KTimeML:
Specification of Temporal and Event Expressions in Korean Text

Seohyun Im
Dept. of Computer Science
Brandeis University
Waltham, MA, USA
ish97@cs.brandeis.edu

Hyunjo You, Hayun Jang, Seungho Nam, Hyopil Shin
Dept. of Linguistics
Seoul National University
Seoul, Korea
youhyunjo, hyan05, nam, hpshin@snu.ac.kr

Abstract

TimeML, TimeBank, and TTK (TARSQI Project) have been playing an important role in enhancement of IE, QA, and other NLP applications. TimeML is a specification language for events and temporal expressions in text. This paper presents the problems and solutions for porting TimeML to Korean as a part of the Korean TARSQI Project. We also introduce the KTTK which is an automatic markup tool of temporal and event-denoting expressions in Korean text.

1 Introduction

The TARSQI (Temporal Awareness and Reasoning systems for QA) Project[1] aims to develop technology for annotation, extraction, and reasoning of temporal information in natural language text. The main result of the TARSQI Project consists of TimeML (Pustejovsky et. al., 2003), TimeBank (Pustejovsky et. al., 2006), and TARSQI Toolkit (TTK, Verhagen and Pustejovsky, 2008). TimeML is a specification language for events and temporal expressions in text. TimeBank is an annotated corpus which was made as a proof of the TimeML specification. TTK is an automatic system to extract events and time expressions, creating temporal links between them[2].

TimeML is an ISO standard of a temporal markup language and has been being extended to other languages such as Italian, Spanish, Chinese, etc. (ISO/DIS 24617-1: 2008). TempEval-2, a task for the Semeval-2010 competition, has been proposed (Pustejovsky et. al. 2008). The task for the TempEval-2 is evaluating events, time expressions, and temporal relations. Data sets will be provided for English, Italian, Spanish, Chinese, and Korean.

The necessity of temporal and event expressions markup for any robust performance such as QA (for Korean QA system, refer to Han et. al., 2004), IE, or summarization is applied to Korean NLP applications as well. Recently, there have been TimeML-related studies for Korean: Jang et. al (2004) show an automatic annotation system of temporal expressions with Timex2 in Korean text. Lee (2008) argues about the semantics of Korean TimeML, specially the EVENT tag. Im and Saurí (2008) focus on the problems of TimeML application to Korean caused by typological difference between English and Korean. Motivated by them, the Korean TARSQI Project[3] started with the purpose of making TimeML, TimeBank and TTK for Korean text[4].

Porting TimeML to other languages can be challenging because of typological difference between languages. In this paper, we present the problems for TimeML application to Korean. Our solution is to change TimeML markup philosophy: a change from word-based in-line annotation to morpheme-based stand-off annotation. Based on the changed annotation philosophy, we decide how to annotate temporal and event-denoting expressions in Korean text. More specifically, it is challenging to decide whether we use LINK tags or attributes to annotate some

[1] Refer to www.timeml.org for details on the TARSQI.

[2] TTK contains GUTime (TIMEX3 tagging, Mani and Wilson, 2000), Evita (event extraction, Saurí et. al., 2005), Slinket (modal parsing, Saurí et. al., 2006b), S2T, Blinker, Classifier, Sputlink, Link Merger, etc.

[3] See http://word.snu.ac.kr/k-tarsqi/doku.php for more information about the KTARSQI Project.

[4] James Pustejovsky gave a talk about TARSQI for KTARSQI Project, visiting Korea for his invited talk at CIL 18 conference in 2008.

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 115–122,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

temporal or event-denoting expressions (see examples in 3.2). In section 4, we describe the specification of Korean TimeML (KTimeML). Section 5 introduces Korean TTK (KTTK). Before discussing the issues of Korean TimeML, we briefly introduce TimeML.

2 The Basics of TimeML

TimeML features four major data structures: EVENT, TIMEX3, SIGNAL, and LINK. The **EVENT** tag encodes event-denoting expressions. The **TIMEX3** tag annotates temporal expressions of different sorts: fully specified dates, times, and durations, or just partially specified dates, times, and durations. The **SIGNAL** tag annotates elements that indicate how temporal objects are related among them (e.g., subordinating connectors such as *when* or *after*).

The LINK tag splits into three main types: (a) **TLINK**, which encodes temporal relations among EVENTs and TIMEX3s; (b) **ALINK**, representing aspectual information as expressed between an aspectual predicate and its embedded event; and (c) **SLINK**, encoding subordination relations conveying evidentiality (e.g. *Mary* ***said*** *[she bought some wine*]), factivity (*John* ***regretted*** *[Mary bought wine]*), or intensionality (*Kate* ***thought*** *[Mary bought beer]*).

