eman ta zabal zazu

universidad
del país vasco

euskal herriko
unibertsitatea

KAF: Kyoto Annotation Framework.

TR 1-2009

Autoreak: Eneko Agirre[1], Xabier Artola[1], Arantza Diaz de Ilarraza[1], German Rigau[1], Aitor Soroa[1], Wauter Bosma’[2]

Afiliazioa: (1) IXA group, University of the Basque Country
(2) Computational Lexicology & Terminology Lab, VU Amsterdam

Konputazio Zientzia eta Adimen Artifiziala Saila
Euskal Herriko Unibertsitatea
Donostia, 2009ko uztaila.

Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco
San Sebastián, julio de 2009.

Txosten Tekniko hau berrikusi dute Unibertsitateko irakasle hauek, zeintzuk adierazi duten txostena argitaratzeak duen interesa:

Este Informe Técnico ha sido revisado por los siguientes profesores de la Universidad, quienes han manifestado el interés de su publicación:

Xabier Arregi Iparragirre   Kepa Sarasola Gabiola

KAF: Kyoto Annotation Framework

Part of deliverable D2.1

Version v0.6

Authors: Eneko Agirre[1], Xabier Artola[1], Arantza Diaz de Ilarraza[1], German Rigau[1], Aitor Soroa[1], Wauter Bosma’[2]

Affiliation: (1) IXA group, University of the Basque Country
(2) Computational Lexicology & Terminology Lab, VU Amsterdam

Knowledge Yielding Ontologies for Transition-based Organization
ICT 211423

Grant Agreement No.	ICT 211423
Project Acronym	KYOTO
Project full title	Knowledge Yielding Ontologies for Transition-based Organization
Funding Scheme	FP7 – ICT
Project website	http://www.kyoto-project.eu/
Project Coordinator	Prof. Dr. Piek T.J.M. Vossen VU University Amsterdam Tel. + 31 (0) 20 5986466 Fax. + 31 (0) 20 5986500 Email: p.vossen@let.vu.nl
Document Number	Part of deliverable D2.1
Status & version	v0.6
Contractual date of delivery	–
Actual date of delivery	July 29, 2009
Type	Working document
Security (distribution level)	Public
Number of pages	19
WP contributing to the deliberable	WP2
WP responsible	German Rigau
EC Project Officer	Werner Janusch

Authors: Eneko Agirre[1], Xabier Artola[1], Arantza Diaz de Ilarraza[1], German Rigau[1], Aitor Soroa[1], Wauter Bosma’[2]

Keywords: Annotation standards

Abstract: This document presents the current draft of KAF: Kyoto Annotation Framework to be used within the KYOTO project. KAF aims to provide a reference format for the representation of semantic annotations.

Contents

List of Tables

1 Introduction

This document presents the first draft of KAF: Kyoto Annotation Framework to be used within the KYOTO project. KAF aims to provide a reference format for the representation of semantic annotations.

There have been numerous attempts to standardize some aspect of natural language processing. To date, the focus of standards (in various stages of development) includes morphosyntactic annotation (MAF) [Clément and Villemonte de La Clergerie, 2005], syntactic annotation (SynAF) [Declerck, 2006], and semantic annotation (e.g. SemAF[Lee *et al.*, 2007]). The beforementioned standards concentrate on a specific stage of annotation. The two meta-models present different degrees of maturity; MAF has entered the last stages of the ISO process, whereas SynAF is at the level of Working Draft standard.

A problem for these formats is that they are difficult to combine. For instance, we might want to do both syntactic annotation and semantic annotation, and integrate the results. The Linguistic Annotation Framework (LAF) [Ide and Romary, 2003] is an ISO standard proposal of a data model for linguistic annotation. It allows individual annotations within the annotation framework to refer to each other, so that the result is a combined analysis of the source text.

Rather than a data model, our aim is a layered annotation format, where several processes can add information without losing anything which is produced by any previous process. KAF provides annotation layers for basic natural language processing and is open to extensions with other annotation layers needed by specific applications, which may be standardized later on. KAF is compatible with LAF but imposes a more specific standardization of the annotation format itself.

Given KYOTO strong vocation towards an open and public system, all data formats have been inspired by standard specifications available in the field of Language Resources. Basic motivations for that were to ensure intra- and inter-operability and portability. MAF and SynAF were investigated as far linguistic annotation for morpho-syntactic and syntactic information, respectively, is concerned.

