Morpheme-based Annotation Tool for Korean Text

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

Table of Contents

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 B**보다 좋다**
  - A-*ka* B-*bota cotha*
  - A-SUB B-COMPARATIVE good-SENTEND
  - A is better than B
- B**보다** 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”)

Jamo					**Syllable/*Kulja***	
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
ha To do (base form)		*-mnida* Sentence ending (formal polite)		*hamnida*

Inflecting Korean Verbs

- Inflections may cause phonologic and orthographic reductions/assimilations:

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

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	Base Morphemes
이 밖에 수입품인 일본 마쓰시타 주서기는 전선의 길이가 98cm로 기준(1백 40cm)보다 짧은 것으로 나타났다고 밝혔다.	이 밖 에 수입품 이 ㄴ 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

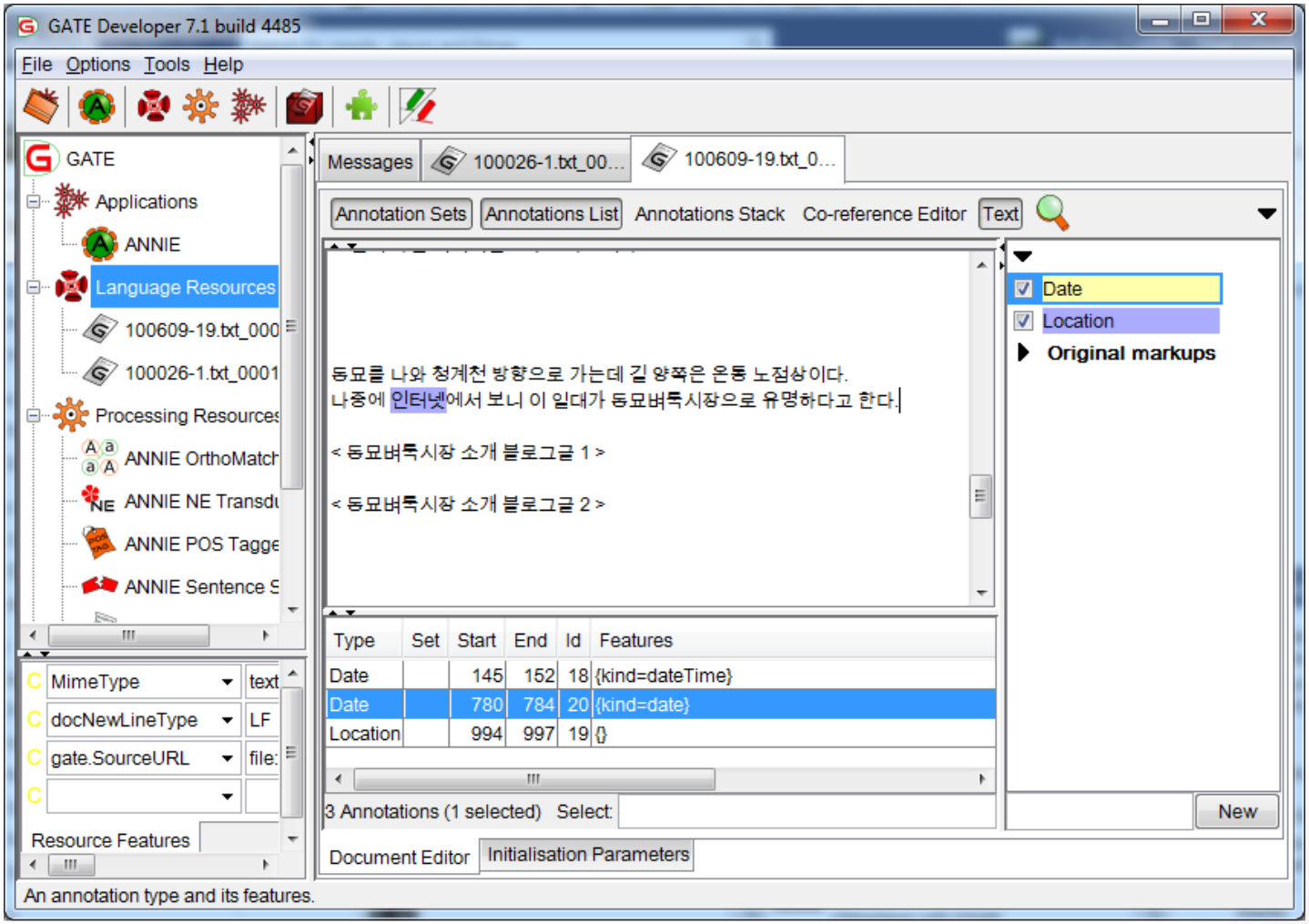
- Ability to switch between fully-inflected text and base morphemes
  - Effectively doubles file size
- Ability to select non-continuous anchors
- Subjectivity Expression, Nested Sources, and Targets should all be created at the same time
  - Separating these attributes incorrectly implies they can exist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


