

# Spatiotemporal annotation: interaction between standards and other formats

Ineke Schuurman\* and Vincent Vandeghinste  
Centrum voor Computerlinguïstiek  
KULeuven  
Leuven, Belgium  
ineke.schuurman@ccl.kuleuven.be

*Abstract* - Standards and the need for standards, for example for annotation purposes, only emerge after a period of time. Before, people just did what they thought was right. This may have resulted in large amounts of data in a format that in the end did not turn out to be on speaking terms with the (new) standard. This format may even have become a de facto standard for a particular language or in a particular domain. In this paper we discuss an approach for situations in which ISOcat is used to mediate between such formats. Another task for ISOcat is to indicate the possible re-use of the output of semantic annotation X using format Y for a new annotation Z. These possibilities are to a large extent determined by the compatibility of the (definitions of the) data categories used in both.

The spatiotemporal annotation schema STEx, as used in the SoNaR-corpus, is central to this paper. Its input consists of other (semantic) annotations. In the TTNWW-project1 STEx is related to relevant standards, like ISO-TimeML, and state-of-the-art formats, like SpatialML. We describe which conditions should be met and how ISOcat can offer a helping hand.

*ISO-standards; semantics; natural language processing; computational linguistics; standardization; spatio-temporal phenomena*

## I. INTRODUCTION

Standards in NLP are very convenient. They allow for a smooth cooperation between tools and resources using the same or related standards. In the SoNaR-corpus, constructed within the Flemish/Dutch STEVIN-programme, and the joint CLARIN-NL and CLARIN-VL project TTNWW, in which, among other things, the tools used in SoNaR are converted into web services in a workflow system, 1) a series of annotation schemas is to be linked to each other, 2) the data categories used are to be introduced in ISOcat, and 3) links with standards are to be established.

---

\* The first author has a second affiliation: UIL-OTS, Universiteit Utrecht, The Netherlands.

SoNaR is funded by both the Flemish and the Dutch governments, via the joint Dutch-Flemish programme STEVIN (<http://taalunieversum.org/taal/technologie/stevin>), TTNWW by the Flemish government and CLARIN-NL.

<sup>1</sup> TTNWW being a joint Flemish/Dutch CLARIN-project .

In this paper we describe these processes from the point of view of STEx, our spatiotemporal annotation schema. In section II STEx is described, in sections III and IV the standards and other formats involved are introduced, whereas section V describes the interactions and relations with STEx. In a last section our plans for the future are presented.

## II. STEx: SPACE AND TIME

Since a few years the STEx (SpatioTemporal Expressions) [1,2] annotation schema is being developed in which spatial<sup>2</sup> and temporal annotations are integrated in one format in order to

- recognize both temporal and geospatial expressions,
- normalize such expressions, and especially
- solve them, i.e. locate spatiotemporal expressions on calendars and maps, mimicking the resolution capacities of an average reader of a text.

The latter often asks for a combined approach: the dates associated with *summer* differ according to the hemisphere in question, when *Mother's day* is celebrated differs almost per country (or even parts of it), whereas the notion *first day of the week* refers to either a Sunday or a Monday, depending on cultural, religious or geospatial factors.

According to ISO 8601 Monday is the first day of the week, but in several countries one may disagree.

Whether *East-Berlin* refers to the capital of the former German Democratic Republic (GDR) or just to the eastern part of the contemporary German capital depends on the period of time involved. However, according to contemporary gazetteers *East-Berlin* is only a city in the US or Canada.

Nevertheless, when one wants to annotate older texts as well, one has to come up with a reasonable solution for such cases.<sup>3</sup> Administrative changes can always occur, cf. the status of Sudan before or after the 9<sup>th</sup> of July 2011. At this date South Sudan became independent of Sudan, the name of the remaining northern part still being Sudan. So the Sudan of before 2011-07-09 was a country in Africa that is located where now both the new

---

<sup>2</sup> Mainly limited to geospatial annotations.

<sup>3</sup> Recently, in Geonames (<http://www.geonames.org>) an attempt was made to handle such cases as well (using Wikipedia), but not yet in a structural way.

republic of Sudan and South Sudan are to be found on updated maps. This is to be reflected in the annotation of texts originating before and after this date. The STE<sub>x</sub>-annotation describes the geospatial and temporal world knowledge of the *intended audience* of a specific text in a way that reflects the state of affairs valid for a *contemporary user*.