KAF can be seen as a three-layer format for text annotation: the first two layers, explicitly dedicated to representing morphosyntactic and syntactic information, are inspired by MAF and SynAF and are implemented “over” the semantic layer. For semantic annotation, the ISO community provides SemAF which is especially dedicated to the representation of events and time. We decided to boost semantic annotation and devised a *dialect* of the ISO standards, where semantic notation is tailored to the specific purposes of the project. KAF layers are to be seen as dialects of the ISO standards, yet maintaining (different degrees of) mappability to them. The KYOTO dialects do not corrupt the compliance with ISO standards and their underlying philosophy; instead, they are in line with the strategy in ISO which provides high-level models (meta-models) able to be adapted, tailored and implemented according to specific needs.

KAF comprises several annotations over a text at different linguistic levels (morphosyntactic, syntactic, semantic) and adopts a stand off strategy for annotating the source text. The following overall rules are followed in all layers:

- `<span>` elements are used for grouping linguistic elements.
- Linguistic annotations of a particular level always spans elements of previous levels.
- Linguistic annotations of different levels are not mixed.

In this document we will describe the annotation levels in turn, using the sentence as a running example:

John taught mathematics 20 minutes every Monday in New York.

2 Root element

All KAF documents have a root element `<KAF>` which has the following attributes:

- `xml:lang`: language identifier[1].

Ex:

```
<KAF xml:lang="en">
<!--- ... --->
</KAF>
```

3 kafHeader

KAF documents may have a header for describing information about the document, such as its original name, URI or a list of the linguistic processors which generated the KAF document. The KAF header is represented within the `<kafHeader>` element, which is optional but highly recommended. The header element has three sub-elements:

3.0.1 fileDesc element

`<fileDesc>` is an empty element containing information about the computer document itself. It has the following attributes:

- `title`: the title of the document (optional).
- `author`: the author of the document (optional).
- `filename`: the original file name (optional).
- `filetype`: the original format (PDF, HTML, DOC, etc) (optional).
- `pages`: number of pages of the original document (optional).

Example:

```
<fileDesc title="3_3012" author="WWF" filename="KYOTO_3_3012"
          filetype="PDF" pages="19"/>
```

[1] as described in `http://www.w3.org/International/articles/language-tags/`

3.0.2 public element

`<public>` is an empty element which stores public information about the document, such as its URI. It has the following attributes:

- `publicId`: a public identifier (for instance, the number inserted by the capture server) (optional).
- `uri`: a public URI of the document (optional).

Example:

```
<public publicId="3_3012" uri="http://kyoto.org/docs/KYOTO_3_3012.pdf" />
```

3.0.3 Linguistic Processors

The header also stores the information about which linguistic processors produced the KAF document, described under `<linguisticProcessors>` elements. There can be several `<linguisticProcessors>` elements, one per KAF layer. KAF layers correspond to the top-level elements of the documents, such as "text", "terms", "deps" etc.

Each `<linguisticProcessors>` element contains one or several `<lp>` elements, each one describing one specific linguistic processor. `<lp>` elements have the following attributes:

- `name`: the name of the processor
- `version`: processor's version
- `timestamp`: a timestamp, denoting the date/time at which the processor was launched. The timestamp follows the XML Schema *xs:dateTime* type[2]. In summary, the date is specified following the form "YYYY-MM-DDThh:mm:ss" (all fields required). To specify a time zone, you can either enter a dateTime in UTC time by adding a "Z" behind the time ("2002-05-30T09:00:00Z") or you can specify an offset from the UTC time by adding a positive or negative time behind the time ("2002-05-30T09:00:00+06:00").

Example:

```
<linguisticProcessors layer="text">
  <lp name="Freeling" version="2.1" timestamp="2009-06-25T10:05:00Z"/>
</linguisticProcessors>
<linguisticProcessors layer="terms">
  <lp name="Freeling" version="2.1" timestamp="2009-06-25T10:10:19Z"/>
  <lp name="ukb" version="0.1.2" timestamp="2009-06-25T16:10:19Z"/>
</linguisticProcessors>
<linguisticProcessors layer="namedEntities">
  <lp name="kybot_NE" version="0.1" timestamp="20090626_00:10:19Z"/>
</linguisticProcessors>
```

[2] See http://www.w3.org/TR/xmlschema-2/#isoformats

Full example of a KAF header:

```
<kafHeader>
  <fileDesc title="3_3012" author="WWF" filename="KYOTO_3_3012"
            filetype="PDF" pages="19"/>
  <public publicId="3_3012" uri="http://kyoto.org/docs/KYOTO_3_3012.pdf" />
  <linguisticProcessors layer="text">
    <lp name="Freeling" version="2.1" timestamp="2009-06-25T10:05:00Z"/>
  </linguisticProcessors>
  <linguisticProcessors layer="terms">
    <lp name="Freeling" version="2.1" timestamp="2009-06-25T10:10:19Z"/>
    <lp name="ukb" version="0.1.2" timestamp="2009-06-25T16:10:19Z"/>
  </linguisticProcessors>
  <linguisticProcessors layer="namedEntities">
    <lp name="kybot_NE" version="0.1" timestamp="2009-06-26T00:10:19Z"/>
  </linguisticProcessors>
</kafHeader>
```

4 Word forms

After tokenization step, all word forms are annotated within the `<text>` element, and each form is enclosed by a `<wf>` element.