GATE Developer 7.1 build 4485
File Options Tools Help
GATE
Applications
ANNIE
Language Resources
100609-19.txt_000
100026-1.txt_0001
Processing Resources
ANNIE OrthoMatch
ANNIE NE Transdu
ANNIE POS Tagge
ANNIE Sentence S
MimeType text
docNewLineType LF
gate.SourceURL file:
Resource Features
Messages
100026-1.txt_00...
100609-19.txt_0...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
Text
동묘를 나와 청계천 방향으로 가는데 길 양쪽은 온통 노점상이다.
나중에 인터넷에서 보니 이 일대가 동묘벼룩시장으로 유명하다고 한다.
< 동묘벼룩시장 소개 블로그글 1 >
< 동묘벼룩시장 소개 블로그글 2 >
Type Set Start End Id Features
Date 145 152 18 {kind=dateTime}
Date 780 784 20 {kind=date}
Location 994 997 19 {}
3 Annotations (1 selected) Select:
Date
Location
Original markups
New
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


brat
192.168.0.153:8001/index.xhtml#/data/news
/data/news
brat
1 입력 : 2012.01.06 01:49
2 버락 오바마 미국 대통령은 5일(현지시간) 새로운 국방전략을 바탕으로 한 군 개혁으로 미군을 보다 슬림화하지만 세계 최강군으로서의 위치를 유지할 것이라고 강조했다.
Person
GPE
Origin
Org
4 오바마 대통령은 이날 펜타곤에서 가진 국방부의 새 국방전략 발표식에 직접 참석, "미국 군대는 군살을 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.
Person
Employment
GPE
6 버락 오바마 미국 대통령이 5일(현지시간) 펜타곤에서 가진 국방부의 새 국방전략 발표식에 참석해 연설을 하고 있다.
7 /AP연합뉴스
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