We set up a knowledge base containing the geospatial and temporal world knowledge an audience in Flanders<sup>4</sup> or the Netherlands is expected to have, now and in the (recent) past. This means that the most detailed information the knowledge base contains concerns the Netherlands and Belgium, and includes local holidays and other festivities, all the smaller villages and hamlets, i.e. everything that could be represented on a map or calendar. The information concerning temporal and geospatial entities in a) the surrounding countries, b) the other countries in Western Europe, c) other European countries, the United States and the former colonies of both Belgium and the Netherlands, and d) the rest of the world is decreasingly fine-grained.<sup>5</sup>

A temporal axis was also taken into account: more information with respect to the present than the past. The rationale is that texts with a Flemish or Dutch *intended audience* provide more explicit information when locations further away (spatial and temporal) are concerned. While the concept behind our approach will work for all intended audiences, the content of the knowledge base needs adaptation to the specific intended audience.

Thus a reference to the word *Dover* without further specification in a Flemish or Dutch newspaper will always be to the town in the UK, although there are other, larger Dovers in the US. When any of these were meant, this would be explicitly mentioned.

When in a larger Flemish newspaper a reference is made to the village of *Haren* without further clarification, it will be understood as Haren near Brussels, while in a Dutch national newspaper it would be understood as Haren near Groningen.

This approach takes into account the Gricean maxims (in short: “*don’t say too much, don’t say too little*”) in order to disambiguate geospatial and temporal notions in compliance to the intended audience of the text. In STE<sub>x</sub> the metadata concerning the source and data of publication of the text are very important, as they allow the STE<sub>x</sub>-tagger to associate the text with the background information (cultural, religious, ...) stored in the STE<sub>x</sub> knowledge base. Although the annotation reflects the interpretation of the intended audience,

even in older texts it should refer to a contemporary map and calendar.

The STE<sub>x</sub> spatiotemporal annotation schema is designed for (academic) research purposes as well as for (industrial) applications like multi-document information retrieval and summarization on a larger scale (like the archives of a news agency, as in the recently finished AMASS++- project).<sup>6</sup>

With respect to the first, in the Flemish/Dutch STEVIN-programme SoNaR, a 500 million word corpus of contemporary written Dutch [3], is being built. 1 million words are being annotated and manually corrected for part-of-speech, syntax, named entity labelling, co-reference, and spatiotemporal expressions [4]. A subset (500K) is annotated for semantic roles as well.

In SoNaR the STE<sub>x</sub> tagger uses the other (manually corrected) annotations except for the semantic roles.<sup>7</sup> Whereas the resulting spatiotemporal annotation will be delivered in a standoff-format, the correctors are enabled to use all the relevant layers in one simplified tree format. Thus correction becomes much easier for them, not having to switch between files or screens, as shown in Fig. 1.

```

16 obj1 np
17 det LID(bep,stan,rest) de
    COREF
    head: floradorp
    id: markable_529
18 mod ADJ(prenom,basis,met-e,stan) Amsterdame
    NE-Loc
    gebruik: lett
    id: markable_133
    subtype: bc
    COREF
    id: markable_529
19 hd N(soort,ev,basis,zijd,stan) wijk
    COREF
    id: markable_529
20 app N(eigen,ev,basis,onz,stan) Floradorp
    NE-Loc
    gebruik: lett
    id: markable_134
    subtype: bc
    COREF
    id: markable_529

```

Figure 1. Simplified tree format for correction purposes

The STE<sub>x</sub> tagger uses a hybrid approach, applying rule-based and machine learning techniques. Rules are mainly used to detect those entities for which temporal and/or geospatial annotation might be considered, and to detect (based on metadata plus matching background information) which entries of the knowledge base might be eligible (often several ones), and to

<sup>4</sup> Flanders being a region in the northern part of Belgium. The official language in Flanders is Dutch, like in the Netherlands.

<sup>5</sup> The level of detail more or less reflects what you learn in school, that being the information the person who wrote a text (like a journalist) can rely on as being known.

