A multi-layer markup language for geospatial semantic annotations
Ludovic Moncla, Mauro Gaio

To cite this version:
hal-01256370

HAL Id: hal-01256370
https://hal.archives-ouvertes.fr/hal-01256370
Submitted on 1 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A Multi-Layer Markup Language for Geospatial Semantic Annotations

Ludovic Moncla
Universidad de Zaragoza
C/ Maria de Luna, 1 - Zaragoza, Spain
moncla.ludovic@gmail.com

Mauro Gaio
LIUPPA - EA 3000
Av. de l’Université - Pau, France
mauro.gaio@univ-pau.fr

ABSTRACT

In this paper we describe a markup language for semantically annotating raw texts. We define a formal representation of text documents written in natural language that can be applied for the task of Named Entities Recognition and Spatial Role Labeling.

The proposal relies on a multi-layer annotation process based on a core generic layer, which can be freely adapted into more specific layers depending on the intended goal. Our markup language is based on the TEI Guidelines\(^1\) to propose a generic and extensible markup language. This language is particularly dedicated for the text mining task and ready to use to be layered with more semantic relationships between elements of the text.

We show the feasibility of this proposal from a generic annotation of texts describing itineraries toward a geospatial semantic annotation.

CCS Concepts

- Information systems \(
\rightarrow\)
  - Data encoding and canonicalization; Language models; Information extraction;
- Applied computing \(
\rightarrow\)
  - Extensible Markup Language (XML);

Keywords

geo-semantic tagging; text annotation; expanded named entity

\(^1\)Laboratoire COGIT
73 av. de Paris - 94160 Saint-Mandé, France

\(^1\)LIUPPA - EA 3000
Av. de l’Université - Pau, France

TEI is an international standard for textual markup defined by the TEI Consortium. It provides a guide to best practices for interchange and encoding of textual material in digital form. http://www.tei-c.org/Guidelines/P5/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GIR ’15, November 26-27, 2015, Paris, France
© 2015 ACM. ISBN 978-1-4503-3937-7/15/11. . . $15.00
DOI: http://dx.doi.org/10.1145/2837689.2837700

1 Motivation and Background

Two categories of markup languages can be considered in geospatial tasks, those dedicated to the encoding of spatial data (e.g., GML, KML) and those dedicated to the annotation of spatial or spatio-temporal information in texts, such as in semantic role parsing [4, 10]. SpatialML [6] is more focused on the annotation of static spatial information and does not provide any support to identify spatio-temporal information such as motion. ISO-Space [12] annotates spatio-temporal information and has been also designed for capturing implicit spatial information. Although ISO-Space seems a comprehensive standard for capturing spatio-temporal information, some of its elements remain complex and are not really suited for a fully automatic process. Otherwise TEI is a standard for textual markup and provides a guide to best practices for interchange and encoding textual information. TEI has been designed to be the more generic and adaptable as possible and provides a generic framework particularly adapted for the customization.

The existing markup languages dedicated to the annotation are more focused on the specification of relations (spatial or spatio-temporal) than in named entities (NE). Standard markup languages consider NE as being only composed of a pure proper name, whereas we consider more complex expressions known as expanded named entities (ENE) [7, 15].

ENE concept is particularly essential in any task that aims to retrieve all or part of evocative context of the NE, such as disambiguation task. For those reasons, we propose the specification of a multi-layer markup language adapted to the annotation of both ENEs and spatial/spatio-temporal relations. The proposed TEI-compliant markup language is based on a core generic layer, dedicated to the annotation of NE, which can be used to share pre-processed corpus of documents. Finally the language is designed to be compliant with an automatic process of annotation.

Although still at an early stage of development, the proposed markup language was applied for a problem of automatic information extraction and toponym resolution described in [8] and for a problem of automatic itinerary reconstruction described in [7].

The remainder of the paper is structured as follows. Section 2 describes the ‘core generic layer’ of our multi-layer annotation language. Then, Section 3 describes the adaptation of the generic language for a specific semantic role. Finally, Section 4 summarises and concludes this paper.
2 A Generic Markup Language for Expanded Named Entity Representation

The main elements of our proposed generic markup specification are categorized in two groups, the first one refers to the elements contained within expanded named entities, the second one refers to the standard elements describing the text segmentation and a set of global attributes defined by the TEI P5 Guidelines which are available for all TEI elements.

