Spatiotemporal annotation using MiniSTEx:
How to deal with alternative, foreign,
vague and/or obsolete names?

Ineke Schuurman

Centrum voor Computerlinguïstiek
K.U.Leuven
ineke.schuurman@ccl.kuleuven.be

Abstract

We are currently developing MiniSTEx, a spatiotemporal annotation system to handle temporal and/or geospatial information directly and indirectly expressed in texts. In the end, the aim is to locate all eventualities in a text on a time axis and/or a map to ensure an optimal base for automatic temporal and geospatial reasoning. A first version of MiniSTEx was originally developed for Dutch, keeping in mind that it should also be useful for other European languages, and for multilingual applications.

In order to meet these desiderata we need the MiniSTEx system to be able to draw the conclusions human readers belonging to the intended audience would also draw, e.g. based on their (spatiotemporal) world knowledge, i.e. the common knowledge such readers share. The world knowledge MiniSTEx uses is contained in interconnected tables in a database. At the moment it is used for Dutch and English. Special attention will be paid to the problems we face when looking at older texts or recent historical or encyclopedic texts, i.e. texts with lots of references to times and locations that are not compatible with our current maps and calendars.

1. Introduction

The information obtained by spatiotemporal annotation of texts (*where* and *when* did *X* happen?) can be useful in for example information retrieval, question answering or multi-document (and multilingual) summarization, in order to determine in which order events happened, whether the same event is discussed in several documents, etc.

The development of a first version of a spatiotemporal protocol meant to be used for corpus annotation was carried out within the context of the STEVIN-project D-Coi,[1] cf. Oostdijk et al. (2008).

In a current project, AMASS++,[2] texts like the ones in digitized archives of Flemish and Dutch news agencies and broadcast companies need to be automatically analysed and indexed in order to secure optimum access to their contents. Reporters, our users, may want to know where beguinages are found. Such a question can be formulated in several degrees of specificity:

(1) In welke ∅/Europese/Belgische/Vlaamse/Vlaams-Brabantse steden staan begijnhoven?

  In which ∅/European/Belgian/Flemish/Flemish-Brabant towns can beguinages be found?

Leuven, having two beguinages, therefore needs to have an annotation in which all these degrees are mentioned, in order to qualify as a candidate in all these questions.

In other occasions the tag should even be more specific:

(2) Het meisje uit Leuven was de winnaar.

  The girl from Leuven was the winner.

(3) Het meisje uit Heverlee was de winnaar.

The girl from Heverlee was the winner.

At first sight there seem to be two different winners (and therefore two different contests?). But as *Heverlee* is a village which is part of the municipality of *Leuven*, the same girl can be referred to. This asks for a more finegrained annotation, saying that the village of *Heverlee* is contained in the municipality of *Leuven*.

In this paper we will explain how, in our knowledge-based approach, we tackle spatiotemporal names making use of a database in combination with the notions *background knowledge, intended audience*, and *present-day user* (Schuurman, 2007b; Schuurman, 2007c), and why we want to be as specific as shown above. Special attention will be paid to the problems we face when looking at older texts or recent historical or encyclopedic texts, i.e. texts with lots of references to times and locations that are not compatible with our current maps and calendars, such as *Czechoslovakia, DDR, Zaire* or *October Revolution, Chinese New Year*.

2. Some characteristics of MiniSTEx

The MiniSpatioTemporal Expressions (MiniSTEx)[3] annotation system, cf. Schuurman (2007b), Schuurman (2007c), is meant to automatically annotate all spatial (including geospatial) and temporal elements in texts. It is also perfectly possible to use only one of the components. The rationale behind tackling these phenomena together is that a) the problems we are facing are largely the same, and b) there are all kinds of connections between both components.

So, a first characteristic of MiniSTEx is that it handles temporal and (geo)spatial annotation in one go, using largely

---

the same approach. It also handles *geotemporal* expressions, i.e. expressions associated with a combination of geospatial and temporal properties (for example in order to express that between the First and the Second World War *Libya*, nowadays an independent country, was a province of Italy).

A second characteristic is that full advantage is taken of the fact that the origin of the texts to be handled is known. The metadata contain at least the date (sometimes even the time) and place of publication, and the origin of the text. From the latter the `background` of the publication can be determined, and thus the `intended audience` of the text can be inferred: who is (or was) adressed? This determines to a large extent how a spatiotemporal expression is interpreted, taking into account Grice's maxims (cf. section 3.). This means that the most obvious interpretation of a (spatiotemporal) expression often will not be clarified by the author, whereas other interpretations will. In this section the notions `background knowledge`, `intended audience`, and `present-day user` will also be explained.