<sup>6</sup> <http://www.cs.kuleuven.be/groups/liir/projects/amass/>

<sup>7</sup> For this layer, PropBank was chosen as annotation format. In PropBank, abstract or figurative expressions like ‘[in his speech]LOC he was talking about ...’, ‘he took it [in his head]LOC to ...’ are considered to be spatial, whereas in STE<sub>x</sub> this is not the case. But especially the fact that semantic roles are assigned to half of the text to be annotated influenced our decision to use this layer only for correction purposes, not as input for the tagger.

guarantee consistency of an interpretation unless indicated otherwise. Memory-based learning<sup>8</sup> is used to select the most plausible interpretation amongst the alternatives offered, taking into account metadata plus background information.<sup>9</sup>

The knowledge base is implemented as a PostgreSQL<sup>10</sup> relational database, to which the STEx tagger connects. The database consists (currently) of 14 different tables, each containing a different type of information of a geospatial or temporal nature:

**Geo:** This is the main table containing geographical information. It is set up as a hierarchy of entries, in which each entry (apart from the top entry 'world') is linked to its holonym through the partof and partofid. The database contains redundant fields, which are there for ease of maintenance and human readability. (55664 entries)

- id: containing a unique identifier number for each entry in the table
- name: the actual name of the entry, as it could appear in a text<sup>11</sup>
- partof: the name of the holonym
- partofid: the id of the holonym
- level: a descriptor of the level in the hierarchy (continent, country, region, province, municipality, place)
- lat: the latitude of the location
- long: the longitude of the location
- iso: the iso abbreviation for continents, countries, and regions
- levelid: a number indicating the level in the hierarchy (1 for continent, 2 for country etc.)
- overrule: to override the default hierarchy. E.g. Istanbul is part-of Turkey, and Turkey is part-of Asia, but Istanbul is part-of Europe. This is explicitly mentioned in order to prevent a deductive Istanbul being part-of Asia
- overrule id: at which level the hierarchy needs to be overruled
- pop: population
- hemisphere: northern or southern hemisphere
- language: language used in the name field
- utc: international time zone

**Adjectives:** list of adjectives referring to place names. The table consists of three fields (1788 entries): name, id and otherid: in some cases the adjective refers to entries in other tables than the Geo table, in which case it is not an integer and therefore requires a separate field

**Altnames:** alternative names (word forms) for

entries in Geo table (or one of the other tables). It contains translations of place names, nicknames or spelling variants (e.g. without accents)

- language: the language of the alternative name
- otherid: in some cases the name refers to entries in other tables than the Geo table, in which case it is not an integer and therefore requires a separate field

**Areas:** list of areas, which are not so clearly or strictly defined, such as West-Europa (Western Europe), 1026 entries. It covers cultural-historical landscapes (e.g. Hageland), conurbations like Randstad (referring to the cities of Amsterdam, Rotterdam, The Hague and Utrecht in the Netherlands), political unions, like European Union (when used metaphorically, as in: *'the cars entered the European Union'*), larger parts of continents (Latin America).

**Historic:** a list of historical (i.e. no longer existing) countries (e.g. Czechoslovakia), counties, duchies, wars etc., allowing STEx to accurately tag texts referring to a different era. It consists of the following fields (1384 entries):

- yearbegin: if identifiable, the start year of the existence of the historic entry
- monthbegin: if identifiable, the start month of the existence of the historic entry
- daybegin: if identifiable, the start day of the existence of the historic entry
- yearend: if identifiable, the end year of the existence of the historic entry
- monthend: if identifiable, the end month of the existence of the historic entry
- dayend: if identifiable, the end day of the existence of the historic entry
- datenose: a boolean field indicating whether the dates are exact or not
- place: the location of the historic entry (in terms of ids referring to other tables)
- placenose: a boolean field indicating whether the places are exact or not
- type: indicating whether it was a country, a duchy, etc.
- partof: indicating where it fits in the geographical hierarchy
- stateof: indicating which contemporary entity nowadays is associated with it (Nazi Germany is associated with Germany, not with Germany plus Austria, Poland, ...)
- year: indicating the year within which a historic event is to be located (Fall of Troy)
- month: idem, for month
- day: idem, for day (Guldensporenslag (Battle of the Gulden Spurs), July 11th, 1302)

**Inhab:** names of the inhabitants of geographical locations.

**Islands:** names of islands, indicating in which sea or ocean they reside, to which country they belong, etc. In STEx we distinguish between countries and islands, even when they fully coincide, as this may

<sup>8</sup> We are using TiMBL [5].