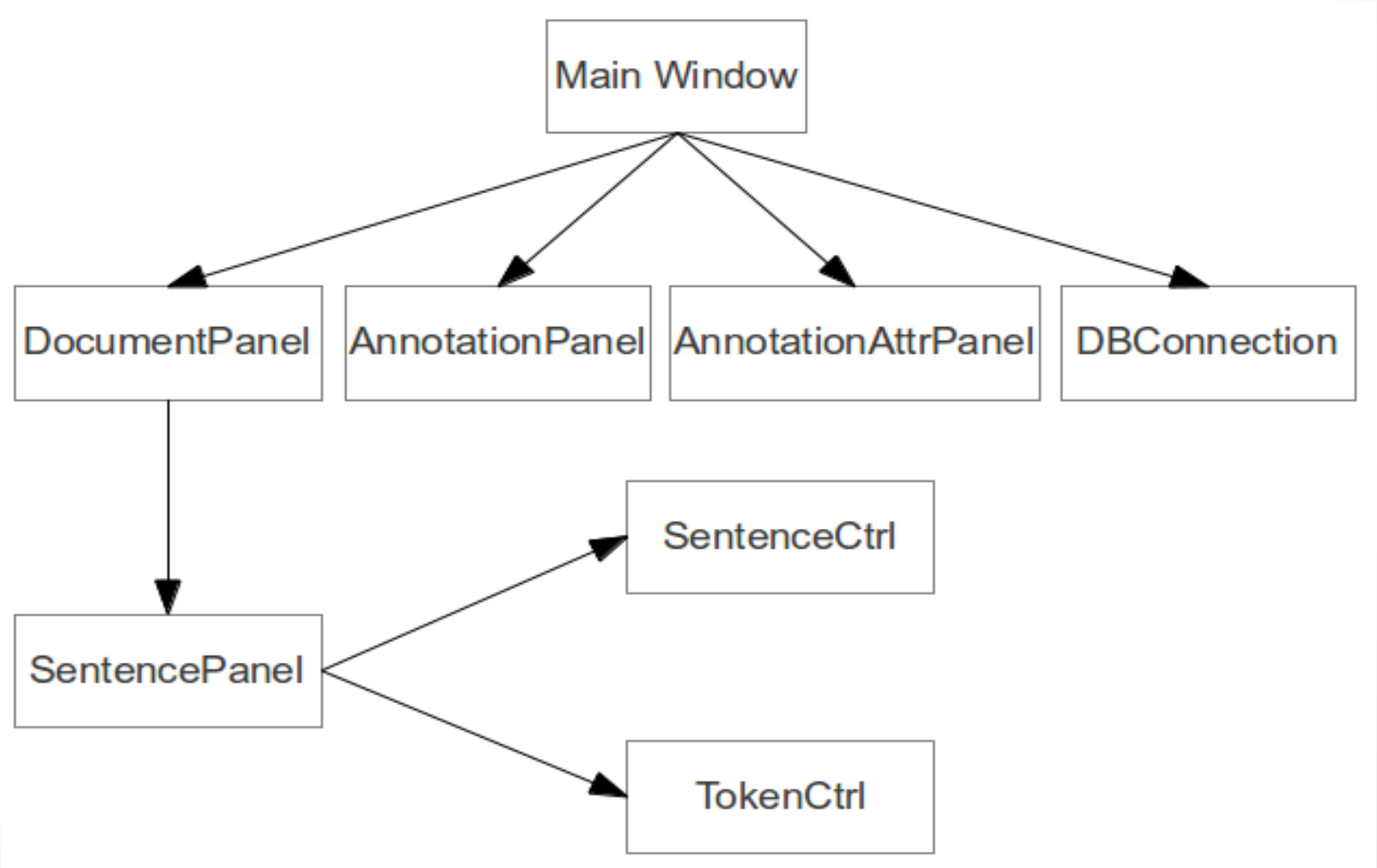
Design Motivation

- 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

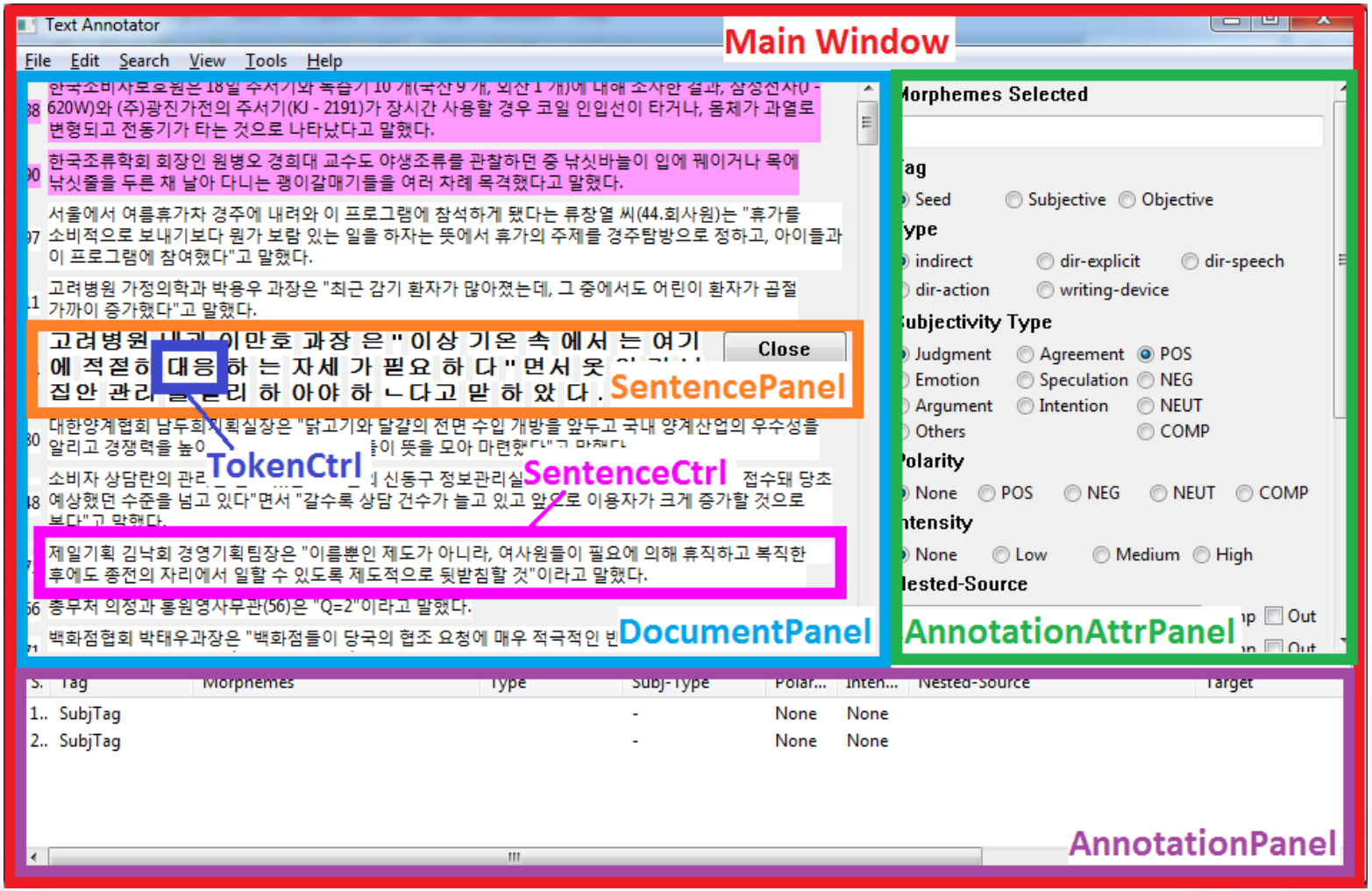