The currently most well known scheme to annotate temporal data is TimeML (Sauri et al., 2006). It allows for the identification of events and the temporal properties they express. The format used for anchoring temporal expressions in MiniSTEx is more or less the same: YYYY-MM-DD for dates and HH:MM:SS for times (combined as YYYY-MM-DDTHH:MM:SS), in which entities may be left out from right to left, while an unknown unit on the left side (such as the year) will be filled out as XXXX. However, in TimeML expressions like *summer* and *Thanksgiving* are not quantified at all, let alone expressions like *Second World War* and the temporal information contained in proverbs like *go to bed with the sun*. In MiniSTEX, however, such expressions are quantified, using a whole series of mechanisms. Considerable effort has been put (and is still being put) in the construction of a database containing the (spatiotemporal) knowledge the audience of a text is expected to have, cf. Schuurman (2007c), as we want MiniSTEx to draw the same kind of spatiotemporal conclusions the human intended audience would draw.

Note that in the first three examples (*summer, Thanksgiving*, and *Second World War*) the value to be assigned clearly depends on the place on earth considered: *summer* in Australia, i.e. the southern hemisphere, is when it is *winter* in Europe (northern hemisphere), *Thanksgiving* in Canada and the USA is celebrated on different dates, and in the Netherlands and Belgium the beginning of the *Second World War* is associated with 1940, whereas this is likely to be 1939 in a country like Poland. The USA, on the other hand, only got involved end of 1941, and therefore many USA-citizens may consider December 1941 as the beginning of World War II.

For annotation of geospatial expressions in natural language up till now no clear standard, like TimeML for temporal annotation, has emerged. Recently the annotation scheme SpatialML, cf. MITRE (2007), turned up. SpatialML is clearly related to TimeML in its design, although there are no clear links between the two schemes as far as the actual tags are concerned, i.e. there is nothing like the combined `geotemp` tag used in MiniSTEx.

In SpatialML expressions like *Czechoslovakia* (we will call this kind of expressions *obsolete* as this entity no longer exists) and *Middle East* (which we will call *vague* as it is difficult to determine its borders) are not quantified.

In MiniSTEx, however, expressions like *Czechoslovakia* and *Middle East* are located on the map on the map, cf. section 6.3.

With respect to geospatial data temporal information might be of importance: the *Democratic Republic of Congo*, an ex-colony of Belgium, was called Zaire between 1971 and 1997; *Suriname* is independent since 1975, before that it was related to the Netherlands for many years, and Brazil got itself another capital: *Brasilia*. Before 1960 Rio de Janeiro used to be its capital. Or entities, like *Czechoslovakia* cease to exist. Temporal data are either necessary in order to decide what the status of an entity is, or the names themselves provide information with respect to the times associated with them: The use of the name *Weimarer Republik* means that the text in which it is used is not written before 1918, and it refers to an entity that did exist between 1918 and 1933.

The handling of spatiotemporal names is just one facet of the whole system. We also take care of tense and aspect, relations between names, shifts of perspective (both temporal and spatial), the classes of verbs involved: reporting, intention, negative,... (the latter inspired by TimeML). For more details, cf. Schuurman (2007a).

In both D-Coi and SoNaR, a project that is expected to follow D-Coi in order to build a 500 million word reference corpus of Dutch,[4] the input consists of syntactically annotated sentences (trees) in XML-format. MiniSTEx adds extra information to these trees. For AMASS++, we will look into the possibilities of using chunked sentences, i.e. a less deeply structured input.

3. Grice's maxims, background knowledge, intended audience and present-day user

In daily life, one way or another, everybody uses the Gricean maxims in written or spoken communication, although many people never saw them formulated as they are a matter of conventional wisdom:

- Maxim of Quantity

- Maxim of Relation (or Relevance)

- Maxim of Manner

- Maxim of Quality

Every author will apply these conversational maxims, often paraphrased as "Don't say too much and don't say too little.", cf. Dale and Reiter (1996).

As said before, this usually means that the most obvious interpretation of a (spatiotemporal) expression often will not be clarified by the author.

---

[4]One million of these will be semantically annotated (named entity recognition, coreference, semantic roles and, as the last component, spatiotemporal semantics), and corrected.

To know what the most obvious interpretation is, we use the `background knowledge` coming with a document: is it a national or a regional newspaper; is it based in Flanders, the Netherlands or elsewhere; does it cover news in general or is it focussing on a more specific topic, like business news; who is the publisher; to what tradition does it belong; in which language is it written; which calendar does it use, ... Its scope is also very important: is it global, national, regional, local?