<sup>9</sup> In a national Flemish paper "Haren" will refer to the village near Brussels, in a local newspaper based in the city of Tongeren to the small village of Haren in its vicinity.

<sup>10</sup> <http://www.postgres.org>

<sup>11</sup> Although the fields 'id' and 'name' occur in all tables, they are only mentioned once.

be relevant for the application at hand. Fields (1034 entries):

- partof: indicate to which country or archipel islands belong
- partofid: indicates the id of the country or archipel (from the geo table)
- location: indicates in which sea or ocean they reside
- locationid: id of the sea or ocean in which they reside
- level: indicates whether it is an individual island, an island group part, or an island group
- levelid: a numeric value indicating the level
- geoid: if the island coincides with an entry in the geo table
- parts: indicating which islands form parts of the island group or group part
- partsid: the ids of the parts

Note that apart from the location on a map (i.e. in the Atlantic Ocean), also the administrative, political characteristics are mentioned (associating Greenland with Denmark).

Lakes: names of lakes and where they are located. 242 entries.

- countries: a list of geo entries (not necessarily restricted to countries) in which the lake is located
- surface: the surface of the lake
- ranking: ranking in the top-100 of biggest lakes

Mountains: contains the names of mountains and mountain ranges plus location. 1262 entries.

- countries: the list of countries in which the mountain (range) is located
- partof: indicating the mountain range of a mountain
- level: indicating the level by name in the mountain hierarchy
- levelid: indicating the level by number in the mountain hierarchy
- height: indicating the height of the mountain

Relations: all lexical items indicating a spatial or temporal relation. It is a list of adjectives, adverbs, multi-word entities, nouns, conjunctions, and prepositions, 453 entries.

- pos: the main part-of-speech category
- label: whether it concerns a temporal, spatial, geospatial or nostex relation
- relatie: indicating the value of the rel feature
- language: indicating the language of the name

Rivers: 718 entries.

- cities: which cities are located on the river
- flowout: id of another river or sea into which the river flows out
- countries: the list of countries through which the river runs
- length: the length of the river

Seas: seas, gulfs, bays, narrows, and oceans and where they are located. 110 entries.

- partof: name of the holonym

- partofid: id of the holonym
- level: indicates the level by name in the sea hierarchy
- levelid: indicates the level by number in the sea hierarchy

Time: a list of words with a temporal meaning, such as days of the week, months, holidays etc., (109 entries)

- language: language used in the name field
- pos: the main part-of-speech category
- identical: link with other name in table
- year: the year of the temporal notion
- month: the month of the temporal notion
- day: the day of the temporal notion
- wday: the weekday of the temporal notion
- yearbegin: the start year of the temporal notion
- monthbegin: the start month of the temporal notion
- daybegin: the start day of the temporal notion
- wdaybegin: the start weekday of the temporal notion
- yearend: the end year of the temporal notion
- monthend: the end month of the temporal notion
- dayend: the end day of the temporal notion
- wdayend: the end weekday of the temporal notion
- freq: how often does it occur ('week' for weekday, 'year' for Christmas)
- dur: the amount of units involved (weekend - 2 (days); Pakjesavond - 6 (hours))
- unit: the units involved (hours, days, weeks, ...)
- form: form used to express a range of possible values, like Prinsjesdag (3rd Tuesday in September): XXXX-09-D2,15.21
- noise: a boolean field indicating whether the times (calendar or clock) are exact or not
- geo: indicates for which geo-entity the values hold
- clock: indicates clock times (avond vs evening)
- religion: indicates which religion an entry represents

Source: describes the origin of documents (newspapers, websites) and their characteristics: (120 entries), providing background information.

- type: newspaper, magazine, website,...
- linked-with: dependencies between sources
- format: paper, web
- domain: general, religion, sport, ...
- scope: national, regional, local (plus indication of region etc)
- distribution: weekly, bi-weekly, D1/D6, D1/D7
- update: 24/24, morning, evening
- since: first issue
- geo-location: UK, London, Brussels, ...
- spec-observance: other than observance usually associated with geo
- spec-language: other than official language associated with geo

This particular table plays a vital role in our approach as far as disambiguation is concerned.