Information relevant to each tag is characterized by means of attribute-value pairs (refer to Pustejovsky et. al. 2003 about specific attributes-value pairs). (1) illustrates an annotated sentence with the TimeML specification:

(1) *John* said$_{e1}$ *that Mary* began$_{e2}$ *to* work$_{e3}$

```
John
<EVENT id="e1" class="REPORTING"
tense="PAST" aspect="NONE" polar-
ity="POS">
said </EVENT>
that Mary
<EVENT id="e2" class="ASPECTUAL"
tense="PAST" aspect="NONE" polar-
ity="POS">
began </EVENT>
to
<EVENT id="e3" class="OCCURRENCE"
tense="NONE" aspect="NONE" polar-
ity="POS">
work </EVENT>

<TLINK eventID="e1" relatedToEvent="e2"
relType="AFTER"/>
<SLINK eventID="e1" subordinatedEvent="e2"
relType="EVIDENTIAL"/>
<ALINK eventID="e2" relatedToEvent="e3"
relType="INITIATES"/>
```

Sentence (1) presents three EVENT expressions (*said, began,* and *work*). SLINK conveys an evidential relation between *e1* (*said*) and *e2* (*began*). TLINK represents a temporal relation – AFTER- between the two same events. ALINK encodes an aspectual relation –initiates– between *e2* (*began*) and *e3* (*work*). Due to space limitations, some EVENT attributes are obviated.

3 Porting TimeML to Korean

3.1 The Characteristics of Korean

Korean is an agglutinative language whose words are formed by joining morphemes together, where an affix typically represents one unit of meaning and bound morphemes are expressed by affixes. For example, the sentence *John-i emeni-kkeyse o-si-ess-ta-te-ra* 'John-Nom mother-Nom come-Hon-Past-Quo-Ret-Dec [5] ' means that (I heard) (John said) that his mother came. Each morpheme has its own functional meaning or content.

As shown above, consideration of morphemes is important for TimeML markup of Korean text. Here, we summarize TimeML-related characteristics of Korean:

(i) In Korean, functional markers (tense, aspect, mood, modality, etc.) are represented morphologically. English as an isolating language uses periphrastic conjugation to represent functional categories.
(e.g. '*-keyss-*'is a conjectural modal morpheme in *pika o-keyss-ta* 'it will rain'. While, '*will*' is an auxiliary verb in *it will rain*.)

(ii) Some subordination is realized morphologically via morpheme contraction.
(e.g. '*-ta-n-ta*' is a morphological contraction which denotes quotation in the sentence *John-i nayil o-n-ta-n-ta* 'John-Nom tomorrow come-Pres-Dec.Quo-Pres-Dec'. Its English counterpart is represented by subordination: *John said that he will come tomorrow*)

(iii) Some connectives in English correspond to morphemes in Korean.
(e.g. Korean counterpart of the English connective '*and*' in *I ate milk and went to sleep* is the morpheme '*-ko*' in the sentence *na-nun wuyu-rul masi-ko ca-re ka-ss-ta* 'I-Top milk-Acc drink-and sleep-ending go-Past-Dec')

(iv) The sentence type of English is represented by word order but that of Korean by ending morphemes
(e.g. Declarative: *pi-ka o-n-ta* 'it is raining' interrogative: *pi-ka o-ni?* 'Is it raining?')

[5] Nom: nominative case, Hon: honorific morpheme, Past: past tense morpheme, Quo: quotative mood morpheme, Ret: retrospective mood morpheme, Dec: declarative sentence ending

These properties of Korean make the porting of TimeML to Korean challenging. In the next section, we discuss the basic issues of KTimeML.

3.2 Basic Issues of Korean TimeML

3.2.1 Morpheme-based standoff annotation

TimeML employs word-based in-line annotation. It poses a challenge at the representation level, since it encodes information mainly based on the structure of the target language, and thus content equivalences among different languages are hard to establish. For example, indirect quotation in Korean offers an example of the mismatch of linguistic devices employed in different languages to express the same meaning. Quotation constructions in English use two predicates, the reporting and the reported, which TimeML marks up as independent EVENTs:

(2) *John* said$_{e1}$ *he* bought$_{e2}$ *a pen.*

```
<SLINK eventID="e1" subordinatedE-
vent="e2"relType="EVIDENTIAL"/>
```

TimeML uses a subordination link (SLINK) in order to convey the evidentiality feature that the reporting predicate projects to the event expressed by its subordinated argument.

