5

Which New York, which Monday?

The role of background knowledge and intended audience in automatic disambiguation of spatiotemporal expressions

Ineke Schuurman
Katholieke Universiteit Leuven

Abstract

The aim of MiniSTEx, a system for automatic spatiotemporal annotation, is to locate eventualities on a time-axis and to disambiguate geospatial information in such a way that geospatial entities can be located on a map. Therefore all kinds of spatiotemporal (geospatial, temporal and geotemporal) expressions are disambiguated. In doing so, the concepts of "background knowledge" and "intended audience", together with the Gricean maxims, play an important role, especially when dealing with indexicals. The system relies on a database containing all kinds of spatiotemporal expressions. At the moment MiniSTEx is used for both Dutch and English texts.

5.1 Introduction

MiniSTEx is a first version of a larger annotation system for spatiotemporal phenomena under construction.[1] It has to handle all types of Dutch texts (both fiction

[1] I would like to thank my colleagues, and especially Vincent Vandeghinste, for all discussions.

Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands
Edited by: Peter Dirix, Ineke Schuurman, Vincent Vandeghinste, and Frank Van Eynde.

and non-fiction, i.e. novels, newspapers, web pages, pamphlets, etc.). The general spatial part of the system still needs to be developed in more detail in the future. The geospatial part, however, is already handled. The same holds for the temporal and geotemporal parts.

The aim of this annotation scheme is to identify spatiotemporal expressions, and to normalize and disambiguate them in order to facilitate reasoning. The approach is meant to be used in applications like (multi-lingual) information retrieval, question answering, and multidocument summarization.

MiniSTEx reflects the state of the art in geospatial and temporal annotation. With respect to the latter, TimeML (Sauri et al. 2006) and TIDES (Ferro et al. 2005) come to mind. Geospatial annotation as such is far less widespread and standardized. However, the subtask of disambiguation is also a subject in geographic information extraction. Some approaches in this field can be found in Ding et al. (2000), Leidner (2006), and Volz et al. (2007).[2]

Typical for MiniSTEx is that it handles a) both geospatial and temporal expressions, and b) also geotemporal expressions, i.e. expressions associated with a combination of geospatial and temporal properties. The system was designed to be used in circumstances in which the background of the texts is known, i.e. not in the first place for web pages and the like. In the annotation process a large spatiotemporal database plays a central role.

And pragmatics, especially when using both the background of a text and its intended audience, plays an important role in deciding which database entry is to be associated with a particular spatiotemporal expression in a text when the tokens as such can refer to several concepts: “which New York, which Monday?”

In the STEVIN-project[3] SONAR (2007-2010) a syntactically analyzed subcorpus[4] of Dutch is being enriched with four types of semantic annotation: a) named entity identification and classification, b) coreference resolution, c) semantic roles and d) spatiotemporal relations (the latter using MiniSTEx). Within SONAR at least part of the expressions to be identified and disambiguated (the so-called timexes) by MiniSTEx are already marked as such.

MiniSTEx is also used in the SBO-project AMASS++ (Advanced Multimedia Alignment and Structured Summarization), funded by IWT. In AMASS++ (2007-2011) we use it both for Dutch and English.

In this paper we will pay special attention to the strategy used to select referents for contextually dependent, non-deictic expressions.

[2] Note that we annotate more phenomena than covered in these papers, cf. Schuurman (2007).

[3] `http://stevin-tst.org`.

[4] SONAR is a 500-million-word reference corpus of contemporary written Dutch. A 1 million subcorpus will be semantically annotated.

5.2 Which New York, which Monday?

There are over 50 Mondays in a year, and, according to Wikipedia[5] (English version, March 2007), 8 geospatial entities called New York, cf. table 5.1.

Table 5.1: Which New York?