In a national Belgian newspaper based in Brussels the use of the notion *summer* without further specification will refer to the months of June, July and August, as Belgium is in the northern hemisphere, whereas the relevant months will be mentioned when a reference is made to summer in countries like Australia or Brazil, i.e. the southern hemisphere. The same holds for toponym resolution: when in the same newspaper no further specifications are given the toponym *Haren* will refer to the village in the Brussels Capital Region (same region), although for example the village with the same name in Germany has a larger population. But when the much lesser known (and smaller) village *Haren* belonging to the municipality of *Borgloon* in the province of Limburg (Flanders) is meant, this will be mentioned explicitly. However, in a Borgloon based local newspaper, the default interpretation will be that of the nearby *Haren* in Limburg.

Considerations like these play a major role in the disambiguation process. In our system, when deciding which town or village is referred to, the number of inhabitants, which is often taken as a major feature in geographic information retrieval (Ding et al., 2000; Leidner, 2006; Leidner, 2007; Volz et al., 2007), only plays a minor role. Other factors are more relevant.

The `intended audience` refers to the people for whom a specific text is written. This may refer to the time, the scope, the country, the tradition, the orientation, ...

We assume that a text always provides its intended audience with all information necessary to understand this text. If not, i.e. when a human reader belonging to the intended audience fails to understand a text, the system can not be blamed for failing. MiniSTEx handles texts by using the `background` and world knowledge the intended audience is supposed to have, storing it in a large database.

Note that for example when dealing with a 1968 newspaper other things are presupposed than in a 2008 news item, as the world did look different those days. Such data will gradually be incorporated in the database, always mentioning when a specific state of affairs is valid. Of course, when a whole corpus/archive of older texts is to be handled, a new table is to be added to the database, covering that period in more detail.

Having such an intended audience seems to be a vital property of a text: a medical text written for British GPs is not likely to be fully understandable for either aerospace engineers, teachers or linguists, nor for Belgian GPs. A text written for people living in Amsterdam may not be understandable for people living in Brussels or Rotterdam when referring to local information: at least at the local level their presupposed geospatial world knowledge is not the same. In a local newspaper many details are supposed to be known, and should therefore be contained in the database. In a national or regional newspaper, aiming at an audience over large parts of the country, such local details will be mentioned explicitly. The same holds for news on other parts of the world. In this, economical, historical (Belgian people are supposed to know things about Congo, the Dutch about Suriname, these countries being their respective ex-colonies), and cultural links come into play: the intended audience is supposed to know more about countries with which there are such links. The knowledge to be presupposed also depends on its source: when reading a very local newspaper one needs to know other things than when reading a large national newspaper, the same difference between a jewish, a general or a communist newspaper. And many Flemish people reading something in a Dutch newspaper about the Dutch *Koninginnedag* (lit. 'queenday', the birthday of the queen) will not know it is celebrated on April 30, and does not celebrate the birthday of the present queen (Beatrix), but that of the former one (Juliana).

Information about the intended audience can be gathered from the background knowledge associated with a publication, together with for example the date of publication, cf. table 1.[5]

Note that the present-day user of the annotated texts (for example a reporter) does not need to belong to the intended audience. The item under consideration can be on the fall of the Berlin Wall in 1989 and all events that followed (like the collapse of Czechoslovakia and Yugoslavia), and written in 1994. So the intended audience was living in 1994, knowing the state of affairs holding at that moment. But the present-day (2008) user is confronted with another state of affairs, for example with respect to Croatia. We therefore want to describe the spatiotemporal state of affairs in 1994 in terms of the state of affairs of 2008. This means that the MiniSTEx database is updated (entries are changed and/or added to tables) whenever major spatiotemporal changes occur.

Questions asked by this present-day user will be interpreted according to the current state of affairs, unless the user states otherwise.

It is clear that MiniSTEX needs to store all this information, for example knowing that *De Morgen* is a Flemish[6] newspaper, cf. tables 1 and 2. It also needs to store especially the common world knowledge both the intended audience and the present-day user are supposed to have in order to be able to analyse and disambiguate the texts, i.e. both the items in the newspapers and the questions formulated by the user.

In table 1, to be used in combination with table 2, some information with respect to newspapers, broadcast companies and the like are collected. The 'neutral' values for a.o. cultural tradition, calendar, and language are shown in table 2. For other cues the news items themselves need to provide information.

