- SEED tags represent subjective statements smaller than a sentence. Contain:
 - Nested-source (the source of the subjectivity)
 - Type (how subjectivity is expressed)
 - Subjectivity-type (specific type of subjectivity)
 - Target (object the subjectivity is directed towards)
- SUBJECTIVE and OBJECTIVE tags denote when an entire sentence is subjective/objective (respectively)

KSML Example

나는 철수를 좋아한다.

Na-nun cheolsu-lul joha-han-ta

I-CASE Chulsoo-CASE LIKE-AUX-SENTEND

<SEED> anchor="좋아하-" nested-source="w-나" type="dir-explicit" subjectivity-type="emotion-pos" target="철수" polarity="pos" intensity="medium" </SEED>

Non-continuous Text Anchors

- Korean has a relatively free word order
- If “보다 좋다” (“*like more*”) is taken as a single sentiment statement:
 - **A가 B보다 좋다**
 - *A-ka B-bota cotha*
 - A-SUB B-COMPARATIVE good-SENTEND
 - A is better than B
 - **B보다 A가 좋다**
 - *B-boda A-ka cotha*
 - B-COMPARATIVE A-SUB good-SENTEND
 - A is better than B

Hangul (한글)

- Phonetic alphabet where individual consonants and vowels are denoted by 자모 (*jamo*)
- *Jamo* combine into blocks representing a complete syllable
 - Called 글자 (*kulja*, “written character”) or 음절 (*umjeol*, “syllable”)

<i>Jamo</i>					<i>Syllable/Kulja</i>	
ㅎ	+	ㅏ	+	ㄴ	=	한
/h/		/a/		/n/		/han/

Complex Verbal Inflections

- **좋다** (*jotha*)
 - good-SENTEND [It's] good.
- **좋아하다** (*johahata*)
 - good-AUX-SENTEND [I] like [it].
- **좋아졌다** (*johajyeotta*)
 - good-PROCESS-PAST-SENTEND [It] became good.
- **좋겠다** (*jokhetta*)
 - good-CONJECTURAL-SENTEND [That] will be good.
- For KSC, **좋** (*joh*) or **좋아** (*joha*, both “good”) is the important part. **The inflection is irrelevant**

Inflecting Korean Verbs

- Inflections may cross syllable boundaries:

하	+	-ㅂ니다	=	합니다
<i>ha</i> To do (base form)		<i>-mnida</i> Sentence ending (formal polite)		<i>hamnida</i>

Inflecting Korean Verbs

- Inflections may cause phonologic and orthographic reductions/assimilations:

하	+	-았	+	-다	=	했다
<i>ha</i> To do (base form)		<i>-at</i> Past tense particle		<i>-ta</i> Sentence ending (written neutral)		<i>haetta</i>

Hangul as Unicode

- Unicode's Hangul Syllables block contains 11172 *kulja*
 - U+AC00 through U+D7AF
 - Different Unicode blocks exist for individual *jamo*
- To a computer **하** (*ha*, U+D558) and **한** (*han*, U+D55C) are as different as A (U+0041) and B (U+0042)
- Unicode Hangul Syllable characters are atomic
 - A formula **does** exist to calculate constituent *jamo* **but** it is unreasonable to expect projects using the KSC to implement this themselves

Inflections and Annotations

- Inflection may not be relevant to annotation or inflection itself may be the only subjectivity expression
 - Tagging entire word may make automated analysis difficult
- Inflections in Korean **can** alter base form's orthography; cannot expect users to solve this problem themselves
- Solution: Represent text using base morphemes

Base Morpheme Example

Raw Text

이 밖에 수입품인 일본
마쓰시타 주서기는 전선의
길이가 98cm로 기준(1백
40cm)보다 짧은 것으로
나타났다고 밝혔다.

Base Morphemes

이 밖 에 수입품 이 나
일본 마쓰시타 주서기 는
전원 전선 의 길이 가 98
cm 로 기준 (1 백 40
cm) 보다 짧 은 것 으로
나타나 았 다고 밝히 었
다 .

*i pakk-ey suimphum-in ilpon masseusitha cuseki-neun censen-ui kili-ka
98cm-lo kicun(1 payk 40cm)-bota ccalb-eun kes-eulo nathanat-tako
balkhyet-ta.*

“Additionally, it was revealed that the imported Japanese Massushita Juicer’s power cable was 98cm, shorter than the standard 140cm.”