2.1 Expanded Named Entity Representation

2.1.1 Principles

According to [5] there are two categories of proper names: pure and descriptive. Pure proper names can be simple (i.e., composed of a single lexeme) or complex (i.e., composed of several lexemes) and are only composed of proper names. Descriptive proper names refer to a composition of proper names and common names (i.e., expansion). In other words, descriptive proper names overlap pure proper names. Descriptive proper names refer to a NE built with a pure proper name and a descriptive expansion. This expansion can change the type (e.g., location, person, etc) of the initial pure proper name.

An ENE may contain an entity built with both categories of proper names (i.e., pure and descriptive), and that can be composed of one or more concepts, whereas most of works of NER are usually only considering pure proper names. This provides us, in particular, to move beyond reduction of place to a name and a set of coordinates, a model still predominant in Geographic Information Science as specified by [11]. We define several levels of overlapping, (0, 1, 2, etc.) for the representation of ENE. Each level is encapsulated in the previous level.

Level 0. refer to pure proper names. It can be seen as the core component of an ENE. Thus, we consider NE as a special kind of ENE. The following examples (1 - 3) show some examples of entity of level 0:

(1) Cardiff → one entity (location)
(2) Balaruc-le-Vieux → one entity (location)
(3) Charles de Gaulle → one entity (person)

Level 1. refer to descriptive proper names composed of a pure proper name (i.e., an entity of level 0) and a common noun (i.e., expansion). The following examples (4 - 6) show the representation of ENE. We can notice that in these cases, descriptive expansions do not change the intrinsic nature of the object described by the proper name.

(4) commune de Balaruc-le-Vieux
(5) comunidad autónoma de Aragón
(6) général Charles de Gaulle

However, when the associated term is not equal to the intrinsic or default type of the pure proper name, it defines a new entity that overlaps the pure proper name. The following examples (7 - 10) illustrate that an entity may contain the name of another entity, and that the new entity may have a different type. For instance, a sentence containing the expanded named entity “sindaco di Venezia” (sindaco = mayor) is referring to the person and not necessarily to the city.

(7) Cardiff Downtown → two entities, Cardiff (location) and Cardiff Downtown (location)
(8) deltà del Ebro → two entities, Ebro (location) and deltà del Ebro (location)
(9) château de Versailles → two entities, Versailles (location) and château de Versailles (location)
(10) sindaco di Venezia → two entities, Venezia (location) and sindaco di Venezia (person)

Level 2. refer to a descriptive proper name composed of another descriptive proper name. ENE of Level 2 are built with ENE of level 1 and with a descriptive expansion. The following examples (11 - 13) show some ENE of Level 2. The behaviour is the same as for the previous level, the expansion can change the nature of the object described by the ENE of level 1. For instance, in example (13) the person ENE ‘général Charles de Gaulle’ becomes a reference to a location with the expansion ‘avenue’ having a geographical sense.

(11) Parque Natural del Delta del Ebro
(12) Cardiff City Football Club
(13) avenue du général Charles de Gaulle

Level 3. ENE at this level are built with ENE of level 2 plus a descriptive expansion. Actually, there is not really a limit to the overlapping. However, it is really rare to find an ENE of level 3 or more. The following examples (14 - 15) show some ENE of Level 3.

(14) la valle glaciale del lago di monte Acuto
(15) the owner of the restaurant of Neuvić Lake
The proposed hierarchy of overlapping of ENE introduces more detailed entities and allows the annotation of more fine-grain NE and less errors of classification.

In our language each level of the ENE can be marked, from the pure proper name to the whole ENE. For instance, considering example (15), using NER method described in [8] produces the following results:

- ‘Neuvic’ as proper name (ENE level 0)
- ‘Neuvic Lake’ as geographical name (ENE level 1)
- ‘restaurant of Neuvic Lake’ as place name (ENE level 2)
- ‘owner of the restaurant of Neuvic Lake’ as person (ENE level 3)

To illustrate the advantage of using the introduced concept of ENE, we have tested four well-known English NER tools (Stanford, Open Calais, Illinois and FreeLing) with example (15). The results are shown below:

- Stanford Named Entity Recognizer\(^2\) annotates one entity:
  - ‘Neuvic Lake’ as location
- Open Calais\(^3\) annotates two NEs:
  - ‘Neuvic Lake’ as natural feature
  - ‘restaurant of Neuvic Lake’ as facility
- NER tool of the Cognitive Computation Group\(^4\) [13] of the University of Illinois annotates only one NE:
  - ‘Neuvic Lake’ as organization
- FreeLing\(^5\) annotates one NE:
  - ‘Neuvic Lake’ as geographical location

We argue that for a powerful task, of marking and of categorizing named entities in context, it is essential to consider ENE (15) as a whole entity. Whereas standard NER tools, considering only the entity ‘Neuvic Lake’, lead to inaccuracies in categorization and this is also true for OpenCalais in despite of taking into account the word ‘restaurant’. As it can be observed in the example (15) the bounding ENE refers to a person (the owner of the restaurant) and not a spatial entity while it is yet the case for all constituent entities. Furthermore, the problem of wrong classification due to bad boundaries detection occurs also for smaller named entities such as for example: the mayor of Paris. Among the four NER tools previously tested (i.e., Stanford, Open Calais, Illinois and FreeLing), only one (Open Calais) succeeds to detect the ENE as a person. All the other, consider only the named entity ‘Paris’ as a location.

2.1.2 TEI Based Annotation

According to our proposal of a TEI compliant XML markup language, we will now describe the TEI elements and their attributes used to represent ENE.

The <name> element provided by TEI annotates proper names or noun phrases which are equivalent to ENE of level 0. Table 1 shows the current set of <name> attributes defined in our customized specification.

In our specification, all the attributes of the <name> element are optional and the type attribute may refer to the

Table 1: Attributes for <name> tag

<table>
<thead>
<tr>
<th>type</th>
<th>attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>category</td>
<td>category of NE (location, person, organization, etc.)</td>
</tr>
<tr>
<td>subType</td>
<td>semantic sub-categorization</td>
</tr>
</tbody>
</table>

\(^1\)http://nlp.stanford.edu/software/CRF-NER.shtml
\(^2\)http://nlp.stanford.edu/software/CRF-NER.shtml
\(^3\)http://new.opencalais.com/
\(^4\)http://cogcomp.cs.illinois.edu/page/demo/view/NER
\(^5\)http://nlp.lsi.upc.edu/freeling/

category of the NE (location, person, organisation, etc.) such as the ENAMEX types proposed in the MUC-6 typology or such as those defined by [15, 16]. The subType attribute for <name> indicates a second level in the classification such as roadName if the type attribute value is equal to ‘location’. Figure 1 illustrates the annotation of pure proper names (i.e., ENE of level 0) using the TEI <name> element.

Figure 1: Examples of Named Entities (NE)

In our annotation scheme, we also take advantage of tags provided by TEI for dates and time. These tags are described in the Core module of the TEI Guidelines and are available for all TEI documents. They refer to the expressions of date, time or duration in texts. Figure 2 shows some examples of annotation of date and time entities.

Figure 2: Annotation of date and time

The <term> element, defined by the TEI Guidelines, contains a single-word, multi-word, or symbolic designation which is regarded as a technical term. In the current generic layer of the language, we use the <term> element for several purposes such as annotating common nouns which refer to the descriptive expansion part of ENR (i.e., common nouns associated with a proper name). Table 2 shows the current set of <term> attributes defined in our customized specification.

Table 2: Attributes for <term> tag

<table>
<thead>
<tr>
<th>type</th>
<th>attribute value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N, offset</td>
<td>N, offset, measure</td>
</tr>
<tr>
<td>measure</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 shows the current set of <term> attributes defined in our customized specification and Figure 3 shows some examples of annotation of <term> elements.