*Streets*: list of frequent street name parts, indicating that the word or multi-word concerns a street name. It has the following fields (84 entries):

- head: head word in street name (e.g. street)
- language: the language in which this head forms a part of the street name

This knowledge base is the core of STEx.

### III. STANDARDS

#### A. ISO-TimeML

A definitive version of ISO-TimeML will soon be approved. It is based on TimeML [6]. The basic tags are EVENT, TIMEX3 and SIGNAL, in which EVENT is more or less synonymous with *eventuality*. [7, p.5] TIMEX3 marks (explicit) temporal expressions and SIGNAL relates temporal expressions with events, mostly expressed by prepositions and conjunctions. LINKs are used to connect and order events (TLINK, SLINK and ALINK). The annotations are rather detailed, and, whenever necessary, the date of publication is taken into account in order to derive explicit dates for expressions like *yesterday* and *one year ago*. Furthermore, there is an optional tag CONFIDENCE, reflecting annotator confidence.

An important characteristic of ISO-TimeML and TimeML is that the annotation labels what is literally expressed: when someone says on May 1<sup>st</sup> 2011 to have been in New York *one year ago*, ISO-TimeML considers this person to have been there exactly one year ago, i.e. at 2010-05-01. The modifier APPROX (part of TIMEX3) is used when the claim was that the visit took place ‘more or less one year ago’, i.e. when an uncertainty is explicitly verbalized.

#### B. ISOcat

ISOcat<sup>12</sup> is the Data Category Registry for ISO Technical Committee 37. This committee develops standards for linguistic resources. Some examples of such ISO-standards are Lexical Markup Framework (LMF), Linguistic Annotation Framework (LAF), Morpho-syntactic Annotation Framework (MAF), and Syntactic Annotation Framework (SynAF). ISO-TimeML is a soon to be approved standard for temporal annotation, cf. above, whereas ISO-Space, for spatial annotation, still has a long way to go. All definitions provided in these standards should become available in ISOcat, some already are. Such definitions can, for example, be part of the description of so-called *data categories*. They are provided in English, information in other languages can be provided as desired.

<sup>12</sup> <http://www.isocat.org>

The idea is that information in ISOcat will remain available for a very long time, data categories come with a persistent identifier (pid), the cool URI,<sup>13</sup> and the information they contain will not change after their acceptance as a standard [8]. When, in the course of time, the definition of a specific linguistic notion changes, a new data category (with the same name) will be constructed, with this new definition and a new pid. Another reason for the presence in ISOcat of several data categories with the same name is that in the various domains, or within various theoretic frameworks, definitions may vary.

Furthermore: ISOcat does not only contain standards as everybody can create their own data categories and share them with the community. Thus all existing annotation schemas can be defined in ISOcat, sometimes with the help of a new addition (in preparation): RELcat relating data categories using *partOf*, *subClassOf*, *sameAs*, *almostSameAs*, etc. [9].

### IV. OTHER FORMATS

#### A. ISO-Space

ISO-Space<sup>14</sup> [10] is a standard under development, paying lots of attention to the general spatial characteristics, and, thus far, not that much to geo-spatial ones. Note that ISO-Space is being developed indepently of ISO-TimeML.

#### B. SpatialML

SpatialML<sup>15</sup> is developed by the same group as TimeML, and shows some of the same characteristics (SIGNAL, LINK, etc). Contrary to the current ISO-Space schema, cf. above, much attention is being paid to geospatial phenomena. The labelling relies on gazetteers. Older states-of-affairs (no longer existing countries and the like) not contained in these gazetteers get a code OTHER.