---

[5]This is just a table to present the information the real database contains. This database is in PostgreSQL (http://www.postgresql.org/)

[6]One of the many roles of Brussels is being the capital of Flanders, although it is located in the Brussels Capital Region.

Table 1: Background-doc

concept	dbid	status	geo-place	trad	cal	lang	time	orient	scope
De Morgen	220000	newspaper	Brussel			Dutch	contemp	gen	regional
De Telegraaf	220003	newspaper	Amsterdam				contemp	gen	national
Joodse Courant	220001	newspaper	Antwerpen	jew	jew	Hebrew,	contemp	gen	national
						Dutch	contemp	gen	national
Ref. Dagblad	220009	newspaper	Apeldoorn	orth-ref			contemp	gen	national
Medisch Contact	220069	prof.journal	Utrecht				contemp	med	national
Dagblad van het Noorden	220015	newspaper	Groningen				contemp	gen	local
New York Times	220051	newspaper	New York				contemp	gen	national
The Times	220053	newspaper	London				contemp	gen	national
TimesOnline	220054	newspaper	London				contemp	gen	national
Vlaamse overheid	230000	web	Brussel			Dutch	contemp	gen	regional
Vlaamse overheid	230000	web	Brussel			English	contemp	gen	global
VRT	230003	broadcast	Brussel			Dutch	contemp	gen	regional
NOS	230005	broadcast	Hilversum				contemp	gen	national

Table 2: Background-geo

concept	dbid	status	trad	cal	hem	UTC[7]	lang	partof	division
España	109	country	chr	greg	north	+1	Spanish Catalan Vasco Gallego	EU	2=comunidade, 3=provincia
Nederland	146	country	chr	greg	north	+1	Dutch Frisian	EU	2=–, 3=provincie
België	137	country	chr	greg	north	+1	Dutch, French, German	EU	2=gewest, 3=provincie
US	199	country	chr	greg	north	-(5/10)	English Spanish	NA	2=state, 3=county
Vlaanderen	102	region					Dutch	BE	

4. A knowledge-based approach: the MiniSTEx database

From the previous it will be clear that a large database plays a central role in our system. The various tables in this database contain spatiotemporal information from single tokens over full NPs and PPs to complete proverbs and the like.[8]

After disambiguation of the spatiotemporal elements contained in a text, both machines and humans should be able to reason based on the spatiotemporal information provided. As we want to present the information in a way useful for both humans and machines, we will for example not provide just the coordinates (lattitude and longitude) of a town like *Leuven*, but also express in a more verbatim way where this town is located:

```
<geo type="place" val="EU::BE::VL::VBR::
Leuven"⁹ coord="+50.87+04.70"¹⁰/ >.
```

This verbatim way of annotating enhances reasoning, as it shows that for example *Leuven* is in *Europe*. As we, at least for the moment, are just using 'point coordinates', which only show roughly where something is located instead of representing its borders, these coordinates are in several cases too imprecise for spatial reasoning.
A date, like *May 9, 1960*, is expressed as

```
<temp type="cal"¹¹ val="1960-05-09"/ >¹²
```

i.e. as "YYYY-MM-DD", cf. TimeML. In addition we use e.g. a notation for NPs expressing a certain period, like *First World War*

```
<temp type="cal" val="1914/1918"/ >¹³
```

or *summer*

---

[8] At the moment we concentrate on geospatial expressions as far as the spatial component is concerned.

[9] Up till provinces we use the two and three letter abbreviations proposed in ISO 3166. For other entities, like towns, we use the full names as used locally (endonym) or their official transliterations in the Latin alphabet, see section 6.2.

[10] In de order `lattitude;longitude`, and format

DD.MM.SS (degrees, minutes, seconds), positive for North and East, negative for South and West, cf. ISO 6709. The seconds, or the minites and the seconds, can be left out.

[11] `Cal` stands for calendar.

[12] These are simplifications. Next to the `val` feature there is a series of other features to be expressed as well, among which a fixed `dbid`, cf. Schuurman (2007a)

[13] The range of years associated with First World War in Belgium.

```
<temp val="M06/08"/ >.
```

NPs like these are not quantified in TimeML.

The spatiotemporal information contained in the databases is based on gazetteers, international (ISO) standards, official (governmental) websites, wikipedia, etc. We also incrementally feed the information in annotated texts back in the system.