In our proposed customization of TEI, the type attribute for the <term> element is mandatory and its value must be: N, offset or measure. The N value means that the <term> element contains a common noun and may refer to the descriptive expansion part of an ENR. Furthermore, in the current version of the language, the offset type refers to the expression of spatial or temporal relations. For instance, in the case of an offset type, the value of the subType attribute
<term xml:id="t1" type="N">refuge</term>
</term>

---
<term xml:id="t2" type="offset" subtype="orientation">
<term lemma="au" type="PREPDETP">au</term>
</term>

---
<term xml:id="t3" type="offset" subtype="direction:final">
<term lemma="jusqu'" type="PREP">jusqu'</term>
</term>

---
<term xml:id="t4" xml:lang="en" type="measure">
<term type="NUM">200</term>
</term>

---
<term xml:id="rs" type="expandedName">
<name type="expandedName">refuge</name>
</term>

Figure 3: <term> elements with different types

may be: orientation, adjacency, inclusion, direction initial or direction final, depending on the nature of the spatial relations. Finally, the measure type annotates distance measures.

As we have seen in the definition of the concept of ENE, there are several levels of expansion. Each level can encapsulate the ENE of lower level. According to the TEI Guidelines, the <rs> element defined in the Core module, contains a general purpose name or referring string. In the generic layer specification of our multi-layer markup language, we use the <rs> element for annotating ENE. Furthermore, in our customized specification, a <rs> element is either composed of a <term> element and another <rs> element; or it consists of a <term> element and a <name> element. <rs> elements interpret ENE in a broad manner and can encapsulate all types of <term> elements.

<table>
<thead>
<tr>
<th>type</th>
<th>expandedName, relative, sequence</th>
</tr>
</thead>
</table>

We specify the <rs> element as having only one attribute: type (Table 3). We define the attribute type as optional and its value must be equal to: expandedName, relative, or sequence. <rs> elements having an expandedName type refer to the expression of ENE (see <rs xml:id="rs1"> in Figure 4). They must be composed of a <name> element or another <rs> element and may contain a <term> element with a type value equal to N. Furthermore, we use the global attribute n to specify the level of encapsulation (e.g., 0, 1, 2, etc.).

The relative type refers to the expression of a proper name (i.e., <name>), or of an ENE of level > 0 (i.e., <rs type="expandedName">), associated with a modifier, i.e., a <term> element having an offset value for the type attribute (see <rs xml:id="rs2"> in Figure 4). Finally, the sequence value for the type attribute refers to a sequence of several <rs> elements (Figure 5). A <term> element contained by a <rs type="sequence"> is applied to all the <rs> elements. For instance, the term ‘les bourgs’ (i.e., small villages) is associated with both Barioz and Bieux entities (which are the names of small villages).

2.2 Text segmentation

This section describes the elements that refer to the text segmentation. Structuring of textual information operates on various levels of discourse. They describe the segmentation of a text into traditional linguistic categories such as sentences, words and punctuation marks. These elements are defined by the TEI Guidelines and belong to the Analysis module (Simple analytic mechanisms) of TEI.

(16) Étes-vous parvenu au refuge du Col de la Vanoise?
Did you reach the refuge of Col de la Vanoise?

The <s> element contains a sentence-like division of a text. It may be used to annotate sentences, or any other non-overlapping segments such as complete and non-nesting segmentation of a text. Figure 6 shows the result of the annotation of sentence (16).

Figure 4: <rs> element: the case of relative type

Figure 5: <rs> element: the case of sequence type

Figure 6: Annotation of a sentence

The <s> element represents a grammatical word. The word segmentation depends on which characters are defined as words dividers. Whereas for most of Indo-European languages the word separator is typically a blank space, the
East Asian languages such as Chinese and Japanese differ, i.e., words are not explicitly delimited. In our work we are mainly focused on Indo-European languages.

Table 4: Attributes for <w> tag

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>lemma</td>
<td>Canonical form of the word</td>
</tr>
<tr>
<td>type</td>
<td>Part-of-speech</td>
</tr>
<tr>
<td>subtype</td>
<td>Semantic sub-categorization</td>
</tr>
</tbody>
</table>

Table 4 shows the current set of <w> attributes defined in our specification and Figure 7 shows the result of the annotation of sentence (16).

Figure 7: Annotation of words, grammatical phrases and punctuation

The lemma attribute for <w> is optional and may contain the canonical form of the word. The type attribute for <w> is mandatory and contains the lexical categories of word (part-of-speech). Although there are significant variations depending on the language, common parts of speech are: noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, numeral, article or determiner. The label for each category is usually defined by abbreviations such as N, V, ADJ, ADV, etc. Different POS tags have been defined and used in well-known projects such as the Brown Corpus, the Penn Treebank and the French Tree Bank. The subtype attribute for <w> is optional and may contain semantic information. For instance, in the current version of the language, subtype is used to classify verbs. The possible values are: motion initial, motion median, motion final, perception and topographic.