On the other hand, a Korean quotative construction, as in (3), has only one verb stem, which corresponds to the subordinated predicate in English. Note that there is no reporting predicate such as *say* in English. Nevertheless, the sentence has a reporting interpretation.

```
(3) John-i  ku-ka   wine-ul  sa-ss-ta-n-ta
    J-Nom   he-Nom  wine-Acc buy-Past-Quo-Pres-Dec
    'John said that he bought some wine'
```

The quotative expression *–ta-n-ta* above is a contracted form of *–ta-ko malha-n-ta* 'Dec-Quo say-Pres-Dec'. Although (3) is a simple sentence involving no subordination at the syntactic level, the two tense markers, '*-ss-*' and '*-n-*', are evidence of the existence of an implicit reporting event. Specifically, the past tense marker '*-ss-*' applies to the main event here (*sa-ss* 'buy-past'), while the present tense marker '*-n-*' is understood as applying to the implicit reporting event (*ta-n-ta* 'report-pres-Dec)[6].

Constructions presented above show a problem for the standard TimeML treatment of a Korean quotative sentence. The relationship between reporting and reported events is expressed morphologically, and thus the SLINK mechanism for word-based annotation is not adaptable here. Because Korean transfers meanings through morphological constructions, morpheme-based annotation is more effective than word-based for TimeML application to Korean[7].

For morpheme-based tagging, we propose stand-off annotation for Korean because it needs two-level annotation: the MORPH tag[8] and TimeML tags. Standoff annotation separates morphologically-annotated data from primary data and saves it in a different file, and then TimeML annotation applies to the data. The following is the proposed morpheme-based stand-off annotation for (3).

```
(4) Morpheme-based stand-off annotation for (3)
   <MORPH id="m7" pos="PV"/>
   <MORPH id="m8" pos="EFP"/>
   <MORPH id="m9" pos="EFP"/>
   <MORPH id="m10" pos="EFP"/>
   <MORPH id="m11" pos="EF"/>
   <EVENT id="e1" morph="m7 m8" yaleRo-
   manization="sa-ss" pred="buy"
   class="OCCURRENCE" tense="PAST" sen-
   tenceMood="DEC"/>
   <EVENT id="e2" morph="m9 m10 m11"
   yaleRomanization="ta-n-ta" pred="say"
   class="REPORTING" tense="PRESENT" sen-
   tenceMood="DEC"/>
   <SLINK eventID="e2" subordinatedE-
   vent="e1" relType="EVIDENTIAL"/>
   <TLINK eventID="e1" relatedToEvent="e2"
   relType="BEFORE"/>
```

In (4), we show the example annotation of the MORPH tag for (3) to help readers to understand our proposal. Standoff annotation makes it possible to extract information about two events without using a non-text consuming EVENT tag. Moreover, each of the two tense morphemes is properly assigned to its related event. Our proposed TimeML annotation scheme is composed of two levels – morphological analysis and TimeML annotation.

[6] Tense markers of the construction can change: *sa-**ss**-tay-**ss**-ta* 'buy-**past**-quo-**past**-dec: *said_bought*'; *sa-**n**-ta-**n**-ta* 'buy-**pres**-quo-**pres**-dec: *say_buy'*, etc.

[7] There can be several ways of annotating morphological constructions: morpheme-based, morpho-syntactic unit-based (refer to MAF: Clément and Clergerie, 2005), character-based, and bunsetsu-based. At present, we adopt morpheme-based annotation because it seems to be enough to introduce the required units for KTimeML markup and we want to avoid the possible redundancy of bunsetsu-based or morpho-syntactic unit-based annotation. Moreover, the criterion for separation of a morphological construction is related with tags such as EVENT, TIMEX3, or attributes like tense, aspect, mood, or modality in KTimeML, not with syntactic or phonological information. Standoff annotation makes it easy to mark up the interval of morphemes. Nevertheless, we consider the possible advantage of morpho-syntactic analysis positively for future work.

[8] The values of the POS attribute are based on a Korean Part_of_Speech Tag Set version 1.0 (Kim and Seo, 1994).

3.2.2 Surface-based annotation

KTimeML adopts the surface-based annotation philosophy of TimeML (Saurí et. al. 2006a), which does not encode the actual interpretation of the constructions it marks up, but their grammatical features. For example, the *leaving* event in the sentence *we are leaving tomorrow* is not annotated as expressing a future tense, but as expressed by means of a present tense form. Several considerations motivate this surface-based approach. As an annotation language, it must guarantee the marking up of corpora in an efficient and consistent way, ensuring high inter-annotator agreement. As a representation scheme, it needs to be used for training and evaluating algorithms for both temporal information extraction and temporal reasoning.

A surface-based approach is the suitable option for meeting such requirements. Nevertheless, it poses a challenge at the representation level. How to represent evidentiality in Korean and English shows the challenge.