The database also contains tables with `background` information of documents, as well as information on their `intended audience`. This information is crucial for disambiguation purposes, cf Schuurman (2007c). As explained above, in a regional Flemish newspaper *Haren* will be considered to refer to the the Belgian village near Brussels, whereas in a national Dutch newspaper a Dutch referent is more obvious. It holds for both countries that when a reference is made to *Dover* without further specification, a Flemish or Dutch intended audience will interpret this as the *Dover* in the UK, although there is at least one larger town with the same name in the US (in the state of Delaware).

As remarked before, the knowledge we are storing in the database is the common knowledge both the intended audience and the present-day user possess. Our intended audience are common people, reading newspapers and watching tv. But not, for example, historians.

Therefore the database contains the knowledge one needs in order to understand an item in the news, be it a newspaper or a broadcast, i.e. the common knowledge the author presupposes the reader to have, which may not be 100% correct from a scientific, for example historical, point of view. We often use approximations, for example when specifying historical names (both temporal and geospatial). But also for contemporary names, especially for areas: who knows exactly the boundaries of the Rocky Mountains? Or of the Randstad, a conurbation in the Netherlands, consisting of the four largest Dutch cities (Amsterdam, Rotterdam, The Hague and Utrecht), and the surrounding areas? Almost nobody, but still the average reader knows where to locate these areas, at least roughly. This will also be good enough for MiniSTEx in this implementation aimed at general news.

The same holds for the temporal component: the Vietnam War, for example, will by most people in our countries be associated with the sixties and early seventies of the last century, i.e. with the war in which the US was directly involved. When in a news item it is relevant to know that it ended April 30, 1975, this will be mentioned.

And when the first 'part' of that war is referred to, starting just after the Second World War, and ending in 1954, this will be mentioned as well.

In MiniSTEx, the *Vietnam War* will strongly be associated with US involvement:

```
<temp type="cal" val="1957/1975"/ >
```

Only when in documents the full war (1945-1975) is referred to as Vietnam War, this stretch of time will be added to a new database entry, which will be marked as only to be used when referred to explicitly.

Especially in Southeast Asia, the first part of the war (1945-1954) is also called the French War, because of the major role of France, or the First Indochina War, while the second part (1957-1975) one is also called the American War or Second Indochina War. Such names, when encountered in a text, will be added to the database as alternative names.

The knowledge contained in the database, and especially its level of depth, also depends on the content of the texts to be annotated. Nobody will expect the intended audience of a Dutch national newspaper to know to which municipalty a specific Russian village belongs, therefore everything one needs to know will be spelled out in the item. It might be sufficient to know that *Zhukovka* is a village near *Moscow*, without knowing about any of the (partial) values inbetween. In that case these will be represented with XX.

```
<geo type="place" id="3"
val="EU::RU::XX::XX::XX::Zhukovka"/>
```

When the tag for *Moscow* is

```
<geo type="place" id="4"
val="EU::RU::XX::XX::XX::Moskva"/>
```

the relation between *Zhukovka* and *Moscow* will be spelled out as

```
<rel name="near">
    <geo arg1="3"/>
    <geo arg2="4"/>
</rel>
```

at the level of their mother node.

On the other hand, it doesn't hurt to mention more information when this information is available:

```
<geo type="place" val="EU::RU::XX::Bryanskaya
oblast::  XX::Zhukovka/ >
```

Note, however, that in general not everything contained in the database is used for a specific annotation. What is actually used is determined in preparatory consultation with the client,[14] depending on the objectives of the specific task.

5. The annotation itself

The format, but especially the level of detail provided for locating geospatial entities on a map are different in SpatialML and MiniSTEx.

In SpatialML, in a sentence like

We just had a meeting at LIIR in Heverlee

Heverlee would be annotated[15] as

```
<PLACE type="PPL"[16] country="BE" form="NAM"
latLong="50.8775N 4.7044E" >Heverlee</PLACE >
```

---

[14] The agency or company commissioning the annotation project, not necessarily the user.

[15] When looking at their annotation of both *Madras* and *Rome*

[16] Where PPL stands for 'populated place'.

or, one field deeper:[17]

```
<PLACE type="PPL" state="FL"[18]
country="BE" form="NAM" latLong="50.8775N
4.7044E">Heverlee</PLACE>
```

In the MiniSTEX `geo`-tag we use a 'deeper' annotation, we do not just mention that it is a village in Flanders, which in turn is a region of Belgium, but also that it is part of the municipality of Leuven, which is in the province of Vlaams Brabant. And we also mention that Belgium is part of Europe.

In our annotation this becomes

```
<geo type="place" val="EU::BE::VL::VBR::Leuven::
Heverlee" coord="+50.87+04.70"/>[19]
```

so `EU contains BE` (i.e. Europe contains Belgium), etcetera.