The <phr> element defined by the TEI Guidelines in the analysis module represents a grammatical phrase. The type attribute may be used to indicate the type of phrase such as noun phrases, prepositional phrases, verbal phrases, etc. Figure 7 shows the annotation of sentence (16).

The <pc> element contains a character or string of characters regarded as constituting a single punctuation mark. Table 5 shows the current set of <pc> attributes.

All the attributes of the <pc> element are optional. The force attribute for <pc> indicates if the considered punctuation mark is a separator for words or phrases. The type attribute for <pc> indicates the kind of punctuation. Figure 7 illustrates the annotation of punctuation.

Table 5: Attributes for <pc> tag

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>force</td>
<td>strong, weak, inter</td>
</tr>
<tr>
<td>type</td>
<td>declarative, imperative, interrogative, exclamatory</td>
</tr>
</tbody>
</table>

All these elements used for the text segmentation and the representation of ENNE define the generic layer of the multi-layer markup language. They have been applied for the text mining and NER tasks.

3 Towards a Geospatial Semantic ML

3.1 Overview

This section describes the adaptation to transform the generic core layer towards a geospatial semantic markup language: the second layer.

We propose some guidelines for a TEI compliant markup language for encoding spatial information. Some elements belonging to the generic layer of our multi-layer markup language (<term> and <rs>) are turned into more specific elements embedding geospatial semantics.

According to the Namesdates module of the TEI Guidelines, the content of <geogFeat> elements is defined as a common noun identifying some geographical feature (e.g., valley, mount, etc.) contained within a spatial NE. This is the equivalent of the definition of a descriptive expansion part of an ENNE (i.e., <term type="N"> element) having a geographical denotation and associated with a spatial NE. Thus, in our customized specification, the <term type="N"> elements having a geographical sense (i.e., city, lake, river, etc.) which are used in conjunction with <name> elements are turned into <geogFeat> elements. Figure 8 shows two examples of <geogFeat> elements.

Figure 8: Annotation of geographical feature names

According to the Namesdates module, the <geogName> element identifies a name associated with some geographical feature such as ‘River Thames’ or ‘col de la Vanoise’. Thus, <rs type="expandedName"> elements which refer to geographical names are turned into <geogName> elements. Figure 9 shows an example of annotation of a <geogName> element.
As we have seen in the definition of the concept of ENE and in the description of the <rs> element, we defined several levels of encapsulation. The global n attribute may be also used to indicate the level of encapsulation of <geogName> elements (in the same way that it was done for the <rs> element). Figure 10 shows an example of encapsulation of two <geogName> elements.

According to our customized specification, the attributes type and subtype are optional and their values refer to the nature of the geographical feature such as lake, mountain, valley, city, etc. In the current version of the language, we propose to follow the classification introduced for the GeoNames Ontology9. The nine categories of feature classes of GeoNames are shown in Table 6. The code column lists the nine possible values for the type attribute and the feature code column shows some examples of feature classes which are used for the subtype attribute. And according to our customized specification, these two attributes can be also used for the <name> element when it is not included in a <geogName> element.

According to the TEI Guidelines, the <offset> element marks that part of a relative temporal or spatial expression which indicates the direction of the offset. Thus, the generic <term type="offset"> elements referring to spatial relations are turned into <offset> elements in the geospatial semantic layer of our proposal.

According to our specification of the geospatial semantic layer, the <offset> element annotates spatial relations expressed in texts. At the current level of development, we distinguish three categories of spatial relations: topological relations [2], directional relations [3] and distances. Currently we consider two main types of topological relations: adjacency and inclusion. Furthermore, distance relations are described with the <measure> element in the paragraph below.

Table 7 shows the current set of <offset> attributes defined by our customized specification and Figure 11 shows three examples of annotation of <offset> element.

The type attribute is mandatory and its value shall be: orientation, direction initial, direction final, meet, inside, ....

The value of the optional subtype attribute depends on the value of the type attribute. For the orientation type, the value of the subtype attribute shall be: south, east, northwest, above, behind, etc. For the adjacency type, the value of the subtype attribute shall be: next, near, etc. And for the inclusion type, the value of the subtype attribute shall be: in, inside, etc.