Design Goals

- 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

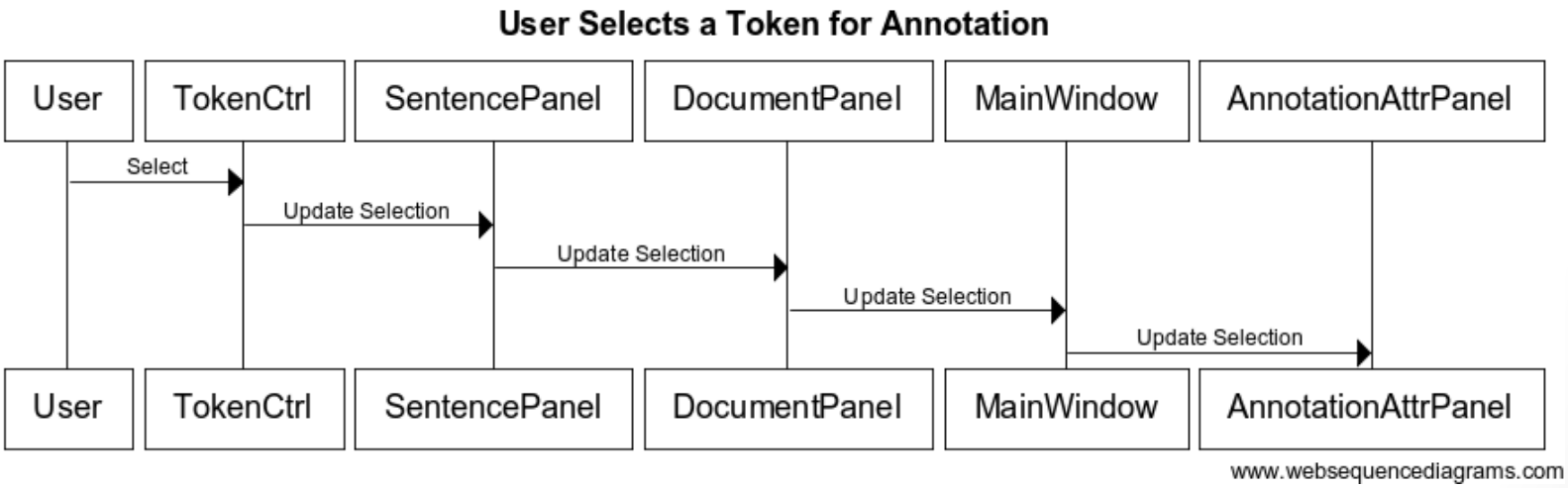
Design Overview