In MiniSTEx we want our annotations to be as informative as possible in order to facilitate reasoning. And we want this information to be easily accessible, i.e. the tagged corpus should be usable as stand alone, not needing the help of an external database. Another desideratum: the annotation format should be useful for several languages, i.e. when both Dutch and English texts have been annotated, the results should be useful when used in combination, for example in multilingual summarization. We therefore want to present the tags as self-relient as possible.

As explained above, in the annotation we are using fixed numbers of fields in the value of the attribute *val* for the range `continent...place`,[20] albeit fields may be left out from right to left. When annotating a noun referring to a village, we are using as many administrative subdivisions as relevant for the Belgian state of affairs:

```
continent::country::region::province::
municipality::place
```

i.e. 6 levels. For the Netherlands 5 levels would have been enough as there are no regions.

This rather finegrained division should be sufficient for other countries as well, al least from the point of view of the intended audience, resp. present-day user of our system: country X may use a deeper division, but in our part of the world we are not confronted with it. The point now is that we use this 6 level deep subdivision for other countries as well, mentioning in the database how a corresponding level is called in the respective countries, i.e. `region` will be `state` in the US, whereas `province` will be `county`. In *Russia* the `oblast` can be considered a `province` and in fact is often called a *provincie* (province) in Flemish and Dutch newspapers. But not all levels will exist in all countries, for example the Netherlands are lacking the

level `region`. This does, however, not imply that we are just using 5 (instead of 6) levels for the Netherlands, as we will fill the field `region` with a dash. This way we can easily see which fields do exist (compare the following examples, where in the second example the non-existing field simply has been left out.

```
<geo type="municipality"
val="EU::NL::-::NH::Amsterdam"/>

<geo type="municipality"
val="EU::NL::NH::Amsterdam"/>
```

In the latter it is unclear what the status of the inbetween fields is.

The type *river* will be associated with the relevant fields of `continent`: in a text on the Netherlands we will associate the river Rhine with the Netherlands, not with the other countries it flows through, unless there is a reason to do so. The same holds for the type `area` (like *Randstad, Rocky Mountains*).[21]

```
<geo type="river" val="EU::NL::Rijn" id="13" >
    <rel name="ends">
        <geo arg1="13"/ >
        <geo arg2="North Sea" type="sea"
        val="AO::North Sea"[22]/ >
    </rel>
</geo>
```

The `ends`-relation expresses that the river Rijn meant is the one flowing into the *North Sea*, and that it is the same one as in

```
<geo type="river" val="EU::DE::Rijn" id="16" >
    <rel name="ends">
        <geo arg1="16"/ >
        <geo arg2="North Sea" type="sea"
        val="AO::North Sea"/ >
    </rel>
</geo>
```

For annotation purposes there are two options: either we always use the terminology used for their Dutch/Flemish counterparts as values for `type`, or the original ones.

For example the Russian concept of *oblast*: according to Wikipedia (Dutch version), an oblast is an entity just below the level of country, and it is something like an area, region or province. In our database the notion is linked with all these, with as default *province* as this is a common further specification.

As can be inferred from table 2 we treat a US county as a province, and a state as a region. The default name in `type` will be the Dutch one, but it is very easy to change it into the original ones (county, state) upon request of the client.

The notation used in the `value` attributes consists mainly of ISO-codes. For the `temp`-tag ISO 8601 is used, for the `geo`-tag ISO 3166. In the latter case we invented new

[17]Based on their example for *New York City*.

[18]In their annotation they are likely to use *FL* (short for the English notion *Flanders*) whereas we opt for the use of the abbreviation *VL* (for the endonym *Vlaanderen.*)

[19]In this notation '`::`' means '`contains`.

[20]For the treatment of oceans, seas, mountains, addresses, . . . , see Schuurman (2007a)

[21]In the database, however, such types are associated with all countries etc that might be relevant.

[22]AO stands for Atlantic Ocean

codes for countries and the like which do no longer exist (like *Soviet Union, Czechoslovakia*), otherwise, especially for lower entities such as municipalities and places, the endonym is conceived as code, cf. section 6.2.

We are using such codes as we want the `value` attributes to be language independent. The `temp`-tags are language-independent by default as these are expressed by numbers, whereas `geo`-tags are much more liable to language influences, especially when no ISO-code is available.

There are in se several ways to achieve language independency, like

1. use one and the same language (like English) for the annotation, regardless the language used in the text and regardless who will use the annotation.