Base Morpheme Readability

- 수입품인 → 수입품 이 ㄴ
 - *suimphum-i-n*
 - imports-COPULA-ADJ
- 나타났다고 → 나타나 았 다고
 - *nathana-at-tako*
 - appear-PAST-REPORTING
- 밝혔다 → 밝히 었 다
 - *palkhy-et-ta*
 - reveal-PAST-SENTEND
- **Difficult to read in large amounts**

Annotation Requirements

- Ability to switch between fully-inflected text and base morphemes
 - Effectively doubles file size
- Ability to select non-continuous anchors
- Subjectivity Expression, Nested Sources, and Targets should all be created at the same time
 - Separating these attributes incorrectly implies they can exist independently
 - Also requires a counter-intuitive order of annotations

Existing Tools

GATE

- Available from <http://gate.ac.uk>
- Supports full range of Text Engineering Processes
 - From Annotation to Interpretation
- User-specifiable annotation schema
- Includes GATE Teamware
 - A web-based parallel annotation solution

GATE

The screenshot displays the GATE Developer interface. The main window shows a text document with the following content:

동묘를 나와 청계천 방향으로 가는데 길 양쪽은 운동 노점상이다.
나중에 인터넷에서 보니 이 일대가 동묘버락시장으로 유명하다고 한다।

< 동묘버락시장 소개 블로그글 1 >
< 동묘버락시장 소개 블로그글 2 >

Annotations are visible in the text, including a date and a location. The Annotations List table below shows the details of these annotations:

Type	Set	Start	End	Id	Features
Date		145	152	18	{kind=dateTime}
Date		780	784	20	{kind=date}
Location		994	997	19	{}

The Annotations List table is currently showing 3 annotations, with 1 selected. The selected annotation is the Date annotation at position 780-784.

Resource Features:

- MimeType: text
- docNewLineType: LF
- gate.SourceURL: file:

Document Editor: Initialisation Parameters

An annotation type and its features.

brat

- Available from <http://brat.nlplab.org>
- Web-based toolkit for annotation and visualization

brat

The screenshot shows a web browser window with the URL `192.168.0.153:8001/index.xhtml#/data/news`. The page content is a news article with several lines of text. Semantic annotations are visible as colored boxes and arrows:

- Line 1: "입력 : 2012.01.06 01:49"
- Line 2: "버락 오바마 미국 대통령은 5일(현지시간) 새로운 국방전략을 바탕으로 한 군 개혁으로 미군을 보다 슬림화하지만 세계 최강군으로서의 위치를 유지할 것이라고 강조했다."
- Line 4: "오바마 대통령은 이날 펜타곤에서 가진 국방부의 새 국방전략 발표식에 직접 참석, "미국 군대는 군사를 보다 없애는 쪽으로 나아갈 것"이라며 "하지만 미국은 모든 종류의 긴급사태와 위협에 준비태세를 갖추고, 날렵하고 유연한 군대를 가진 군사적 우위를 지속적으로 유지할 것"이라고 말했다." Annotations include "Person" (orange box) over "오바마", "GPE" (blue box) over "미국", "Origin" (blue arrow) from "GPE" to "Org", and "Org" (blue box) over "군대".
- Line 6: "버락 오바마 미국 대통령이 5일(현지시간) 펜타곤에서 가진 국방부의 새 국방전략 발표식에 참석해 연설을 하고 있다." Annotations include "Person" (orange box) over "버락 오바마" and "Employment" (blue arrow) from "Person" to "GPE", with "GPE" (blue box) over "미국 대통령".
- Line 7: "/AP연합뉴스"
- Line 9: "미국 역대 대통령이 펜타곤에서 회견을 갖고 연설을 한 것은 극히 이례적으로 이번 새 국방전략이 오바마 대통령의 뜻에 따른 것이라는 점을 분명히 하려는 의지로 보인다."