According to the TEI Guidelines, the <measure> element contains a word or phrase referring to some quantity, usually comprising a number, a unit, and a commodity name. Thus, according to our specification of the geospatial semantic layer, the <term type="measure"> elements referring to distance relations are turned into <measure> elements in the

---

Table 6: GeoNames feature classes

<table>
<thead>
<tr>
<th>Name</th>
<th>Code</th>
<th>Feature code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative boundary</td>
<td>A</td>
<td>first-order administrative division (ADM1).</td>
</tr>
<tr>
<td>Area</td>
<td>L</td>
<td>locality (LCTY), park (PRK).</td>
</tr>
<tr>
<td>Hydrographic</td>
<td>H</td>
<td>stream (STM), lake (LK).</td>
</tr>
<tr>
<td>Hypsographic</td>
<td>T</td>
<td>valley (VAL), pass (PASS).</td>
</tr>
<tr>
<td>Populated place</td>
<td>P</td>
<td>populated place (PPL), farm village (PPLF).</td>
</tr>
<tr>
<td>Road / Railroad</td>
<td>R</td>
<td>trail (TRL), street (ST), road (RD).</td>
</tr>
<tr>
<td>Spot</td>
<td>S</td>
<td>school (SCM), resthouse (RHSE).</td>
</tr>
<tr>
<td>Undersea</td>
<td>U</td>
<td>canyon (CNYU), reef (RFU).</td>
</tr>
<tr>
<td>Vegetation</td>
<td>V</td>
<td>forest (FRST), cultivated area (CULT).</td>
</tr>
</tbody>
</table>

---

Table 7: Attributes for <offset> tag

<table>
<thead>
<tr>
<th>type</th>
<th>orientation, direction initial, direction final, meet, inside, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>subtype</td>
<td>sub-categorization</td>
</tr>
</tbody>
</table>

---

Figure 9: Annotation of geographical names

Figure 10: Encapsulation of <geogName> elements

---

Figure 11: Annotation of <offset> elements

---

9http://www.geonames.org/ontology/
geospatial semantic layer of the language.

<table>
<thead>
<tr>
<th>type</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>unit</td>
<td>unit identifier</td>
</tr>
<tr>
<td>quantity</td>
<td>numeric value</td>
</tr>
</tbody>
</table>

Table 8: Attributes for `<measure>` tag

Table 8 shows the current set of `<measure>` attributes defined in our customized specification. All attributes are optional. The `unit` attribute indicates the units used for the measurement. The value shall be expressed in the International System Units (SI)\(^\text{10}\) such as meter and second. The value of the `quantity` attribute shall be a numeric value. Figure 12 shows an example of annotation of `<measure>` element.

Figure 12: Annotation of `<measure>` element

The `<placeName>` element is defined by the TEI Guidelines as containing an absolute or relative place name. Thus, according to our specification of the geospatial semantic layer, the `<rs>` elements of the generic layer of our proposal referring to geographical places (i.e., containing a `<geogName>` or `<name type="place">` are turned into `<placeName>` elements. Furthermore, we specify that `<geogName>` elements must be included into `<placeName>` elements.

Table 9: Attributes for `<placeName>` tag

<table>
<thead>
<tr>
<th>type</th>
<th>absolute or relative</th>
</tr>
</thead>
</table>

Figure 13: Annotation of an absolute `<placeName>`

We distinguish between two types of `<placeName>`: absolute and relative and we define the `type` attribute as optional\(^\text{10}\) http://www.bipm.org/en/publications/si-brochure/

Table 9 shows the current set of `<measure>` attributes defined in our customized specification. All attributes are optional. The `unit` attribute indicates the units used for the measurement. The value shall be expressed in the International System Units (SI)\(^\text{10}\) such as meter and second. The value of the `quantity` attribute shall be a numeric value. Figure 12 shows an example of annotation of `<measure>` element.

| `<measure` xml:lang="en" type="distance" unit="m" quantity="200">
| `<w` lemma="m`e tre" type="N" meters"/>
| `</measure>`