New York	U.S. state (population
New York	city in the above state
New York	county, generally referred to as Manhattan
New York	metropolitan area
New York	Lincolnshire
New York	Tyne and Wear
New York	Missouri
New York	Texas

Even as a geospatial expression *New York*[6] is ambiguous. Even more ambiguous than shown in table 5.1: in GeoNet Names Server[7] (a gazetteer) there are already 12 hits outside the US. And in the Getty Thesaurus of Geographic Names[8] 15 instances inside the US are mentioned. In our database-driven approach this means that an expression like *New York* might get several entries in the spatiotemporal database (up to 27+).

So which one to choose when annotating a particular text?

One of the basic assumptions of MiniSTEx is that in order to facilitate reasoning quantification of information is essential. Therefore, in contrast with common practice, cf. Sauri et al. (2006), expressions like *winter* are also normalized in terms of the months people associate with *winter*, for example *december, january* and *february*: “XXXX-12/02”[9] (instead of “XXXX-WI”).[10]

Note that in Schuurman (2007) some spatiotemporal expressions may have various, in se correct values, depending for example on the hemisphere (*winter*), or on religion and/or tradition (*Christmas*). Others are often used in a sloppy way, like *winter, week* or *Christmas*[11]. Reliability features (`noise` and `soft`) are added to indicate such behaviour, cf. Schuurman and Monachesi (2006), and especially Schuurman (2007), when it is not clear enough which referent is meant.

People do succeed in detecting the correct referent from context. MiniSTEx is able to do so as well, i.e. to identify spatiotemporal expressions, and to disambiguate them.

Before we describe the MiniSTEx approach, let us have a look at the kind of spatiotemporal data we typically are confronted with when annotating (or reading)

[5] `http://en.wikipedia.org`.

[6] There are also lots and lots of hotels, ships, songs, albums, etc with this name. Within the Stevin-programme named entitiy recognition is to filter out these.

[7] `http://earth-info.nga.mil/gns/html/index.html`.

[8] `http://www.getty.edu/research/conducting_research/vocabularies/tgn/`.

[9] In combination with a ’reliability’ feature when necessary, cf. Schuurman (2007).

[10] In which ‘WI’ is an abbreviation of *winter*.

[11] For example: when you are going somewhere for Christmas, is it just the 25th of December, or does it include the 26th as well?

a text in order to detect anchors enabling the location of eventualities (events, states, processes) on a time-axis and/or on a virtual map.

5.3 Types of times and places in need of disambiguation

Whereas the expressions in section 5.3.1 contain all information needed to interpret them themselves, this is not the case for those in section 5.3.2.

5.3.1 Independent temporal, resp. geospatial expressions

Examples of independent temporal, resp. geospatial expressions are expressions like those in a) *March 1st, 2003; Washington D.C.; the Netherlands* and b) *the first Tuesday in May 2000, the capital of Sweden*.

Of the expressions mentioned only those in a) are really easy to describe formally. They (or their constituting elements, as for *March 1st, 2003*) are contained as such in the database, cf. table 5.5.[12]

In expressions like *the first Tuesday in May 2000* or *the capital of Sweden*[13] the constituting elements need to be solved before a specific date or town can be associated with them. For *the first Tuesday in May 2000*, etc. this means that the constituting elements are contained in the database as forms, to be combined and solved when applied: “2000-05-02”.

The common characteristic of all these expressions is that there is just one possible solution, even when part of the construction can refer to several temporal or geospatial entities.

5.3.2 Indexicals

Indexicals are context dependent expressions, usually deictic ones, like *today, this week, now; here, in this country*.

But note that also the meaning of *March 1st, Monday, Easter 2003, winter 2002* and *New York, Dallas* and *Washington* depends on the broader context or even other information coming with the text under consideration (metadata). In the first two expressions the year is lacking, *Easter* comes on another date in the orthodox church, *winter* depends on the hemisphere, *New York* can be the city or the state, etc.

Such indexicals need to be solved, taking the context into consideration. This not only is necessary for deictic expressions, but is also explicitly necessary for non-deictic expressions like *Monday* or *New York*: which *Monday*, which *New York*?

It are expressions like these non-deictic ones that are the subject of this paper.