(5) *I saw$_{e1}$ that John bought$_{e2}$ some wine.*

```
<SLINK lid="sl1" eventID="e1" subordinatedEvent="e2" relType="EVIDENTIAL"/>
```

English, as an isolating language, expresses evidentiality in a periphrastic manner. Hence, the TimeML treatment of these constructions consists in marking the two involved predicates as EVENTs, and introducing an SLINK between them. Korean has both periphrastic and morphological ways for expressing evidentiality. Annotating the periphrastic version with the standard TimeML treatment poses no problem because it has two predicates denoting events like its English counterpart. Morphological constructions however, are harder to handle, because the retrospective mood morpheme '*-te-*' brings about the implicit reference to a seeing event.

```
(6) Vietnam-un   tep-te-ra
    Vietnam-Top  hot-Ret-Dec
    '(as I saw) Vietnam was hot'
```

They are similar to quotative constructions in the sense that, although there is only one predicate expressed on the surface, the sentence refers to more than one event. Unlike quotative constructions, there is no morphological evidence of the implicit event; e.g. tense or sentence mood markers independent of those applied to the only verbal predicate in the sentence. The issue to consider is therefore whether to treat the evidential constructions by introducing an EVENT tag for the retrospective mood marker as in (7) or to handle them by specifying the evidential value of the main predicate at the MOOD attribute of its EVENT tag, as illustrated in (8).

(7) SLINK tagging for (6)

```
<EVENT id="e1" morph="m3" yaleRomanization="tep" class="STATE" pos="ADJECTIVE" tense="NONE"/>
<EVENT id="e2" morph="m4 m5" yaleRomanization="te-ra" class="PERCEPTION" pos="NONE" tense="NONE"/>
<SLINK lid="sl1" eventID="e2" subordinatedEvent="e1" relType="EVIDENTIAL"/>
```

(8) Mood-attribute tagging for (6)

```
<EVENT id="e1" morph="m3 m4 m5" yaleRomanization="tep-te-ra" pred="hot" class="STATE" pos="ADJECTIVE" tense="NONE" mood="RETROSPECTIVE"/>
```

As in (7), adding an EVENT tag for the retrospective morpheme corresponds semantically to English-based TimeML. However, it is not surface-based, because the perception event is an implicit event entailed by the retrospective morpheme. While, the annotation in (8) is a surface-based annotation of the evidential construction which uses the MOOD attribute for retrospective mood, thus respects the surface-based philosophy of TimeML. This is different from the English counterpart that presents two EVENTs related with a TLINK signaling their relative temporal order. KTimeML follows the surface-based annotation philosophy of TimeML ((8) here).

3.2.3 Cancellation of the head-only rule

TimeML employs the head-only markup policy in order to avoid problems derived from tagging discontinuous sequence (e.g. *we* **are** *not fully* **prepared**). If the event is expressed by a verbal phrase, the EVENT tag will be applied only to its head, which is marked in bold face in the examples (e.g. *has been* ***scrambling****, to* ***buy****, did not* ***disclose***). However, Korean does not have the discontinuity problem. See Korean examples:

```
(9) a.*na-nun cwunpitoy-e    wanpyekhakey  iss-ta
       I-Top  prepared-e    fully         exist-Dec
      'we are fully prepared'
    b. *John-un  ca-ko        anh-iss-ta
        J-Top    sleep-ko     Neg-exist-Dec
        'John is not sleeping'
```

In the above sentences, '*-e iss-*' and '*-ko iss-*' are respectively perfective and progressive aspect markers. No word can make discontinuous sequence by being embedded into the middle of the verb phrases. As we saw from the examples, Korean does not have discontinuity problem in verbal phrases. Thus, KTimeML does not need to follow the head-only annotation rule. By cancellation of the head-only rule, we annotate various verbal clusters (main verb + auxiliary verb construction: e.g. *mek-ko iss-ta* 'eat-progressive-dec'). It makes the KTimeML more readable by showing the progressive aspect-denoting expression *-ko iss-* in one unit of annotation.

4 Specification of the Korean TimeML

Based on the proposed annotation principles of KTimeML, we present the specification of the first version of KTimeML (KTimeML 1.1) with changed tags, attributes, and their values. We assume that the MORPH-tagged data are separately saved in a different file. KTimeML contains EVENT, TIMEX3, SIGNAL, and LINK tags. Some new attributes such as `mood` and `sType` are added to the attributes of the EVENT tag. The other tags have no changes from the TimeML tags[9].