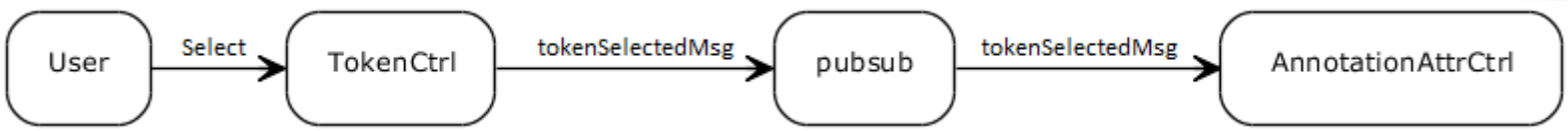
Design Overview



Token Selection w/o PubSub



Token Selection w/ PubSub

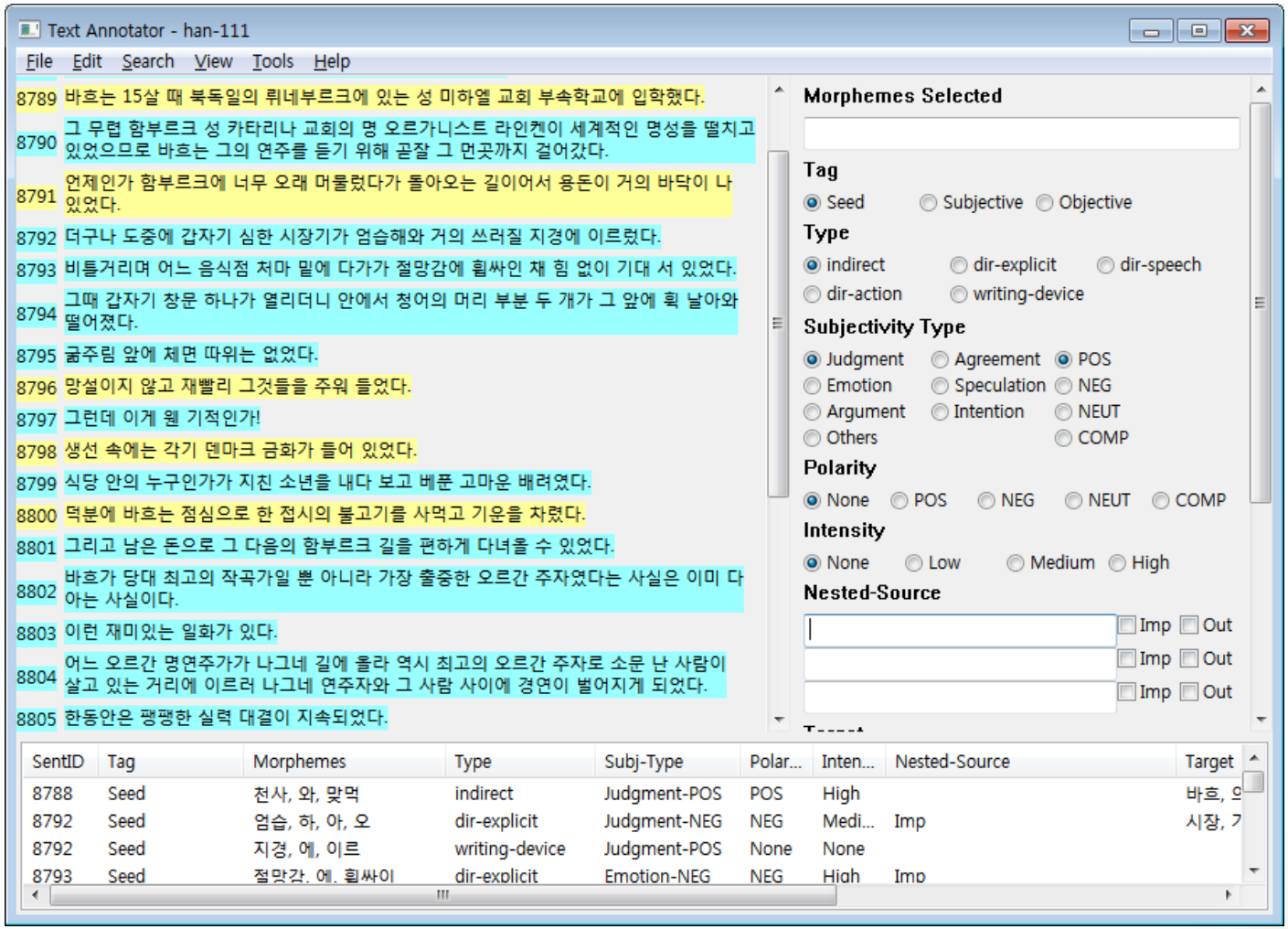


- MainWindow, DocumentPanel, and SentencePanel no longer involved in token selection
- Fewer dependencies
- Asynchronous