[12] The database as presented here is a simplified one.

[13] In order to solve this construction, we need an additional (optional) feature in the `geo`-tag of *Stockholm*, expressing that it is a capital.

5.4 Indexicals and the interpretation thereof

The problem of how to annotate *Monday*, *New York*, and the like mainly concerns the interpretation of both temporal and geospatial indexicals, cf. section 5.3.2.

Table 5.2: Which one to choose?

Groningen	province or town in the Netherlands
Den Haag,'s Gravenhage	several names for the same town
Vecht	2 rivers in the Netherlands
Rijn	same river in several countries
Luxemburg	country, town in that country, or province in Belgium
Haren	2 villages in the Netherlands, one in Belgium
Kerst (Christmas)	on different dates depending on religion
vaderdag (father's day)	many dates possible, depending on country/region
winter	different months, depending on hemisphere; different dates (meteorological vs astronomical winter)
Koninginnedag	April 30 since 1949; August 31st from 1890 till 1949
november revolutie (November revolution)	same as October revolution
namiddag (afternoon)	different periods of time in the Netherlands and Belgium
Rio de Janeiro	town, region or Earth Summit[14]

A look at table 5.2 shows us that there is a variety of cases to disambiguate. The examples all show different instances of what in (geographical) information extraction is called[15]

1. *multi-referent ambiguity* (or homonymy): two or more concepts share the same name (*Groningen, Haren, hofstad, vaderdag, winter*)

2. *name-variant ambiguity* (or synonymy): the same concept comes with several names (*Den Haag – 's Gravenhage, november revolution – october revolution*[16])

Whereas we are not aware of attempts to solve problems like these as far as temporal concepts are concerned, there are a few attempts with respect to geospatial concepts in the field of information extraction, cf. section 5.1.

5.4.1 Other (geospatial) approaches

In Volz et al. (2007) a novel approach is presented to disambiguate geographic names based on an ontology. Their ontology contains data from publicly available

[14] This is an example of a geotemporal expression. Whereas geospatial expressions are in fact a subset of spatial expressions, geotemporal expressions are expressions associated with both temporal and geospatial properties. These are typically expressions concerning larger events like (*fall of the Berlin wall, Earth Summit, 9/11, . . .*), the temporal and/or geotemporal details of some of these may even be considered common knowledge (*World War II*). Albeit sometimes there is some uncertainty whether it started in 1939 or 1940, or about the exact end date, the intended audience knows where to situate World War II on a time axis.

[15] We will bypass the third ambiguity: geoname (short for geographical name) – non-geoname.

[16] Depending on the calendar used.

gazetteers (like GeoNet Names Server) and common world/linguistic knowledge obtained from WordNet[17] and EuroWordnet.[18] When they have spotted all candidates for geonames, they first try to narrow down the selection by looking in a window of 2 consecutive geographical terms whether there are clues to be found (like *Paris, France* vs. *Paris, Texas*), in a second step a window of 11 consecutive terms (t_i(+|-)5) is taken into consideration to find instances of the same geographic feature class (like *country, populated place*).

The remaining candidates are ranked according to the weights attached to the concepts in the ontology. A country gets the weight +3000, a populated place the same weight (+3000), but in this case the number of inhabitants (divided by 1000) is added. This would mean that, when no further information is availble via the first steps, the city of Luxemburg will be ranked higer than the country with the same name, and that Lancaster (California) will be ranked higer than Lancaster (UK).

Ding et al. (2000) are closest to our approach in that they try to determine the intended audience of a webpage, i.e. its geographical scope. They use two methods to determine this scope: 1) look what the geographic location is of hosts referring to the website under consideration, and 2) look what the scope is of all geographical places mentioned in this website. This will give a clue where the intended audience is located.

5.4.2 The pragmatic MiniSTEx approach

In order to develop a system dealing with disambiguation of temporal and geospatial data we asked ourselves “What makes a reader understand the geospatial and temporal data contained in a specific text?” as such characteristics may be useful for our design as well.