Drawbacks of Existing Tools

- Neither GATE nor brat allow for...
 - Switching between raw and parsed text
 - Annotating non-continuous segments of text
- In both Text Anchor, Nested Source, and Target must all be annotated separately then then...
 - Reference these separate annotation in a further “Sentiment Tag” annotation (GATE)
 - Create links between these separate annotations using a simple drag and drop technique (brat)

Design and Implementation

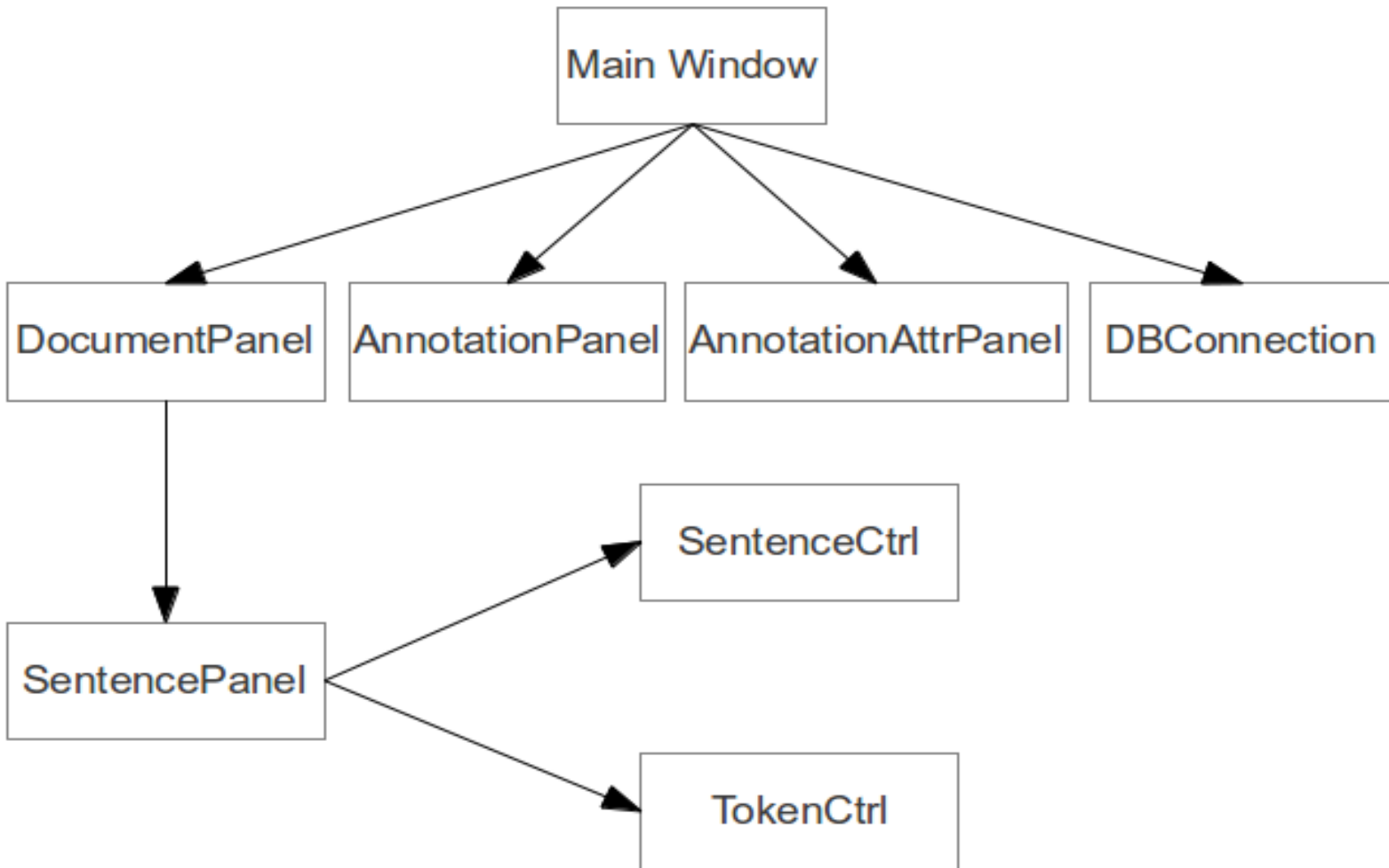
Design Motivation

- Structure of Korean Language and Orthography causes problems with many tools designed for Latin Alphabet Languages
- Lab undertook several text annotation-based projects
- Previous projects used custom quick-and-dirty tools that were purpose-built and difficult to adapt or maintain

Design Goals

- Focus on future adaptability and maintainability
- Modular design
 - Few inter-dependencies between components
- Publisher-Subscriber Pattern (PubSub)
 - Components don't need to know where input will come to/where output goes to, only the shape of the data

Design Overview



Design Overview

The screenshot shows the Text Annotator application window. The main text area contains several paragraphs of Korean text. A red box highlights the entire window, labeled "Main Window". A green box highlights the right-hand side panels, labeled "AnnotationAttrPanel". A blue box highlights the bottom table, labeled "AnnotationPanel". A purple box highlights the bottom right corner, labeled "AnnotationPanel". An orange box highlights a specific sentence in the text, labeled "SentencePanel". A yellow box highlights a word in the text, labeled "TokenCtrl". A pink box highlights another word in the text, labeled "SentenceCtrl".