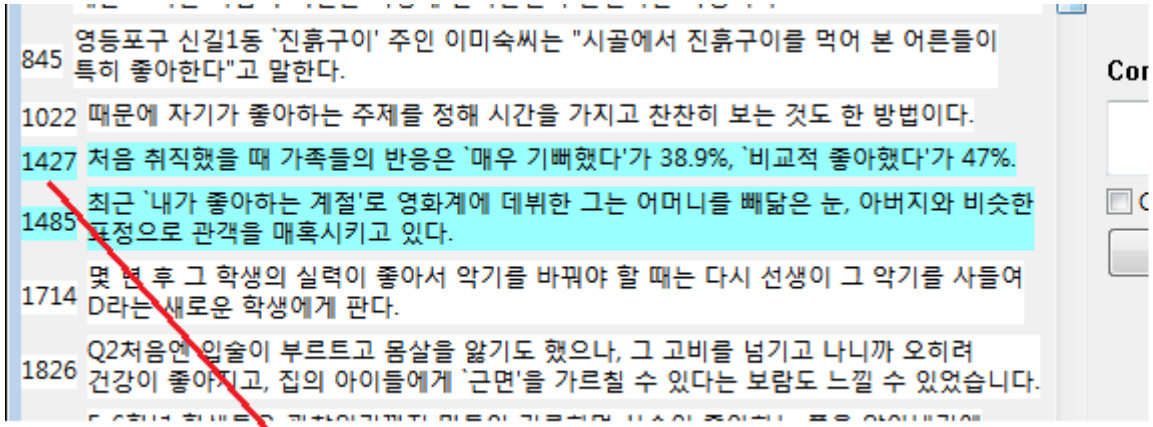
Implementation

- 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

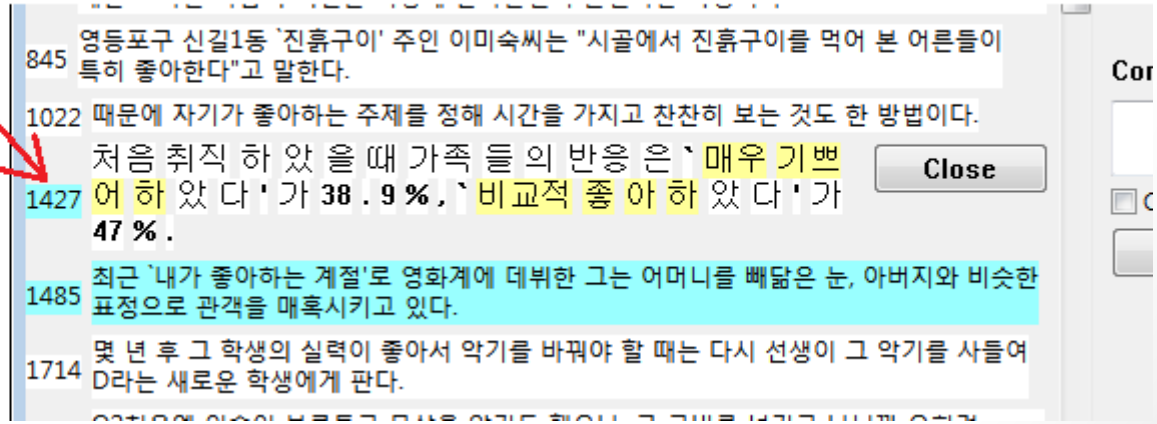


Raw Text vs. Base Morpheme

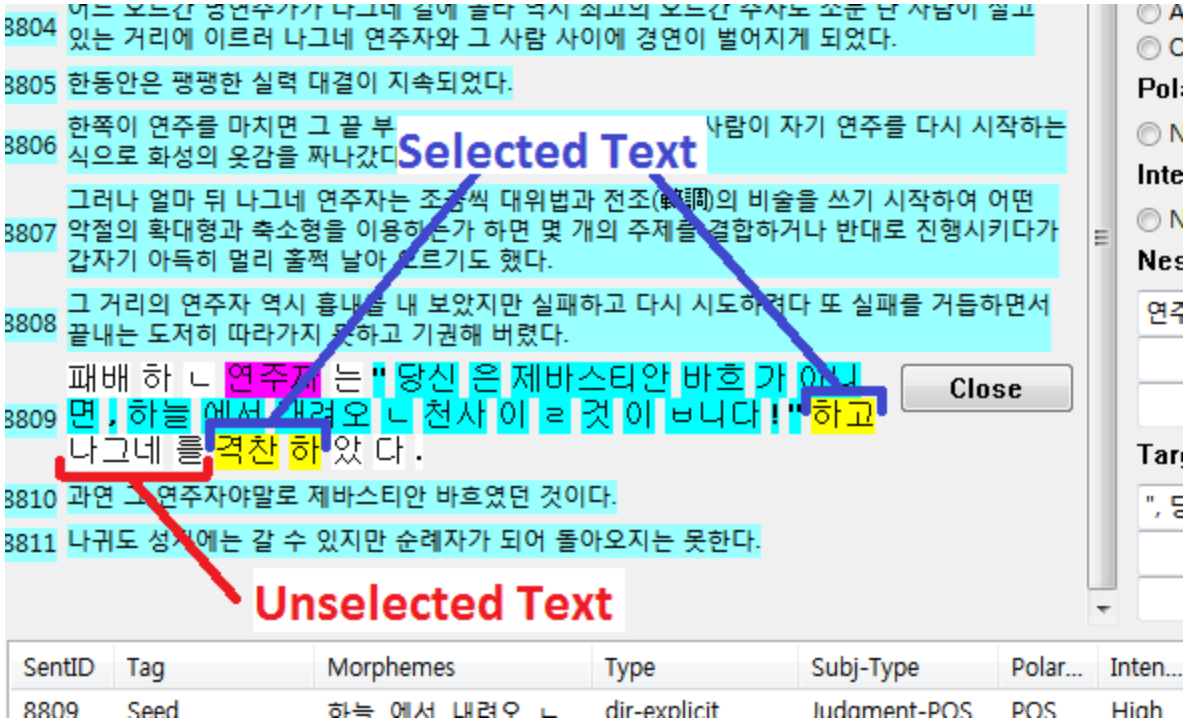


Raw Text View

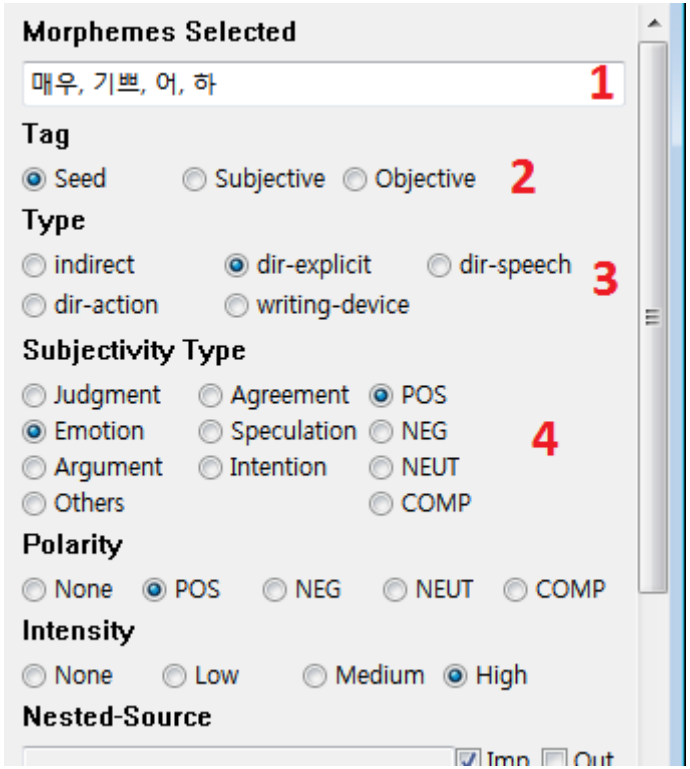
Morpheme View



Non-continuous Text Anchors

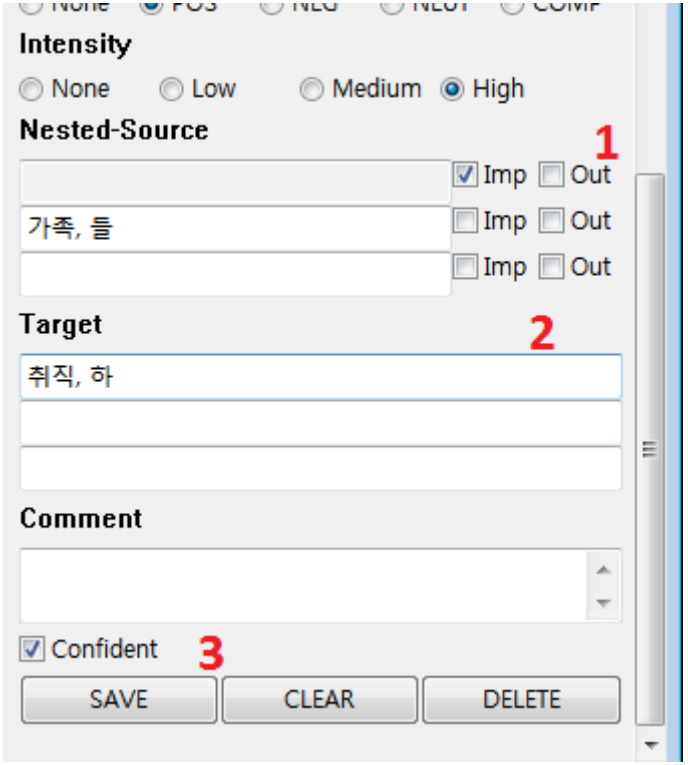


SEED Tag Attributes 1



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

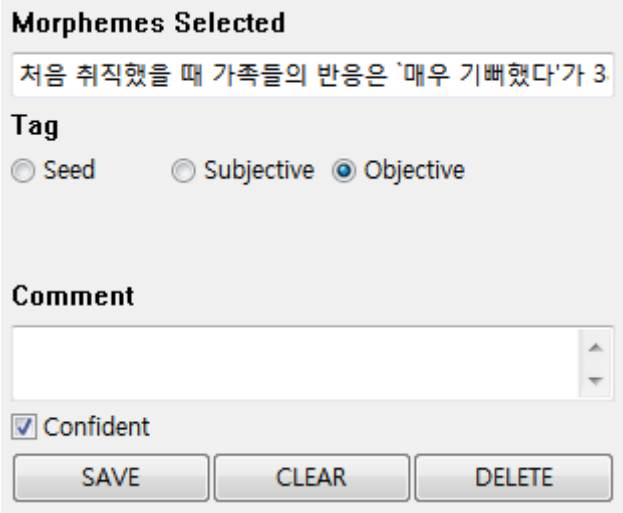
SEED Tag Attributes 2



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

Objective Tag Attributes



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-complient 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.