The vital property of a text seems having an intended audience: a medical text written for British GPs is likely not to be fully understandable for either aerospace engineers, teachers or linguists. Nor for Belgian GPs. And a text written for people living in Amsterdam, be it a newspaper or a bulletin by the city council, may not be understandable for people living in Brussels or Rotterdam when refering to local information.

This is the case because every speaker (author) will apply conversational maxims as formulated by Grice (1975), often paraphrased as “Don’t say too much and don’t say too little.”, cf. Dale and Reiter (1996), without as much as thinking.

These maxims are

1. Maxim of Quantity:

   (a) Make your contribution as informative as required;

   (b) Do not make your contribution more informative than is required.

2. Maxim of Relation (or Relevance):

[17] `http://wordnet.princeton.edu/w3wn.html`.

[18] `http://www.illc.uva.nl/EuroWordNet/`.

    (a) Be relevant.

3. Maxim of Manner:

    (a) Be perspicious:

        i. avoid obscurity of expression,
        ii. avoid ambiguity,
        iii. avoid unnecessary wordiness,
        iv. be orderly.

4. Maxim of Quality:

    (a) Do not say what you believe to be false;

    (b) Do not say that for which you lack evidence.

In MiniSTEx, we assume that a text always provides the (intended) reader 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, a system can neither be blamed for failing. MiniSTEx handles texts by using the background and world knowledge the intended audience is supposed to have.
Therefore problems we are faced with are:

(A) Determination of the intended audience of a text

(B) Determination of the corresponding spatiotemporal background knowledge

(C) Exploitation of this background knowledge

5.5 Determination of intended audience and spatiotemporal background knowledge

As far as problem (A) is concerned, note that our approach is not designed to primarily deal wit web pages, but rather with digital archives (broadcasting companies, news agencies), corpora and the like. Of the latter kind of resources the background is more often known. This is very important as it helps us a lot in determining both (A) the intended audience and (B) the spatiotemporal background knowledge this audience may be supposed to possess. So, unlike Ding et al. (2000) working with web pages in English, we do not solely rely on the distribution of web links in determining the intended audience. A first clue is provided by the language used: a text written in Dutch is in all probability meant for Dutch and/or Flemish readers, a text in Hebrew for Israelis or Jews around the world. For texts in English, the intended audience is more difficult to discover as these are either meant for a British (or an American, Australian, Canadian,…) audience, i.e. the text has a national scope, or for “the rest of the world” (global scope). But, especially for the smaller languages, data with respect to the intended audience can thus be derived even when details with respect to the source of the text are unknown. However, for known resources many more details are available, making use of the spatiotemporal data associated with the title (like *De Morgen*, *Daily Telegraph*, *Boston Globe*, *www.vlaanderen.be* etc)., cf. table 5.3.[19]

Table 5.3: Background-doc

concept	dbid	status	geo	trad	cal	lang	scope
De Morgen	220000	newspaper	Brussel			Dutch	national
De Telegraaf	220003	newspaper	Amsterdam				national
Ref. Dagblad	220009	newspaper	Apeldoorn	orth-ref			
Vl.overheid	230000	web	Brussel			Dutch	regional
Vl.overheid	230000	web	Brussel			English	global

Other information relevant for determining the intended audience are *tradition* (Christian, Islamic, Jewish, Eastern Orthodox, ...), and *calendar*: (Gregorian, Hebrew, Hindu, ...). Nowadays the Gregorian calendar is widely used in Israel, but the Hebrew calendar can also still be used (and is in fact used in a religious or cultural context). And the November Revolution in Russia (1917-11-07 according to the Julian Calendar used in Russia at that moment) is known in the western world as the October Revolution (Gregorian calendar: 1917-10-25). The intended audience of a Jewish newspaper or an older Russian text is supposed to be familiar with such traditions and calendars.