Figure 14: Annotation of a relative `<placeName>`

The `<place>` element is defined by the TEI Guidelines as a generic element containing data about a geographic location. In the current version of our specification of the geospatial semantic layer, we use this element to replace the generic `<rs type="sequence">` element. Thus, we consider that the `<place>` element refers to the definition of a spatial area from the association of several locations (Figure 15). According to our specification, `<place>` elements can contain the `type` and `subtype` attributes described in the `<geogName>` element and referring to the feature class of the spatial object.

| `<place` type="P" subtype="PPLS">
| `<geogFeat`>
| `<w` lemma="le" type="DET">les</w`/>
| `<w` lemma="bourg" type="N">bourgs</w`/>
| `</geogFeat`>
| `<w` lemma="de" type="PREP">de</w`/>
| `<placeName` xml:id="ps2">
| `<name`>
| `<w` type="NPr">Barioz</w`/>
| `</name`>
| `</placeName`>
| `<w` lemma="et" type="PREP">et</w`/>
| `<w` lemma="de" type="PREP">de</w`/>
| `<placeName` xml:id="pl1">
| `<name`>
| `<w` type="NPr">Bieux</w`/>
| `</name`>
| `</placeName`>
The function attribute (optional) indicates the motion class. In the current version of our proposal we consider a set of six motion classes based on the classifications of verbs of motion proposed by [9]: leave, hit, reach, external, internal, and cross. Figure 16 shows the result of the annotation of sentence (16) using the elements available in our geospatial semantic layer.

Figure 16: Annotation of a <phr> element

3.2 Encoding Geometric Properties of Spatial Features

The TEI Guidelines describe also some elements for encoding geometric properties of spatial features. According to the Namesdates module, the <location> element defines the location of a place as a set of geographical coordinates and the <geo> element contains any expression of a set of geographic coordinates. However, geographic coordinates such as latitude and longitude values are not often available directly in the textual description and must be retrieved from external geographic resources.

Figure 17 shows an example of annotation using the <location> and <geo> elements. The <bloc> and <country> elements (optional) indicates the continent and the country respectively to which the location belongs.

As we have defined the concept of ENE as an encapsulation of several levels of expansion (see Section 2.1), according to our specification the <location> and <geo> elements can be nested in various elements depending on the ENE to which it refers. For instance, in Figure 17 the <location> element is nested in the <name> element and refer to the location of the spatial NE ‘Pau’, whereas for relative ENE (e.g., ‘sud de Pau’) the <location> element is nested in a <placeName> element.

Furthermore, a location may be specified by using a non-TEI XML vocabulary such as GML and KML. Then, we also propose a mapping via gazetteer unique identifiers, i.e. the use of RDF identifiers to interlink with resources on the Web of Linked Data such as Geonames or DBpedia.

3.3 Indication of Uncertainty

The <certainty> element indicates the degree of certainty associated with some aspects of the text markup. This element is described in the Certainty module of the TEI Guidelines.

Table 11: Attributes for <certainty> tag

<table>
<thead>
<tr>
<th>target</th>
<th>URI data pointer</th>
</tr>
</thead>
<tbody>
<tr>
<td>locus</td>
<td>name, start, end, location, value</td>
</tr>
<tr>
<td>assertedValue</td>
<td>alternative value</td>
</tr>
<tr>
<td>degree</td>
<td>degree of confidence</td>
</tr>
</tbody>
</table>

Table 11 shows the current set of <certainty> attributes defined in our specification. The target attribute indicates the element to which the certainty is applied using the URI syntax. The target attribute is optional and if it is not expressed, the certainty relies to its parent element. The locus attribute indicates the aspect concerning which certainty is being expressed. The locus attribute is mandatory and its value shall be one of the following: name, start, location, value. The name value indicates that the uncertainty relies on the name of the element to which the <certainty> element refers. The start, end and location values indicate respectively whether the start, the end or both the start and the end of the element are correctly identified. The value value indicates that the uncertainty concerns the content of the element. The assertedValue attribute provides an alternate value for the aspect of the considered markup. For instance, when the value of the locus attribute is equal to name, the value of the assertedValue refers to the alternate name of the element in question. And finally, the degree attribute indicates the degree of confidence expressed by the <certainty> element.