#### C. Named Entity Labelling in Dutch

Labelling of named entities for Dutch [11]<sup>16</sup> distinguishes between persons, organisations, locations, products, events, and miscellaneous. For STEx, the locations-label is very informative, especially because metonyms are associated with a distinctive label using the feature *meto*:

- 1) The major of Leuven decided to build a new town hall
- 2) Leuven decided to build a new town hall
- 3) They played their last match in Nijmegen
- 4) They played their last match against Nijmegen

<sup>13</sup> <http://www.w3.org/TR/cooluris>

<sup>14</sup> <http://sites.google.com/site/wikiisospace>

<sup>15</sup> <http://www.docstoc.com/docs/48973729/SPATIALML>

<sup>16</sup> [http://lt3.hogent.be/sonar/images/4/44/SoNaR\\_NE\\_Richtlijnen\\_20091019.pdf](http://lt3.hogent.be/sonar/images/4/44/SoNaR_NE_Richtlijnen_20091019.pdf)

At the level of spatiotemporal annotation *Leuven* and *Nijmegen* in 1) and 3) should be labelled as geospatial entities, whereas in 2) and 4) they should get a tag *nostex*, indicating that it does not concern a STExpression. The tag *meto* is providing more than enough training material to apply machine learning techniques in future applications.

#### D. COREA

Within SoNaR, as far as annotation of co-reference relations is concerned, the COREA annotation schema<sup>17</sup> distinguishes ten types of co-reference, such as strict co-reference, part/whole, type-token, metonymy, bound anaphora, and R-pronouns [12].

The occurrence of a co-reference relation between two constructions of which one is recognized at the level of STEx as spatial or temporal is very informative, as is the label ‘*metonym*’.

### V. INTERACTING AND LINKING

#### A. Interacting with other formats

The question is: when are other annotation layers useful?

With respect to named entity labelling: The definition of *event* in [11] differs from the definition used in STEx. Although this is not really an issue as STEx only re-uses the *locations*, there are some types of locations that are considered metonyms in STEx (and therefore will get the label ‘*nostex*’), whereas these are labelled as standard locations at the level of named entity labelling, fully in line with the guidelines. An example:

5) Vietnam heeft diepe wonden nagelaten  
(Vietnam left deep scars)

LOC.x.meto.EVE.x<sup>18</sup>

6) Hij heeft Wimbledon gewonnen (he won Wimbledon)

LOC.bc.lett<sup>19</sup>

In 6) the tournament was named after an inhabited place, therefore it is not considered a metonym at the level of named entities. In 5) on the other hand *Vietnam* stands for *Vietnam War*, and therefore is a metonym at that level. In STEx both are considered metonyms, and therefore the label ‘*nostex*’ is assigned to both. Contrary to the divergence between both layers with respect to the interpretation of the notion *event*, this is a mismatch that might have consequences when it is disregarded. Within the SoNaR-project the correctors are told to pay attention to these cases, a

structural solution is not yet implemented as the number of occurrences is rather small.

Also with respect to co-reference annotation a mismatch shows up: the lacking of co-reference relations concerning adverbs with an incorporated R-pronoun, like ‘*erin, daarop*’ (in it, on it) which is not explicitly mentioned in the manual. These are very important for spatiotemporal analysis in case one wants to reason at the level of the text.

Therefore, at the level of spatiotemporal annotation using STEx one needs to know exactly what is being annotated at the previous layers in order to decide whether the output can be re-used with no further ado. One therefore is to know what the annotations intend to cover by checking their definitions. But not all manuals contain such clear definitions. Often descriptions are rather verbose.

This is where ISOcat may come in as in ISOcat rather short, explicit definitions are asked for.

Such definitions can even be more informative when they are complemented with telling examples, also negative ones. For example, when defining the NE-annotation schema used in SoNaR, it would be good to have a negative example showing that “Wimbledon” in 6) is not considered a metonym.

The ‘*Note*’-section coming with the data category in ISOcat would be a proper place to state which cases are not handled contrary to the expectations a user might have. In the case of the COREA annotation schema in SoNaR this concerns the so-called R-adverbs: ‘*erin, daarop, ...*’ (in it, on it). In this case there is no mismatch between formats because of different perceptions, but because certain phenomena are not covered.

The proposed use of ISOcat heavily relies on the quality of the definitions provided in the data categories. This means that all linguistic notions used in the definitions also need to be described, even when they are not contained in the annotation schema. Otherwise the resulting definition is too vague to be useful.

In order to avoid a proliferation of annotation schemas, ISOcat can also be used to select a series of data categories for a new language or domain. Note that one should not combine randomly: the selected data categories should share their theoretical background. But how to find out which instantiations of data categories in ISOcat belong together?