Table 5.4: Background-geo

concept	dbid	status	trad	cal	hem	UTC[20]	lang	partof	division
Spanje	109	cntry	chr	Greg	north	+1	ES	EU	2=region, 3=province
Nederland	146	cntry	chr	Greg	north	+1	DU	EU	2=–, 3=province
België	137	cntry	chr	Greg	north	+1	DU, FR, GE	EU	2=region, 3=province
VS	199	cntry	chr	Greg	north	-(5/10)	EN, ES	NA	2=state, 3=county
Vlaanderen	102	region					DU	BE	

The MiniSTEx database consists of more tables than presented in this section, cf. the tables in section 5.5.1. Those tables provide the data to connect the concepts in these background tables: in table 5.3 the geo-column refers to geospatial entities. Via table 5.5 these entities can be linked with entities in table 5.4. This table defines the spatiotemporal backgroundknowledge associated with a geospatial entity, unless it is superseded by information in table 5.3 itself. These columns in table 5.3 are only filled out in case they contain information that is to overrule the general information. So, *Reformatorisch Dagblad* is said to belong to the

[19] For convenience of the reader most tables as they are presented here contain the concepts. This is only for matter of presentation. In reality the only column all tables contain is the one with the `dbid`. The real tables also contain more columns, i.e. more types of data. And there are more tables.

[20] Coordinated Universal Time.

orthodox-reformatoric tradition instead of the more general christian tradition. For *De Morgen* and *De Telegraaf* the values for `geo` and `trad` are those of *Brussel* and *Amsterdam* respectively. For *De Telegraaf* `lang` is also that of *Amsterdam*, whereas for *De Morgen* the values for *Brussel* are overruled by the statement that only *Dutch* is used.

In MiniSTEx the spatiotemporal background knowledge the intended audience is supposed to have is contained in a series of tables.

5.5.1 The design of the MiniSTEx spatiotemporal database

As might be expected from the previous sections, the MiniSTEx database is meant to mimick the spatiotemporal knowledge of an intended audience. It is not the case that a new database is built for every new audience (one for *De Morgen*, another one for texts by the *Vlaamse overheid*, still another one for *Reformatorisch Dagblad*, etc.). This is not necessary, although parts of the database, like `ranking`, will need to be adapted for other 'supertypes' of intended audience (other countries etc). This issue will be researched in AMASS++.

In the end the Dutch database, consisting of a series of tables. will contain lots of temporal and (geo)spatial data with respect to the Dutch language, the Netherlands and Belgium, but far less with respect to, say French Guyana, Peru and Macedonia, the jewish calendar and the orthodox culture. Of these it will contain only those data relevant for a Flemish/Dutch audience. It may, for example, only contain two instantiations of New York (the state and the metropole).[21] This makes our approach a pragmatic one.

The central table in our database, cf. table 5.5, contains the concepts, their `dbid` and the `tag` associated with them, together with their `background`, `rank` and `parts`.

The `background` of a concept refers to specific conditions associated with it. *Thanksgiving* is celebrated both in the USA and in Canada, but on different dates. Apart from such geospatial conditions, references may be made to `tradition`, `calendar`, `hemisphere`, `language` (this one albeit rather seldom), …[22] It might come as a surprise that language doesn't play a more important role. But it turns out that the role of the country, or the region, is by far more important. The case of *vaderdag* is illustrative in this respect. At least three values for *vaderdag* are valid in the Dutch-speaking regions, cf. table 5.5. But when an item on, say, the Antwerp *vaderdag* is translated into English the term used will become *Father's day*, although the date it refers to is still to be the Antwerp one, not the UK one! The geographical background is of importance, not the language used.

The ranking indicates that, when *Thanksgiving* is mentioned in a Dutch or Flemish context without further details, it is likely to refer to the American instantiation (see below).

As alluded to above, our database for Dutch contains many (corrected) data relevant for the Netherlands and Belgium, based on gazetteers, Wikipedia,

[21] Others to be added when need arises.

[22] Cf. below, the paragraph on background.