2. use one and the same language for the annotation, regardless the language used in the text. The language used should be that of the (intended) user.

3. use the name in the language used in the territory dealt with in the text, i.e. use endonyms, if necessary transposed into the roman alphabet (in case of languages like Japanese, Russian, Arabic,...)

In the first two situations described above, often exonyms will be used: the original name (toponym) is adapted to the language used in another country (the name of the Dutch town of *Vlissingen* becoming *Flushing* in English).

In MiniSTEx we opt for the use of endonyms at the level of places and municipalities:[23] even when in an English text the name *The Hague* is used, the annotation will be using *Den Haag*. But when in a Dutch text the name *Londen* is used, it will be annotated as *London*, i.e. using the proper endonym.

6. Consequences for alternative, foreign, vague and obsolete names

Below we will provide a short overview of the consequences of our approach when alternative, foreign, vague and obsolete names are concerned.

6.1. Alternative names

Sometimes a spatiotemporal entity is referred to with several names, even within the same language and at the same moment in time. In such cases, the official name will be used in de `val` feature. We are using a rather large table with all kinds of alternative names in order to enable the system to come up with the correct tag, i.e. to identify and disambiguate the names, most of them dealing with `geo` names but part of them with `temp` as well, like alternative names for traditional festivities. The codes used in the `value` attributes will be the official ones.

Note that quite often only the official name will be mentioned on maps, or in gazetteers, which makes it difficult to identify such alternative names. We therefore had to collect such names and their official counterparts ourselves. This was only done for those countries our intended audience and present-day user might be presupposed to be familar

---

[23]And also for area's which can occur at higher levels as well.

with, cf. preceding sections.

Another name of the well known Dutch town of *Den Haag* (The Hague), for example, is *'s Gravenhage*. The official name is *Den Haag*. Also the name *Hofstad* (litt. 'court town') is sometimes used. So *'s Gravenhage* and *Hofstad* may both turn up in the `alt` feature:

```
<geo type="place" val="EU::NL::--::ZH::Den Haag"
alt="Hofstad" coord="+52.05+04.18"/>
```

Whether or not the feature `alt`, displaying the name used in the sentence, will actually be used in the annotation depends on the client.

A name can at the same time be the official name of geographic entity X and also be used as an alternative name for entity Y. *Roosendaal*, for example, is an alternative name for the Dutch village of *Rozendaal* in the province of *Gelderland*, but also the official name of the city of *Roosendaal* in the province of *Noord Brabant*. In MiniSTEx this does not entail that the official reference is to be preferred (cf. the issue with the number of inhabitants). The intended audience etc. are much more important in the disambiguation process.

Sometimes, an alternative name only holds a limited time a year, like the names used during carnival. In such a case a temporal tag is added, reflecting the time of the year carnival is celebrated.

6.2. Foreign names

Texts in a particular language will quite often contain the names of provinces, towns, rivers etc in other countries using the name in the language the text is written in (exonym) or in the official language of that other country (endonym). Sometimes the English exonym is used in other languages as well.

We will use the endonym, cf. above.

In case several countries are involved, for example when the location of a river is concerned, this might lead to endonyms in several languages. In such a case we use the endonym used in the country in which this river flows into another river or in the ocean or the sea. So the endonym to be used for *Rhine, Rijn, Rhein* will be *Rijn*, even when the text deals exclusively with the Rhine in Germany, because the mouth of the river, i.e. its end, is in the Netherlands. A relevant language-dependent exonym, if any, may be mentioned in the feature `exonym`:

```
<geo type="river" val="EU::DE::Rijn" id="16"
exo="Rhein">
    <rel name="ends">
        <geo arg1="16"/>
        <geo arg2="North Sea" type="sea"
        val="AO::North Sea"/>
    </rel>
</geo>
```

The rationale behind adding a `rel`-tag is that there may be several rivers with the same name, even within the same country: There are two rivers *Vecht* in the Netherlands, which are distinguished using the `ends`-relation.

Note that for North Sea, an international territory, we use a name in English because there is no valid endonym.

6.3. Vague and obsolete names

Obsolete names are names for no longer existing entities. These may be vague, but don't need to be: *Czechoslovakia* is obsolete, but by no means vague as its location can perfectly be described, referring to the present-day countries Czech Republic and Slovakia. But a name like Mesopotamia is considered vague as it refers to an area between the rivers Tigris and Euphrates, i.e. Iraq, western Iran and eastern Syria. But what exactly denotes western Iran and eastern Syria?