Main Window

AnnotationAttrPanel

AnnotationPanel

SentencePanel

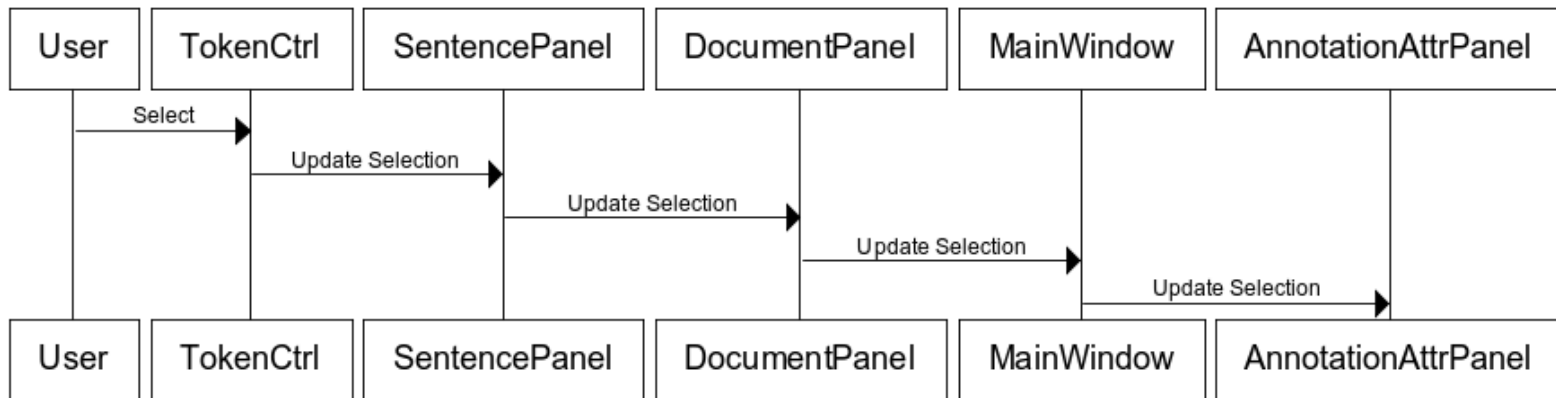
TokenCtrl

SentenceCtrl

S.	Tag	Morphemes	Type	Subj-type	Polar...	Inten...	Nested-source	Target
1..	SubjTag			-	None	None		
2..	SubjTag			-	None	None		

Token Selection w/o PubSub

User Selects a Token for Annotation



Token Selection w/ PubSub



- MainWindow, DocumentPanel, and SentencePanel no longer involved in token selection
- Fewer dependencies
- Asynchronous

Implementation

- Standalone application easier to develop in short timeframe
- Cross platform technologies
 - Python 2.7
 - wxPython 2.9
- Emulates environment's native GUI API
- **MORE?**

Annotation Tool Screenshot

Text Annotator - han-111

File Edit Search View Tools Help

8789 바흐는 15살 때 북독일의 뤼네부르크에 있는 성 미하엘 교회 부속학교에 입학했다.

8790 그 무렵 함부르크 성 카타리나 교회의 명 오르가니스트 라인켄이 세계적인 명성을 떨치고 있었으므로 바흐는 그의 연주를 듣기 위해 곧잘 그 먼곳까지 걸어갔다.

8791 언제인가 함부르크에 너무 오래 머물렀다가 돌아오는 길이어서 용돈이 거의 바닥이 나 있었다.

8792 더구나 도중에 갑자기 심한 시장기가 엄습해와 거의 쓰러질 지경에 이르렀다.

8793 비틀거리며 어느 음식점 처마 밑에 다가가 절망감에 휩싸인 채 힘 없이 기대 서 있었다.