(Euro)WordNet, etc. New data are added constantly. For other countries it contains only those data we consider relevant (like all continents, all countries, main cities in the neighbouring countries, and the US, main rivers etc.), based on the same kind of resources, plus some others, like The World Factbook[23]. More will be added when necessary on basis of the texts handled by the system. Therefore it is likely that for *New York*, cf. table 5.1, only the top two will ever be contained in it.[24] This means that names that in se could be ambiguous according to a gazetteer or Wikipedia can be unambiguous in our (Dutch) database.

A second table, cf. table 5.6, contains the name variants of the concepts contained in table 5.5, like synonyms. But still only in Dutch. Here again we use ranking to indicate the most likely referent(s).

There also is a, rather small, table with language-sensitive concepts, cf. table 5.7. Above we have explained that in general all and every of the background factors is of greater importance than the language[32]. There are just a few exceptions, in which a language only allows one value to be associated with a concept, while in other languages these concepts are associated with other values. An example that comes to mind is *avond – evening* vs *nacht – night*.

Table 5.7: Language-sensitive concepts

concept	dbid	language	tag
avond (evening)	1302	Dutch	<temp type=”cal” val=”T18/24”>
nacht (night)	1303	Dutch	<temp type=”cal” val=”T22/06”>
evening	1308	English	<temp type=”cal” val=”T18/21”>
night	1309	English	<temp type=”cal” val=”T21/06”>

Although the values given for many concepts may vary to some extent from person to person, from region to region, or from season to season, the ranges are relatively small, i.e. it is a matter of `noise`, not from a completely different value for a different, albeit related, concept. But in case of *avond – evening* vs *nacht – night* the Dutch and British concepts are different ones (reflected by the `dbid` they get).

The first step in selecting referents is to determine all non-ambiguous expressions. On basis of these the value of the remaining expressions is calculated in the next steps, keeping in mind the background of the text, cf. tables 5.3 and 5.4, and the type of the surrounding names: when *Kerst* (Christmas) appears in a text with a background in the christian tradition, it will be solved as referring to the

[23] https://www.cia.gov/library/publications/the-world-factbook/geos/be.html.

[24] *Manhattan*, for example, is unlikely to be referred to as *New York*, although it will be linked.

[25] The ’::’-sign is only used in geospatial entities. A::B indicates that B is part of A.

[26] ‘greg’ refers to the Gregorian calendar

[27] ‘form’ is used instead of ’val’ when variables are involved (in this case for the year, indicated by XXXX).

[28] A ‘|’ is used to indicate a non-exclusive ‘*or*’, the brackets indicate the scope.

[29] ‘X..Y’ indicates one or more elements out of a range X till Y.

[30] ‘X/Y’ means the whole range X up to and including Y

[31] An example of a geotemporal concept, thus including other concepts.

[32] Although the language is important in determining the intended audience.