The preferred way of making clear which instantiations are connected is by both attaching the ISOcat-pids to the resources they are used in and to refer to them in the corresponding manual as well. For the time being one also could mention the name of the full set to which a specific data category

<sup>17</sup> <http://www.cnts.ua.ac.be/~iris/corea.html>

<sup>18</sup> LOC.x.meto.EVE.x : a location metonymically used as event

<sup>19</sup> LOC.bc.lett: literally interpreted location being an inhabited place

belongs in the ‘Note section’ in ISOcat itself.<sup>20</sup> In the future there will be an extension of ISOcat, called SCHEMACat in which annotation schemas can be stored permanently using their pids as handles [9].

### B. Linking with the standard

As standards only emerge after a while, there will be annotation formats around not complying with them. Whereas LAF, MAF, SynAF and the like are rather ‘relaxed’, i.e. they allow for many annotation schemas, ISO-TimeML and ISO-Space tend to be rather strict. This makes it more difficult to adapt existing formats to these new standards,<sup>21</sup> provided one wants to. Especially when an existing annotation format serves as a kind of *de facto* standard for a domain or a language,<sup>22</sup> or when transforming an annotation schema would have a serious impact, for example because the basic principles are rather different, one may want to refrain from a transformation.

Nevertheless, having (more or less) the same annotations for a plethora of languages and in various domains would also be a big asset. Is there a middle course in formally relating the standard format and another format?

### C. Relating STEx with ISO-TimeML

When we transform the STEx annotation into the formats used in ISO-TimeML and SpatialML or ISO-Space, in se independent schemas, some information would get lost. Combined annotations would no longer be possible, for example denoting the former country of Czechoslovakia.<sup>23</sup> It would also become more difficult to assign temporal values taking into account geospatial characteristics, and the other way around.

```
<stex>
  <geo type="country" val="EU::CS">
    <parts>
      <geo type="country" val="EU::CZ"/>
      <geo type="country" val="EU::SK"/>
    </parts>
  </geo>
  <temp type="cal" val="1918/1990"/>
</stex>
```

Nevertheless, relating the STEx annotation scheme and the standard would be informative, for example in the context of CLARIN. A researcher might be confronted with one resource annotated using STEx, and another one using ISO-TimeML. For many phenomena in both schemas a formal relation can be formulated, sometimes in an easy way (event and eventuality in STEx being two different notions, contrary to ISO-TimeML), sometimes more complex (use of formulas in STEx,

LINKs in ISO-TimeML). Defining all data categories in ISOcat (and RELcat) is feasible, and these can be used to express the relations formally.

However, there seems to be one real issue, a kind of paradigm clash: the role of the feature ‘noise’ in STEx.<sup>24</sup>

### D. Noise

STEx is not designed to annotate what is expressed, i.e. to come up with literal interpretations, but to annotate what is expected to be meant. In STEx, when someone says on the first of May 2011 to have been in New York “one year ago”, this person is considered to have been there *more or less* one year ago, i.e. around 2010-05-01. And this is expressed by adding `noise="yes"` to the annotation. Only when the claim would have been that the visit took place ‘exactly one year ago’, the noise-feature would have been left out (or represented as `noise="no"`). This is also the case when someone would claim to have been in New York “sixteen days ago”, whereas “two weeks ago” or “fourteen days ago” will get `noise="yes"`. The point being here that certain numbers are often used in a vague way. We call these ‘containers’. For the unit ‘day’, such containers are 10, 14, 20, 30, 100, ..., 365, for the unit ‘week’ these are 0.5, 1, 1.5, 2, ... . In a similar way, saying that it is 9:20 PM often means that it is more or less that time, whereas 9:19 PM will be considered an accurate expression.

This characteristic of STEx turns out to be essential in applications like multi-document summarization, but also for information-extraction and the like: “people talk sloppy”. It is essential for both temporal and geospatial expressions. The notion “England”, for example, is used to refer to the region England (correct), but also to Great Britain or the United Kingdom as a whole.