`<wf>` elements have the following attributes:

- `wid`: the unique id for the word form.
- `sent`: sentence id of the token (optional)
- `para`: paragraph id (optional)
- `page`: page id (optional)
- `xpath`: in case of source xml files, the xpath expression identifying the token (optional)

Example of word level annotations:

```
<text>
  <wf wid="w1" sent="s1" para="p1">John</wf>
  <wf wid="w2" sent="s1" para="p1">taught</wf>
  <wf wid="w3" sent="s1" para="p1">mathematics</wf>
  <wf wid="w4" sent="s1" para="p1">20</wf>
  <wf wid="w5" sent="s1" para="p1">minutes</wf>
  <wf wid="w6" sent="s1" para="p1">every</wf>
  <wf wid="w7" sent="s1" para="p1">Monday</wf>
  <wf wid="w8" sent="s1" para="p1">in</wf>
  <wf wid="w9" sent="s1" para="p1">New</wf>
  <wf wid="w10" sent="s1" para="p1">York</wf>
  <wf wid="w11" sent="s1" para="p1">.</wf>
</text>
```

5 Terms

Terms refer to previous word forms (and groups multi word forms) and attach lemma, part of speech, synset and name entity information.

`<term>` elements have the following attributes:

- `tid`: unique identifier
- `type`: type of the term. Currently, 3 values are possible:
  - `open`: open category term
  - `close`: close category term
  - `entity`: term is a named entity
- `lemma`: lemma of the term
- `pos`: part of speech. The first letter of the pos attribute must be one of the following:

N	common noun
R	proper noun
G	adjective
V	verb
P	preposition
A	adverb
C	conjunction
D	determiner
O	other

  more complex pos attributes may be formed by concatenating values separated by a dot ".". For example, in Basque we have `"V.ADI.SIN"` for simple verbs or `"V.ADI.KON"` for complex verbs.
- `netype`: if the term is a named entity, the type of the entity (only if `type="entity"`).
- `case`: declension case of the term (optional).
- `head`: if the term is a compound, the id of the head component (see Section 5.2).

5.1 External References

The `<externalReferences>` element is used to associate terms to external resources, such as elements of a Knowledge base, an ontology, etc. It consists of several `<externalRef>` elements, one per association.

`<externalRef>` elements have the following attributes:

- `resource`: indicates the identifier of the resource referred to.
- `reference`: code of the referred element. If the element is a synset of some version of WordNet, it follows the pattern:

  `[a-z]3-[0-9]2-[0-9]+-[nvars]`

  which is a string composed by four fields separated by a dash. The four fields are the following:

  - Language code (three characters).
  - WordNet version (two digits).
  - Synset identifier composed by digits.
  - POS character:

    n noun
    v verb
    a adjective
    r adverb

  examples of valid patterns are: “ENG-20-12345678-n”, “SPA-16-017403-v”, etc.
- `confidence`: a floating number between 0 and 1. Indicates the confidence weight of the association.