8794 그때 갑자기 창문 하나가 열리더니 안에서 청어의 머리 부분 두 개가 그 앞에 휙 날아와 떨어졌다.

8795 굶주림 앞에 체면 따위는 없었다.

8796 망설이지 않고 재빨리 그것들을 주워 들었다.

8797 그런데 이게 웬 기적인가!

8798 생선 속에는 각기 덴마크 금화가 들어 있었다.

8799 식당 안의 누군인가가 지친 소년을 내다 보고 베푼 고마운 배려였다.

8800 덕분에 바흐는 점심으로 한 점시의 불고기를 사먹고 기운을 차렸다.

8801 그리고 남은 돈으로 그 다음의 함부르크 길을 편하게 다녀올 수 있었다.

8802 바흐가 당대 최고의 작곡가일 뿐 아니라 가장 출중한 오르간 주자였다는 사실은 이미 다 아는 사실이다.

8803 이런 재미있는 일화가 있다.

8804 어느 오르간 명연주가 나그네 길에 올라 역시 최고의 오르간 주자로 소문 난 사람이 살고 있는 거리에 이르러 나그네 연주자와 그 사람 사이에 경연이 벌어지게 되었다.

8805 한동안은 팽팽한 실력 대결이 지속되었다.

Morphemes Selected

Tag

Seed Subjective Objective

Type

indirect dir-explicit dir-speech
 dir-action writing-device

Subjectivity Type

Judgment Agreement POS
 Emotion Speculation NEG
 Argument Intention NEUT
 Others COMP

Polarity

None POS NEG NEUT COMP

Intensity

None Low Medium High

Nested-Source

Imp Out
 Imp Out
 Imp Out

SentID	Tag	Morphemes	Type	Subj-Type	Polar...	Inten...	Nested-Source	Target
8788	Seed	천사, 와, 맞먹	indirect	Judgment-POS	POS	High		바흐, 오
8792	Seed	엄습, 하, 아, 오	dir-explicit	Judgment-NEG	NEG	Medi...	Imp	시장, 기
8792	Seed	지경, 예, 이르	writing-device	Judgment-POS	None	None		
8793	Seed	절망감, 예, 휩싸이	dir-explicit	Emotion-NEG	NEG	High	Imo	

Raw Text vs. Base Morpheme

845 영등포구 신길1동 '진흙구이' 주인 이미숙씨는 "시골에서 진흙구이를 먹어 본 어른들이 특히 좋아한다"고 말한다.

1022 때문에 자기가 좋아하는 주제를 정해 시간을 가지고 찬찬히 보는 것도 한 방법이다.

1427 처음 취직했을 때 가족들의 반응은 '매우 기뻐했다'가 38.9%, '비교적 좋아했다'가 47%.

1485 최근 '내가 좋아하는 계절'로 영화계에 데뷔한 그는 어머니를 빼닮은 눈, 아버지와 비슷한 표정으로 관객을 매혹시키고 있다.

1714 몇 년 후 그 학생의 실력이 좋아서 악기를 바꿔야 할 때는 다시 선생이 그 악기를 사들여 D라는 새로운 학생에게 판다.

1826 Q2처음엔 입술이 부르트고 몸살을 앓기도 했으나, 그 고비를 넘기고 나니까 오히려 건강이 좋아지고, 집의 아이들에게 '근면'을 가르칠 수 있다는 보람도 느낄 수 있었습니다.

Raw Text View

Morpheme View

845 영등포구 신길1동 '진흙구이' 주인 이미숙씨는 "시골에서 진흙구이를 먹어 본 어른들이 특히 좋아한다"고 말한다.

1022 때문에 자기가 좋아하는 주제를 정해 시간을 가지고 찬찬히 보는 것도 한 방법이다.

1427 처음 취직 하 았 을 때 가 족 들 의 반 응 은 '매우 기뻐 어 하 았 다 ' 가 38.9%, '비교적 좋아 하 았 다 ' 가 47% .

1485 최근 '내가 좋아하는 계절'로 영화계에 데뷔한 그는 어머니를 빼닮은 눈, 아버지와 비슷한 표정으로 관객을 매혹시키고 있다.

1714 몇 년 후 그 학생의 실력이 좋아서 악기를 바꿔야 할 때는 다시 선생이 그 악기를 사들여 D라는 새로운 학생에게 판다.