Names do also change over time. The war that was usually called the *Great War* was named *First World War* after the *Second World War* came into being, *Leningrad* was again called *St. Petersburg* in the years after the collapse of the USSR. In these cases, the position on a timeline or a map remains the same. But these names are used in different times, i.e. they are not alternative names as mentioned above. In such cases we use the present-day name as official name in the `val` feature, adding a feature `hist`oric. Also the coordinates mentioned will be those associated with the current name. Note that in such cases temporal information is also relevant.

So `disambiguation` of esp. obsolete names is done with respect to the `intended audience` of a text. For an item in a Dutch newspaper of March 26, 1938 this will be Dutch readers living in those days. But `localization` (on a map or time line) is done for the `present-day user` of the annotation, using contemporary maps and calendars.

Things become more complicated when there are no present-day corresponding entities. This is for example the case for *Czechoslavakia* and the *USSR*, which fell apart in two or more parts. Mentioning these parts in the `val`, however, will do the trick: we can thus locate *Czechoslavakia* on a present-day map. Locating entities like *East Germany* is more difficult. In the `val` feature we list the present-day 'Bundesländer' (States) which used to belong to the *Deutsche Demokratische Republik*, whereas a `hist` feature may be used to mention the 'old' name *Deutsche Demokratische Republik* in the tag itself.

Old names like *Mesopotamia* can only vaguely be located on contemporary maps and time line: *Mesopotamia* was an area in the surroundings of present-day Iraq, and existed roughly between 3500 B.C.and 500 B.C.

Also temporal entities may be rather complex. The Russian *October Revolution* (1917) occured in October, according to the Julian calendar which was in use in Russia in those days. But in November, according to the Gregorian calendar used in most European countries. In the `temp` tag associated with *October Revolution* we will mention <temp val="1917-11-07"/ > (i.e. using the Gregorian calendar), but the name will remain *October Revolution*.[24]

Entities like *Chinese New Year*, which may fall within a specific range of dates, will be treated like Easter, which shares this characteristic and may fall between March 22 and April 25th:

```
<temp type="cal" val="XXXX-(03-22..04-25)"[25]
amount="1" unit="D"/ >.
```

7. Conclusions and further work

MiniSTEx tries to make the spatiotemporal knowledge humans are supposed to have available for the machine as well, in order to enhance reasoning.

Intended audience and background knowledge are central notions, when it comes to disambiguation and location. Although only part of the system is already implemented (the geospatial component), the results are encouraging.

In the near future, the temporal part of the system will be implemented as well. We will also experiment with an implementation based on chunked strings of input, instead of on a full syntactic analysis,

8. References

R. Dale and E. Reiter. 1996. The role of the Gricean maxims in the generation of referring expressions. In B. Di Eugenio and N. Green, editors, *AAAI Spring Symposium on Computational Implicature: Computational Approaches to Interpreting and Generating Conversational Implicature*, pages 16–20, Menlo Park, California.

J. Ding, L. Gravano, and M. Shivakumar. 2000. Computing geographical scopes of web resources. In *26th International Conference on Very Large Databases, VLDB 2000*, Cairo, Egypt, September 10–14.

J. Leidner. 2006. Toponym Resolution: A First Large-Scale Comparative Evaluation. Technical report, School of Informatics, University of Edinburgh, July.

J. Leidner. 2007. *Toponym Resolution in Text*. Ph.D. thesis, University of Edinburgh.

MITRE, 2007. *SpatialML: Annotation Scheme for Marking Spatial Expressions in Natural Language*, October 1.

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. In *Proceedings of LREC 2008*, Marrakech, Morocco.

R. Sauri, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky, 2006. *TimeML Annotation Guidelines, version 1.2.1*.

I. Schuurman. 2007a. MiniSTEx Protocol, version 0.2. KULeuven 2007, March.

I. Schuurman. 2007b. Spatiotemporal Annotation on Top of an Existing Treebank. In K. De Smedt, J. Hajic, and S. Kuebler, editors, *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, pages 151–162, Bergen, Norway.

I. Schuurman. 2007c. Which New York, which Monday? The role of background knowledge and intended audience in automatic disambiguation of spatiotemporal expressions. In *Proceedings of CLIN 17*.

R. Volz, J. Kleb, and W. Mueller. 2007. Towards ontology-based disambiguation of geographical identifiers. In *WWW2007*, Banff, Canada, May 8-12.

---

[24]An expression like *October Revolution* will be annotated with the complex `geotemp` tag, containing full `geo` and `temp` tags.

[25].. stands for 'one or more units out of a range'.