Example of term level annotations:

```
<terms>
  <term tid="t1" type="entity" lemma="John" pos="R" netype="person">
    <span>
        <target id="w1"/>
    </span>
  </term>
  <term tid="t2" type="open"  lemma="teach" pos="V">
    <span>
        <target id="w2"/>
    </span>
    <externalReferences>
```

```
      <externalRef resource="WN-1.7" reference="ENG-17-00861095-v" confidence="0.80"/>
      <externalRef resource="WN-1.7" reference="ENG-17-00859568-v" confidence="0.20"/>
    </externalReferences>
  </term>
  <term tid="t3" type="open"  lemma="mathematics" pos="N">
    <span>
        <target id="w3"/>
    </span>
    <externalReferences>
      <externalRef resource="WN-1.7" reference="ENG-17-04597590-n" confidence="1.0"/>
    </externalReferences>
  </term>
  <term tid="t4" type="entity" lemma="20" pos="N" netype="number">
    <span>
        <target id="w4"/>
    </span>
  </term>
  <term tid="t5" type="open" lemma="minute" pos="N">
    <span>
        <target id="w5"/>
    </span>
  </term>
  <externalReferences>
    <externalRef resource="WN-1.7" reference="ENG-17-12621100-n" confidence="0.80"/>
    <externalRef resource="WN-1.7" reference="ENG-17-12631889-n" confidence="0.06"/>
    <externalRef resource="WN-1.7" reference="ENG-17-12630443-n" confidence="0.01"/>
    <externalRef resource="WN-1.7" reference="ENG-17-11241911-n" confidence="0.01"/>
    <externalRef resource="WN-1.7" reference="ENG-17-05339359-n" confidence="0.01"/>
    <externalRef resource="WN-1.7" reference="ENG-17-04316149-n" confidence="0.01"/>
  </externalReferences>

  <term tid="t5" type="close" lemma="every" pos="D">
    <span>
        <target id="w6"/>
    </span>
  </term>

  <term tid="t6" type="entity" lemma="Monday" pos="N" netype="date">
    <span>
        <target id="w7"/>
    </span>
    <externalReferences>
      <externalRef resource="WN-1.7" reference="ENG-17-12557842-n" confidence="1.0"/>
    </externalReferences>
  </term>
  <term tid="t7" type="close" lemma="in" pos="P">
    <span>
        <target id="w8"/>
    </span>
  </term>
  <term tid="t8" type="entity" lemma="New_York" pos="R" netype="location">
    <span>
        <target id="w9"/>
        <target id="w10"/>
    </span>
  </term>
</terms>
```

5.2 Compound terms

Compound terms can be represented in KAF by including `<component>` elements within `<term>` elements. For example, the term *landbouwbeleid* (English: agriculture policy) would look like this:

```
<term tid="t7" head="t7.1" lemma="landbouwbeleid" pos="N" type="open">
  <span><target id="w7"/></span>
  <component id="t7.1" lemma="landbouw" pos="N">
    <externalReferences>...</externalReferences>
  </component>
  <component id="t7.2" lemma="beleid" pos="N">
    <externalReferences>...</externalReferences>
  </component>
  <externalReferences>...</externalReferences>
</term>
```

`<component>` elements have the following attributes:

- `id`: unique identifier
- `lemma`: lemma of the term
- `pos`: part of speech
- `case`: declension case

6 Dependency relations

Dependencies represent dependency relations among terms. Each dependency is represented by an empty `<dep>` element and span previous terms.

`<dep>` element have the following attributes:

- `from`: term id of the source element

- `to`: term id of the target element
- `rfunc`: relational function. One of:
  - `mod`: indicates the word introducing the dependent in a head- modifier relation. For instance:

`mod(by,gift,Peter)`	the gift of a book by Peter
`mod(of,examination,patient)`	the examination of the patient

  - `subj`: indicates the subject in the grammatical relation Subject-Predicate. For instance:

`subj(arrive,John,_)`	John arrived in Paris
`subj(employ,Microsoft,_)`	Microsoft employed 10 C programmers
`subj(employ,Paul,obj)`	Paul was employed by Microsoft

  - `csubj, xsubj, ncsubj`: The Grammatical Relations (RL) s csubj and xsubj may be used for clausal subjects, controlled from within, or without, respectively. ncsubj is a non-clausal subject. For instance:

`xsubj(win,require,_)`	to win the America's Cup requires heaps of cash

  - `dobj`: Indicates the object in the grammatical relation between a predicate and its direct object. For instance:

`dobj(read,book,_)`	read books

  - `iobj`: The relation between a predicate and a non-clausal complement introduced by a preposition; type indicates the preposition introducing the dependent. For instance:

`iobj(in,arrive,Spain)`	arrive in Spain
`iobj(into,put,box)`	put the tools into the box
`iobj(to,give,poor)`	give to the poor

  - `obj2`: The relation between a predicate and the second non-clausal complement in ditransitive constructions. For instance:

`obj2(head,dependent)`	
`obj2(give,present)`	give Mary a present
`obj2(mail,contract)`	mail Paul the contract

- `case` (optional): declension case

Example of dependency relation annotations:

```
<deps>
  <!-- subj(teach, John) -->
  <dep from="t1" to="t2" rfunc="subj" />
  <!-- dobj(teach, Mathematics) -->
  <dep from="t3" to="t2" rfunc="dobj" />
  <!-- iobj(teach, New_York) -->
  <dep from="t8" to="t2" rfunc="iobj" />
</deps>
```

7 Chunks

Chunks are noun or prepositional phrases, spanning terms.