Close

Non-continuous Text Anchors

3804 어느 오느간 영연수가 나그네 집에 들다 극사 최고의 오느간 수사도 소운 난 사람이 알고 있는 거리에 이르러 나그네 연주자와 그 사람 사이에 경연이 벌어지게 되었다.

3805 한동안은 팽팽한 실력 대결이 지속되었다.

3806 한쪽이 연주를 마치면 그 끝 부 사람이 자기 연주를 다시 시작하는 식으로 화성의 옷감을 짜나갔다. **Selected Text**

3807 그러나 얼마 뒤 나그네 연주자는 조금씩 대위법과 전조(轉調)의 비술을 쓰기 시작하여 어떤 약절의 확대형과 축소형을 이용하는가 하면 몇 개의 주제를 결합하거나 반대로 진행시키다가 갑자기 아득히 멀리 훌쩍 날아 오르기도 했다.

3808 그 거리의 연주자 역시 흥내를 내 보았지만 실패하고 다시 시도하려다 또 실패를 거듭하면서 끝내는 도저히 따라가지 못하고 기권해 버렸다.

3809 패배 하 나 연주자는 "당신은 제바스티안 바흐가 아닌 면, 하늘에서 내려오 나 천사 이 는 것이 아니다!" 하고 나그네 를 격찬 하 았 다. **Unselected Text**

3810 과연 그 연주자야말로 제바스티안 바흐였던 것이다.

3811 나귀도 성서에는 갈 수 있지만 순례자가 되어 돌아오지는 못한다.

SentID	Tag	Morphemes	Type	Subj-Type	Polar...	Inten...
8809	Seed	하늘 에서 내려오 나	dir-explicit	Judgment-POS	POS	High

SEED Tag Attributes 1

Morphemes Selected
매우, 기쁘, 어, 하 **1**

Tag
 Seed Subjective Objective **2**

Type
 indirect dir-explicit dir-speech **3**
 dir-action writing-device

Subjectivity Type
 Judgment Agreement POS **4**
 Emotion Speculation NEG
 Argument Intention NEUT
 Others COMP

Polarity
 None POS NEG NEUT COMP

Intensity
 None Low Medium High

Nested-Source
 Imp Out

1. Subjectivity Expression's text anchor
2. Type of tag being made
3. How the subjectivity is expressed
4. Specific type of subjectivity being expressed

SEED Tag Attributes 2

Intensity
 None Low Medium High

Nested-Source

	<input checked="" type="checkbox"/> Imp <input type="checkbox"/> Out
가족, 들	<input type="checkbox"/> Imp <input type="checkbox"/> Out
	<input type="checkbox"/> Imp <input type="checkbox"/> Out

Target

취직, 하

Comment

Confident

SAVE CLEAR DELETE

1. Source of the subjectivity (Imp means an implicit source. Out means generic source)
2. Object the subjectivity is directed towards
3. Annotators can mark annotations for review

Subjective Tag Attributes

Morphemes Selected

최근 '내가 좋아하는 계절'로 영화계에 데뷔한 그는 어마

Tag

Seed Subjective Objective

Polarity

None POS NEG NEUT COMP

Intensity

None Low Medium High

Comment

Confident

Objective Tag Attributes

Morphemes Selected

처음 취직했을 때 가족들의 반응은 '매우 기뻐했다'가 3

Tag

Seed Subjective Objective

Comment

Confident

Future Directions and Improvements

Data Independence

- No technical reason why the tool cannot be used with other morphologic rich languages as-is
- **BUT** tool currently requires raw and parsed text to be manually linked in database before annotation can begin
- The ability to populate database from user-specified sources would greatly increase usability
 - How? Must pay careful consideration to data's expected structure

User Defined Schemas

- Currently uses hard-coded KSC annotation schema
- User Defined Schemas needed for project independence
- GATE uses W3C-compliant XML schemas

Collaborative Tools

- Currently must be run locally on a local database
- Manual synchronization becomes more impractical as number of users increases
- Web-based interfaces like GATE Teamware and brat more convenient for users

Increased Tool Support

- Support more database solutions
- Support a wider range of tasks, like GATE
 - Currently only supports Visualization and Annotation

References

- Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* (formerly *Computers and the Humanities*), 39(2/3):164–210.
- Wiebe, J. 2002. Instructions for annotating opinions in newspaper articles. Department of Computer Science Technical Report TR-02-101, University of Pittsburgh.