KTimeML 1.1 adds the attributes of predicate_content (`pred`), `mood`, verb_form (`vForm`), and sentence type (`sType`) to the attributes of EVENT in TimeML (For Korean grammar, refer to Sohn, 1999, Nam and Ko, 2005). The BNF of EVENT is shown below:

```
attributes ::= id pred morph yaleRomanization
                class pos tense [aspect][mood]
               [sType][modality] vForm
id ::= ID
{id ::= EventID
 EventID ::= e<integer>}
morph ::= IDREF
{morph ::= MorphID}
yaleRomanization ::= CDATA
pred ::= CDATA
class ::= 'OCCURRENCE'|'ASPECTUAL'|'STATE'|
          'PERCEPTION'|'REPORTING'|'I_STATE'|
          'I_ACTION'
pos ::= 'ADJECTIVE'|'NOUN'|'VERB'|'OTHER'
tense ::= 'PAST'|'NONE'
aspect ::= 'PROGRESSIVE'|'PERFECTIVE'|
           'DURATIVE' | 'NONE'
mood ::= 'RETROSPECTIVE' | 'NONE'
          {default, if absent, is 'NONE'}
sType ::= 'DECLARATIVE'|'INTERROGATIVE'|
          'IMPERATIVE'|'PROPOSITIVE'| 'NONE'
         {default, if absent, is 'DECLARATIVE'}
modality ::= 'CONJECTUAL'|'NONE'
          {default, if absent, is 'NONE'}
vForm ::= 'S_FINAL'|'CONNECTIVE'|'NOMINALIZED'|
          'ADNOMINAL'
          {default, if absent, is 'S_FINAL'}
polarity ::= 'NEG'|'POS'
          {default, if absent, is 'POS'}
```

KTimeML puts the semantic content of EVENT-tagged expressions for international communication. Because mood is not an important grammatical category for English, TimeML does not markup a mood attribute, but KTimeML adds the mood attribute since there are morphemes that express mood like many other languages. Unlike English, different sentence ending morphemes represent sentence types in Korean. Hence, KTimeML adds `sType` to attributes of the EVENT tag. We put `vForm` to distinguish between different subordinated clauses[10].

Event classes in KTimeML are the same as TimeML. Korean tense system does not have distinction between present and future unlike English, and thus the tense attribute has PAST and NONE values. We add DURATIVE to aspect attribute values in KTimeML for the durative expression such as combination of stative verb + progressive aspect marker (e.g. *al-ko iss-ta* 'know-durative-Dec').

For `mood`, KTimeML 1.1 puts the retrospective mood ('*-te-*'). The values of `vForm` attribute are S_FINAL, CONNECTIVE, and NOMINALIZED, and ADNOMINAL. The sentence types in Korean are DECLARATIVE, INTEROGGATIVE, IMPERATIVE, and PROPOSITIVE (e.g. *cip-ey ka-**ca*** 'Let's go home'). KTimeML puts CONJECTURAL (e.g. *nayil pi-ka o-**keyss**-ta* '(I guess) It will rain tomorrow') as a modality value and default is NONE. The sentence in (10) is an interesting example that includes all attributes of an EVENT tag for Korean TimeML except for aspect.

```
(10)ecey  Seoul-un  pi-ka  o-ass-keyss-te-ra
     yesterday Seoul-Top rain-Nom come-Past-Conj-Ret-Dec
     '(From that I saw), I guess that it rained in Seoul
      yesterday'

   <EVENT id="e1" morph="m6 m7 m8 m9 m10"
    yaleRomanization="wa-ss-keyss-te-ra"
    pred="come" pos="VERB"
    class="OCCURRENCE" tense="PAST"
    aspect="NONE" mood="RETROSPECTIVE"
    modality="CONJECTURAL" vForm="S_FINAL"
    sType="DECLARATIVE" polarity="POS"/>
```

Each of the morphemes above has its own functional meaning, which is represented as a value of an attribute in the EVENT tag.

The major types of TIMEX3 expressions are: (a) Specified Temporal Expressions, *2009-nyen 5-wol 1-il* '2009-year 5-month 1-day', (b) Underspecified Temporal Expressions, *wolyoil* 'Monday', *caknyen* 'last year', *ithul cen* 'two days ago'; (c) Durations, 2 *kaywol* '2 months', 10 *nyen* 'ten years'.

```
attributes ::= tid type [functionInDocument]
              [temporalFunction] morph
               yaleRomanization
              (value|valueFromFunction)
              [mod][anchorTimeID|anchorEventID]
```

[9] Nevertheless, how to annotate various morphological constructions in the specific texts is not trivial. The annotation guideline, which will be published on the web, will handle the issues in detail.