With respect to the problem of NE classification, the <certainty> element can be used to indicate the degree of certainty of the type assigned to a NE. Figure 18 shows an example of a <certainty> element applied to a <placeName>
element. The name value of the locus attribute and the rs value of the assertedValue attribute indicate that the <placeName> element may be a <rs> element.

Figure 18: Annotation using the <certainty> element

Furthermore, the <certainty> element can be also used to indicate the certainty degree of the toponym disambiguation task. In this case, the <certainty> element must be associated with the <geo> element. Figure 19 shows an example in which the uncertainty relies on the geographical location of the spatial entity.

Figure 19: Annotation using the <certainty> element

4 Discussion

The main idea of this paper is to propose a general framework for people to create their own specific markup language based on a core generic layer particularly dedicated for the text mining tasks such as information extraction, data mining, but also for deeper linguistic processing such as semantic parsing and co-reference resolution. The core generic layer is also ready to use to be layered with more semantic relationships between elements of the text.

To our knowledge NER is an important pre-processing step for most of these tasks and automatic NER process for Indo-European languages might be more or less challenging (e.g. for German it is especially challenging). This is the reason why our contribution is based on examples from this task and our proposal relies on the TREC standard which is widely used in digital humanities and linguistics for Indo-European languages. Thus, the objective is to define several specific languages, each one adapted to a specific need and all based on the same generic core layer. The proposed generic core layer may be used to create and share pre-processed corpus.

Table 12 shows a summary of the different elements defined by the generic and the geospatial layers of our multi-layer markup language.

Table 12: Summary of the tagset defined for the Generic and the Geospatial layer of our multi-layer markup language

<table>
<thead>
<tr>
<th>Textual elements</th>
<th>Tagset of the Generic layer</th>
<th>Tagset of the Geospatial layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>&lt;w&gt;</td>
<td></td>
</tr>
<tr>
<td>sentence</td>
<td>&lt;s&gt;</td>
<td></td>
</tr>
<tr>
<td>punctuation</td>
<td>&lt;p&gt;</td>
<td></td>
</tr>
<tr>
<td>spatial and temporal relations</td>
<td>&lt;term&gt;</td>
<td>&lt;offset&gt;</td>
</tr>
<tr>
<td>measure expressions</td>
<td>&lt;measure&gt;</td>
<td></td>
</tr>
<tr>
<td>expansion of ENE</td>
<td>&lt;rs&gt;</td>
<td></td>
</tr>
<tr>
<td>NE (level 0)</td>
<td>&lt;nPr&gt;</td>
<td></td>
</tr>
<tr>
<td>ENE (level &gt; 0)</td>
<td>&lt;rs&gt;</td>
<td>&lt;geogName&gt;</td>
</tr>
<tr>
<td>verbal phrase</td>
<td>&lt;phr&gt;</td>
<td></td>
</tr>
</tbody>
</table>

Figure 20 shows the result of the annotation of sentence (16) using all the elements and attributes defined in our customized specification of a geospatial semantic language.

Figure 20: Example of annotation

We applied our proposal of a multi-layer markup language for the annotation of geospatial information from textual descriptions of itineraries. Then, we use this encoding of geospatial information in order to automatically reconstruct the described itinerary. The automatic reconstruction of itineraries task is based on a calculation (multi-criteria ap-
proach) using all the elements annotated from the textual description. Given the work in progress of our modeling work, the information concerning the reconstructed itinerary is not yet introduced in the annotation. Despite this we have shown how it is possible to define new layers according to the needs and aims. For instance, users can define a third layer derived from the second layer of the multi-layer markup language using non-tei elements and also non-consuming tags which may interlink already tagged elements such as ne or ene with verbs or any other kind of element (e.g., spatial or temporal relations, etc.). On our side we are considering in the short term to specify a third layer introducing more semantic and dedicated to the representation of motion using some ISO-Space tags such as the spatial link elements (QSLINK, GLINK, MOVELINK, MLINK).

5 Acknowledgments

This work has been partially supported by: the Communauté d’Agglomération Pau Pyrénées (CDAPP) and the Institut National de l’Information Géographique et Forestière (IGN) through the PERDITO project; the Spanish Government (project TIN2012-37826-C02-01); and the Aragon and Aquitaine Regional Governments through the transborder Aragon-Aquitaine cooperation programme 2014.

References