\*\*\*

At the moment, it does not seem to be possible to define such top-level characteristics, i.e. the basic assumptions of an annotation scheme in ISOcat, as it is a characteristic of the schema as a whole, and not of a data category it contains. However, it is expected that SCHEMACat will offer the possibility to characterize a schema as a whole.<sup>25</sup>

## VI. PLANS FOR THE FUTURE

As far as work on STEx is concerned, we need to extend the eventualities we handle. At the moment we neglect eventualities expressed by deverbal nouns. Furthermore, we want to cover the spatial expressions in general, not mainly those

<sup>20</sup> At this moment it is not yet possible to search for the name of a tagset or the like in ISOcat using the web interface. This will be remedied.

<sup>21</sup> Unless these are based on timex2 or the like, cf. [12].

<sup>22</sup> In such a case the costs might be high, especially when this format is used in several corpora and other applications (as is the case in the STEVIN-programme).

<sup>23</sup> This is a simplified version without coordinates etc.

<sup>24</sup> Note that “noise” and “confidence” are no synonyms. The notion ‘confidence’ is used, also in STEx, when a toponym like “Dover” is used while it is unclear which Dover is meant.

<sup>25</sup> This is not an isolated case. In the TTNWW-project we also need an characterization at the top-level of a schema, for example to distinguish between function- and form-driven annotation schemas. (is ‘poor’ in ‘The poor may have had no time to ...’ a noun or an adjective?)

referring to geospatial entities. We also have to implement a fall back strategy to use when our knowledge base does not contain the proper information, using, for example, DBpedia.<sup>26</sup>

#### ACKNOWLEDGMENT

We would like to thank Menzo Windhouwer for his valuable contributions regarding ISOcat. All remaining errors are ours.

#### REFERENCES

- [1] I. Schuurman (2007). "Which New York, which Monday? The role of background knowledge and intended audience in automatic disambiguation of spatiotemporal expressions," in P. Dirix, I. Schuurman, V. Vandeghinste and F. Van Eynde (eds) Computational Linguistics in the Netherlands. Selected papers from the seventeenth CLIN meeting (CLIN-17, Leuven), LOT Utrecht, pp. 67-81.
- [2] I. Schuurman, and V. Vandeghinste (2010), "Cultural aspects of spatiotemporal analysis in multilingual applications," Proceedings of LREC 2010, Malta.
- [3] N. Oostdijk, M. Reynaert, P. Monachesi, G. van Noord, R. Ordelman, I. Schuurman, and V. Vandeghinste (2008), "From D-Coi to SoNaR: A reference corpus for Dutch," Proceedings of LREC 2008, Marrakech.
- [4] I. Schuurman, V. Hoste, and P. Monachesi (2010), "Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch," Proceedings of LREC 2010, Malta..
- [5] W. Daelemans and A. Van den Bosch (2005). Memory-based language processing. Cambridge, UK. Cambridge University Press.
- [6] R. Sauri, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky (2005), "TimeML annotation guidelines." Tech. report, version 1.4.
- [7] C. Tenny and J. Pustejovsky (2000), "A history of events in linguistic theory," in C. Tenny and J. Pustejovsky (eds) Events as grammatical objects. The converging perspectives of lexical semantics and syntax. CSLI Publications, Stanford.
- [8] M. Windhouwer (2010), "Semantic interoperability of linguistic resources now and in the future," LAT, <http://www.latmpi.eu/latnews/tag/isocat>.
- [9] M. Windhouwer and I. Schuurman (2011). "RELcat and friends," (<http://lux12.mpi.nl/isocat/files/manual/>, item: CLARIN-NL ISOcat tutorial presentation, 4).
- [10] J. Pustejovsky, J. Moszkowicz, and M. Verhagen (2011), "ISO-Space: The annotation of spatial information in language," Proceedings of ISA-6: ACL-ISO International Workshop on Semantic Annotation, Oxford, England, January 2011.
- [11] B. Desmet and V. Hoste (2009), Annotatierichtlijnen voor Named Entities in het Nederlands, versie 0.1. L3 technical paper, Ghent, Belgium.
- [12] G. Bouma, W. Daelemans, I. Hendrickx, V. Hoste, and A.-M. Mineur (2007), The COREA-project. Manual for the annotation of coreference in Dutch texts.
- [13] E. Saquete and J. Pustejovsky (2011), "Automatic transformation from TIDES to TimeML annotation," in Language Resources and Evaluation., Springer.

---

<sup>26</sup> <http://wiki.dbpedia.org/OnlineAccess>