[10] ISO-TimeML also has `pred, mood, and vForm.`

```
tid ::= ID
{tid ::= TimeID
 TimeID ::= t<integer>}
morph ::= IDREF
{morph ::= MorphID}
yaleRomanization ::= CDATA
type ::= 'DATE'|'TIME'|'DURATION'
functionInDocument ::= 'CREATION_TIME'|
         'EXPIRATION_TIME'|'MODIFICATION_TIME'|
         'PUBLICATION_TIME'|'RELEASE_TIME'|
         'RECEPTION_TIME'|'NONE'
temporalFunction ::= 'true'|'false'
         {temporalFunction ::= boolean}
value ::= CDATA
         {value ::= duration|dateTime|
                    time|date|gYearMonth|
                    gYear|gMonthDay|
                    gDay|gMonth}
valueFromFunction ::= IDREF
{valueFromFunction ::= TemporalFunctionID
TemporalFunctionID ::= tf<integer>}
mod ::= 'BEFORE'|'AFTER'|'ON_OR_BEFORE'|
        'ON_OR_AFTER'|'LESS_THAN'|'MORE_THAN'|
        'EQUAL_OR_LESS'|'EQUAL_OR_MORE'|'START|
        'MID'|'END'|'APPROX'
anchorTimeID ::= IDREF
        {anchorTimeID ::= TimeID}
comment ::= CDATA
```

Although the BNF of TIMEX3 in Korean TimeML is same as that of TimeML, we point out that Korean time expressions also have the issue of how to treat morphological representations of temporal meaning. For example, *pwuthe* 'from' and *kkaci* 'to' in 3*ilpwuthe* 5il*kkaci* 'From 3rd to 5th' both are the counterparts of prepositions in English (Jang et. al., 2004). We do not tag temporal morphemes as SIGNALs, in principle. Instead, we mark up 3*ilpwuthe* 'from 3rd' with one TIMEX3 tag. However, temporal connectives such as *ttay* 'when' in *ku-ka o-ass-ul ttay younghee-nun ttena-ss-ta* 'When he came, Younghee left' are tagged as SIGNALs.

SIGNAL is used to annotate sections of text - typically function words - that indicate how temporal objects are to be related to each other. It includes temporal connectives (e.g. *ttay* 'when', *tongan* 'during'), and temporal noun (e.g. *hwu* 'after', *cen* 'before'). See the BNF of SIGNAL below:

```
attributes ::= sid morph yaleRomanization
sid ::= ID
{sid ::= SignalID
SignalID ::= s<integer>}
morph ::= IDREF
{morph ::= MorphID}
yaleRomanization ::= CDATA
```

We show an annotated example which describes the difference of Korean TimeML markup from the English-based TimeML. The sentence below is a compound sentence.

```
(11) ku-nun hankwuk panghan-ul maci-n hwu,
     Ku-Top  Korea    visit-Acc   finish after

     onul  cwungkwuk-uro ttena-ss-ta
      today China-for     leave-Past-Dec
      'He finished his visit to Korea
      and left for China today'
```

```
<Document time: March, 20, 2009>

<EVENT id="e1" morph="m4 m5" yaleRomaniza-
   tion="pangmwun-ul"
 pred="visit" class="OCCURRENCE"/>
<EVENT id="e2" morph="m6 m7" yaleRomaniza-
   tion="machi-n" pred="finish"
   class="ASPECTUAL" pos="VERB"
   tense="NONE" vForm="ADNOMINAL"/>
<SIGNAL sid="s1" morph="m8" yaleRomaniza-
   tion="hwu"/>
<TIMEX3 tid="t1" morph="m9" yaleRomaniza-
   tion="onul" type="DATE" value="2009-03-
   20" temporalFunction="true"/>
<EVENT id="e3" morph="m14 m15 m16" yaleRo-
   manization="ttena-ss-ta"
 pred="leave" class="OCCURRENCE"
 tense="PAST" sType="DECLARATIVE"
 vForm="S_FINAL"/>
```

LINK types splits into TLINK, SLINK, and ALINK. The BNF of TLINK is as follows:

```
attributes ::= [lid] (eventID|timeID)
               [signalID] (relatedToEvent|
               relatedToTime) relType [comment]
lid ::= ID
{lid ::= LinkID
 LinkID ::= l<integer>}
eventID ::= IDREF
{eventID ::= EventID}
timeID ::= IDREF
{timeID ::= TimeID}
signalID ::= IDREF
{signalID ::= SignalID}
relatedToEvent ::= IDREF
{relatedToEvent ::= EventID}
relatedToTime ::= IDREF
{relatedToTime ::= TimeID}
relType ::= 'BEFORE'|'AFTER'|INCLUDES'|
            'IS_INCLUDED'|'DURING'|
            'SIMULTANEOUS'|'IAFTER'|'IBEFORE'|
            'IDENTITY'|'BEGINS'|'ENDS'|
            'BEGUN_BY'|'ENDED_BY'|'DURING_INV'
comment ::= CDATA
```

TLINK is a temporal link among EVENTs and TIMEX3s. For example, three TLINKs are tagged between the events in (11). We show those together with other LINKs in (12). Now, we show the BNF of SLINK.

```
attributes ::= [lid] eventID [signalID]
               subordinatedEvent relType
               [comment]
lid ::= ID
{lid ::= LinkID
 LinkID ::= l<integer>}
eventID ::= IDREF
{eventID ::= EventID}
subordinatedEvent ::= IDREF
{subordinatedEvent ::= EventID}
signalID ::= IDREF
{signalID ::= SignalID}
```

```
relType ::= 'INTENTIONAL'|'EVIDENTIAL'|
            'NEG_EVIDENTIAL'|'FACTIVE'|
            'COUNTER_FACTIVE'|'CONDITIONAL'
comment ::= CDATA
```

The subordination link is used for contexts involving modality, evidentials, and factives.