`<chunk>` elements have the following attributes:

- `cid`: unique identifier
- `head`: the chunk head's term id
- `phrase`: typo of the phrase
- `case` (optional): declension case

Example of chunk annotations:

```
<chunks>
  <!-- John -->
  <chunk cid="c1" head="t1" phrase="NP">
    <span>
<target id="t1"/>
    </span>
  </chunk>
  <!-- taught -->
  <chunk cid="c2" head="t2" phrase="V">
    <span>
<target id="t2"/>
    </span>
  </chunk>
  <!-- Mathematics -->
  <chunk cid="c3" head="t3" phrase="NP">
    <span>
<target id="t3"/>
    </span>
  </chunk>
  <!-- 20 minutes -->
  <chunk cid="c5" head="t5" phrase="NP">
    <span>
<target id="t4"/>
<target id="t5"/>
    </span>
  </chunk>
  <!-- every -->
  <chunk cid="c6" head="t6" phrase="R">
    <span>
<target id="t6"/>
    </span>
  </chunk>
  <!-- every Monday -->
  <chunk cid="c7" head="t7" phrase="NP">
    <span>
<target id="t6"/>
<target id="t7"/>
    </span>
  </chunk>
  <!-- in New York -->
  <chunk cid="c9" head="t9" phrase="PP">
    <span>
<target id="t8"/>
<target id="t9"/>
    </span>
  </chunk>
</chunks>
```

8 Events

Events provide event information, including roles, spanning chunks. The specific semantics of `<event>` elements is defined in [Lee *et al.*, 2007].

`<events>` elements have the following attributes:

- `eid`: unique identifiers
- `span`: chunk id of the main event
- `lemma`: lemma of the event
- `pos`: part of speech (see pos attribute of `<term>` elements. The same rules apply here.)
- `eiid`:
- `class`: event class
- `tense`:
- `aspect`:
- `polarity`:

Example of event annotation:

```
<events>
  <event eid="e1" span="c2" lemma="teach" pos="V" eiid="ei1" class="OCCURRENCE"
       tense="PAST" aspect="NONE" polarity="POS">
    <roles>
<role cid="c1" role="agent"/>
<role cid="c2" role="subject"/>
<role cid="c3" role="location"/>
```

```
      </roles>
  </event>
</events>
```

9 Quantifiers

Quantifiers are annotated within `<quantifiers>` element. Normally, they are further used for specifying relations. The specific semantics of `<quantifier>` elements is defined in [Lee *et al.*, 2007], the sole difference being that on KAF quantifiers refer to chunks.

`<quantifier>` elements have the following attributes:

- `qid`: unique identifier
- `span`: chunk id of quantifier

Example of quantifier annotation:

```
<!-- every -->
<quantifiers>
  <quantifier qid="q1" span="c6"/>
</quantifiers>
```

10 Time expressions (timex)

Time expressions are annotated within `<timexs>` element. The specific semantics of `<timex>` elements are defined in [Lee *et al.*, 2007], the sole difference being that on KAF quantifiers refer to chunks.

Example of time expression annotations:

```
<!-- 20 minutes every Monday -->
<timexs>
  <timex3 texid="tex1" type="DURATION" value="P20TM">
    <span>
<target id="c5"/>
    </span>
  </timex3>

  <timex3 texid="tex2" type="SET" value="xxxx-wxx-1" quant="EVERY">
    <span>
<target id="c7"/>
    </span>
  </timex3>

  <tlink timeID="tex1" relatedToTime="tex2" relType="IS_INCLUDED"/>
  <tlink eventInstanceID="ei1" relatedToTime="tex1" relType="SIMULTANEOUS"/>
</timexs>
```

References

[Clément and Villemonte de La Clergerie, 2005] Lionel Clément and Éric Villemonte de La Clergerie. Maf: a morphosyntactic annotation framework. In *Proceedings of the 2nd Language & Technology Conference*, page 90–94, April 2005.

[Declerck, 2006] Thierry Declerck. Synaf: Towards a standard for syntactic annotation. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth Conference on International Language Resources and Evaluation*, pages 229–233. European Language Resources Association (ELRA), May 2006.

[Ide and Romary, 2003] Nancy Ide and Laurent Romary. Outline of the international standard linguistic annotation framework. In *Proceedings of ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*, pages 1–5. Association for Computational Linguistics, July 2003.

[Lee *et al.*, 2007] K. Lee, J. Pustejovsky, H. Bunt, B. Boguraev, and N. Ide. *Language resource management - Semantic annotation framework (SemAF) - Part 1 :Time and events.* International Organization for Standardization, Geneva, Switzerland, 2007. http://lirics.loria.fr/doc_pub/SemAFCD24617-1Rev12.pdf.