In Korean, various morphemes bring about subordination clauses. Nominal endings such as -*um/-ki* make nominal clauses (e.g. *na-nun John-i o-ass-**um**-ul al-ko iss-ta* 'I-Top John-Nom come-Past-Nominal ending-Acc know-Durative-Dec'; *na-nun kongpwuha-**ki**-ka shilh-ta* 'I-Top study-nominal ending-Nom hate-Dec'). Adnominal endings such as *-n/-un/-nun* make adnominal clauses (e.g. *na-nun John-i kaci-e-o-**n** kwaca-rul mek-ess-ta* 'I-Top John-Nom bring-adnominal ending cookies-Acc eat-Past-Dec'). Conditional clauses are also triggered by morphemes (e.g. *na-nun John-i o-**myen** ka-keyss-ta* 'I-Top John-Nom come-Conditional go-Conj-Dec'). All the above morphemes are not separately tagged as SIGNALs. The words with the morphemes – *o-ass-um-ul, kongpwuha-ki-ka, kaci-e-o-n,* and *o-myen* – are tagged as EVENTs.

ALINK is an aspectual link which indicates an aspectual connection between two events.

```
attributes ::= [lid] eventID [signalID]
               relatedToEvent relType
               [comment]
lid ::= ID
{lid ::= LinkID
 LinkID ::= l<integer>}
eventID ::= IDREF
{eventID ::= EventID}
relatedToEvent ::= IDREF
{relatedToEvent ::= EventID}
signalID ::= IDREF
{signalID ::= SignalID}
relType ::= 'INITIATES'|'CULMINATES'|
            'TERMINATES'|'CONTINUES'|
            'REINITIATES'
comment ::= CDATA
```

Now we show the ALINK and TLINKs of the sentence in (11).

(12) LINKs between the events in (11)

<ALINK eventID="e2" relatedToEvent="e1" **relType="CULMINATES"**/>

<TLINK eventID="e3" relatedToEvent="e2" **relType="AFTER"**/>
<TLINK eventID="e2" relatedToEvent="e1" **relType="ENDS"**/>
<TLINK eventID="e3" relatedToEvent="e1" **relType="AFTER"**/>

That is, the *visiting* event and the *finishing* are related aspectually and its relation type is culminating. The *finishing* event is related temporally with the *leaving* event by the signal '후'('after'). Naturally, the relation type of the TLINK is AFTER. From ALINK, additional TLINKs are derived between *visiting, finishing*, and *leaving* events.

5 Korean TARSQI ToolKit

Based on the specification of KTimeML, we started to develop KTTK[11].

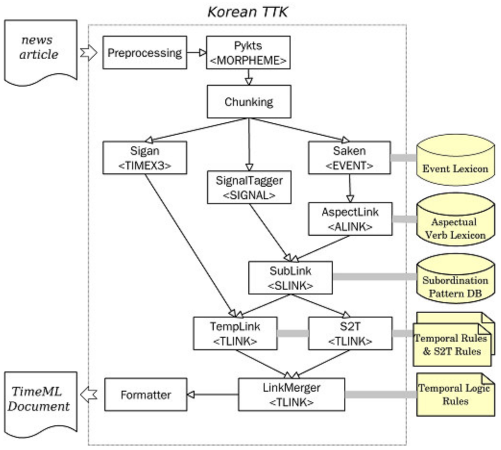

Figure 1. Korean TARSQI Architecture

At first, the normalization of the raw document is done in the preprocessor module. Here the raw text is separated into sentences, wide characters are substituted by regular characters, punctuation symbols are normalized (specially quotation marks), sino-korean characters (hanja) are transcribed in hangul, and, the encoding is also normalized to unicode.

The next module is called Pykts (Python Wrapper for KTS). Here, sentences are parsed in order to get their morphological components, which is achieved by means of a program called KTS. With the exception of this morphological parser, which was programmed in C, all the other components of our project are being written in Python in order to achieve good results in less time. The output of Pykts is a Document object composed by a hyerarchical data structure of document, sentences, words and morphemes, which is passed to the Event Tagger.

The Event Tagger consists of three modules: a preprocessor where the chunking of Time Expressions is done; a module called Saken, which does the tagging of events; and, a module called Sigan for TIMEX3 tagging. Then, LINK

[11] The architecture mainly relies on that of TTK. However, KTTK introduces a morphological analyzer for morpheme-based standoff annotation. KTTK uses the Aspectual Verb Lexicon for ALINK extraction.

taggers add TLINK, ALINK, SLINK tags. A module S2T changes the annotated SLINKs and ALINKs into TLINKs. In the final step, the LINK Merger merges all TLINKs with temporal closure.

6 Conclusion and Future Work

Temporal and event information extraction is an important step for QA and other inference or temporal reasoning systems. Korean TARSQI Project aims at (1) making KTimeML; (2) building Korean TimeBank as a gold standard, and (3) developing KTTK as an automatic markup tool of temporal and event expressions in Korean text.

In this paper, we presented problems in porting TimeML to Korean and proposed changes of TimeML philosophy. Since consideration of morphological issues is a basic step for KTimeML, we introduce a morpheme-based two-level stand-off annotation scheme. We adopt the surface-based annotation of TimeML, but do not follow the head-only annotation.

The tags of KTimeML are EVENT, TIMEX3, TLINK, ALINK, and SLINKs. The morphological annotation is saved as separate data. The EVENT tag has the attributes such as `vForm, sType, mood`, and `modality` in addition to the attributes of TimeML. We showed the architecture of KTTK.

This work will be a help for QA, IE, and other robust performance for Korean. In addition, KTimeML will be, hopefully, a model for porting TimeML to other agglutinative languages such as Japanese.

Aknowledgements

The authors wish to thank Kiyong Lee (ISO-TimeML) and James Pustejovsky, a director of the TARSQI Project, for their help for the specification of KTimeML.

References

Lionel Clément and Éric Villemonte de la Clergerie. 2005. MAF: a Morphosyntactic Annotation Framework. In *Proceedings of the Language and Technology Conference*, Poznan, Poland, pages 90-94.

Han, Kyoung-Soo, Hoojung Chung, Sang-Bum Kim, Young-In Song, Joo-Young Lee, and Hae-Chang Lim. 2004. TREC 2004 Question Answering System at Korea University. In *Proceedings of the 13rd Text REtrieval Conference*, Pages 446-455. Gettysburg, USA.

Im, Seohyun and Roser Saurí. 2008. TimeML Challenges for Morphological Lanuages: A Korean Case Study. In *Proceedings of CIL* 18, Seoul, Korea.

ISO DIS 24617-1:2008. *Language resources management – Semantic annotation framework (SemAF) – Part1: Time and events. ISO 2008*. Unpublished.

Jang, Seok-Bae, Jennifer Baldwin, and Inderjeet Mani, 2004. Automatic TIMEX2 Tagging of Korean News. In *Proceedings of ACM Transactions on Asian Language Information Processing*. Vol. 3, No. 1, Pages 51-65.

Kim, Jae-Hoon and Seo, Jung-yeon. 1994. ms. *A Korean Part-of-Speech Tag Set for Natural Language Processing* Version 1.0. KAIST. Seoul, Korea.

Kiyong, Lee, 2008. Formal Semantics for Temporal Annotation, An invited plenary lecture for CIL 18. In *Proceedings of the 18th International Congress of Linguists*, CIL 18, Seoul, Korea.

Inderjeet Mani and George Wilson. 2000. Processing of News. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Pages 69-76.

Nam, Ki-Shim and Yong-Kun Ko, 2005. *Korean Grammar (phyojwun kwuke mwunpeplon)*. Top Publisher. Seoul, Korea

Pustejovsky, J., M. Verhagen, X. Nianwen, R. Gaizauskas, M. Happle, F. Shilder, G. Katz, R. Saurí, E. Saquete, T. Caselli, N. Calzolari, K.-Y. Lee, and S.-H. Im. 2008. *TempEval2: Evaluating Events Time Expressions and Temporal Relations: SemEval Task Proposal.*

James Pustejovsky, Jessica Littman, Roser Saurí, Marc Verhagen. 2006. *TimeBank 1.2. Documentation.*

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. IWCS-5. *Fifth International Workshop on Computational Semantics*.

Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006a. *TimeML Annotation Guidelines* Version 1.2.1.

Roser Saurí, Marc Verhagen, and James Pustejovsky. 2006b. SlinkET: A Partial Modal Parser for Events. In *Proceedings of LREC* 2006. Genova, Italy.

Roser Saurí, Robert Knippen, Marc Verhagen and James Pustejovsky. 2005. Evita: A Robust Event Recognizer for QA Systems. In *Proceedings of HLT/EMNLP 2005*, Pages 700-707.

Sohn, Ho-Min. 1999. *The Korean Language*. Cambridge University Press.

Marc Verhagen and James Pustejovsky. 2008. Temporal Processing with the TARSQI Toolkit. In *proceedings Coling 2008: Companion volume - Posters and Demonstrations*, Pages 189-192.