Table 5.5: Concepts

concept	dbid	background	tag	rank	parts
Spanje	109	EU[25]	<geo type=”country” val=”EU::Spanje”/>		
Brussel	130	BE::BR	<geo type=”place” val=”EU::BE::BR::-::Brussel” />		
Den Haag	135	NL::ZH	<geo type=”place” val=”EU::NL::-::ZH::Den Haag” />		
Apeldoorn	145	NL::GE	<geo type=”place” val=”EU::NL::-::GE::Apeldoorn” />		
Haren	142	BE::BR	<geo type=”place” val=”EU::BE::BR::-::Haren />	2	
Haren	143	NL::GR	<geo type=”place” val=”EU::NL::-::GR::Haren />	1	
Haren	144	NL::NB	<geo type=”place” val=”EU::NL::-::NB::Haren />	3	
augustus	10057	greg[26]	<temp type=”cal” form=”XXXX-08”[27]/>		
vaderdag	1500	EU::(NL|UK|FR)[28]	<temp type=”cal” form=”XXXX-06-D07,15..21”[29]/>		
vaderdag	1501	EU::BE	<temp type=”cal” form=”XXXX-06-D07,08..14” />		
vaderdag	1502	BE::AN	<temp type=”cal” form=”XXXX-03-19” />		
St. Jozef	1550	chr	<temp type=”cal” form=”XXXX-03-19” />		
Thanksgiving	210074	NA::VS	<temp type=”cal” form=”XXXX-11-D04,22..28” />	1	
Thanksgiving	210075	NA::CA	<temp type=”cal” form=”XXXX-10-D01,08..14” />	2	
avond	1302	DU	<temp type=”clock” form=”T18/24” />		
nacht	1303	DU	<temp type=”clock” form=”T22/06” />		
middag	1291	EU::NL	<temp type=”clock” val=”T12/18”[30]		
namiddag	1292	EU::NL	<temp type=”clock” val=”T16/18”		
namiddag	1293	EU::BE	<temp type=”clock” val=”T12/18”		
Kerst	1310	chr	<temp type=”cal” form=”XXXX-12-25” />		
Kerst	1311	orth	<temp type=”cal” form=”XXXX–01-07” />		
winter	100562	north	<temp type=”cal” form=”XXXX-12/02” />		
Rio de Janeiro	101	BR::RJ	<geo type=”place” val=”SA::BR::RJ::-::Rio de Janeiro” />	1	
Rio de Janeiro	141	SA::BR	<geo type=”region” val=”SA::BR::Rio de Janeiro” />	2	
UNCED[31]	500010	UN|conf	<stex> <temp type=”cal” val=”1992-06-3/14” /> </stex>		101

Table 5.6: Name-variants of concepts

name-variant	dbid	concept	rank
’s Gravenhage	135	Den Haag	
hofstad	135	Den Haag	1
hofstad	145	Apeldoorn	2
oogstmaand	10057	augustus	
Rio-conferentie	500010	UNCED	
Rio	500010	UNCED	
Rio de Janeiro	500010	UNCED	
VN-conferentie inzake ontwikkeling en milieu	500010	UNCED	
wereldmilieu- en ontwikkelingsconferentie	500010	UNCED	

25th of December, unless it refers to *Kerst* in for example Russia. Russia comes with an orthodox background, and therefore *Kerst* will be solved as occuring in January, which is to be indicated. *Haren* will be associated with the village in the Brussels Capital Region when mentioned in an item in De Morgen (unless stated otherwise) because of its background.

5.6 The disambiguation steps

One could say that according to the Gricean maxims the intended audience of a text, cf. problem (A), in fact determines the way the content of a text is articulated. It is the intended audience, together with its (spatiotemporal) background, that makes an author mention things explicitly, or leave them out.
For example, in Belgium everybody knows that the official languages of the country are Dutch, French and German; or that Leuven is a town in Flanders, one of the three regions in Belgium. This is not mentioned in a text for, say, a Flemish audience. Indeed, a text becomes almost unreadable when it contains such unnecessary details. But in a British newspaper, one will have to mention such Belgian details explicitly. Another example: for a Flemish audience it is obvious that *Sinterklaas*[33] is celebrated on the 6th of december, whereas in the Netherlands the 5th will be associated with it. So, when in a Dutch newspaper an item would occur on the celebration of *Sinterklaas* in Belgium, the date will be mentioned explicitly. Otherwise, the intended Dutch readers will assume it is the 5th, as they are used to.

As a consequence, the intended reader by default prefers one reading over another one as a consequence of his spatiotemporal background knowledge, and the expectations evoked by it. In the same way the reader expects the referent of a geospatial expression to be the one that is most relevant for him (for example the most nearby *Haren* or the most well-known *New York*), he also expects temporal referents to be as relevant as possible. A plain reference to *Monday* is taken to

[33] The name day of Saint Nicholas, patron saint of, among others, children.

refer to last Monday or next Monday[34]; a reference to Christmas to the dates he himself is used to celebrate it. When another instantiation is desired, this should have been made clear.

The last refuge for both human reader and MiniSTEx is ranking: when all other steps fail one has to look at the importance of a specific referent for a intended audience. Clues may be nearness, importance, size, ...

Unlike Volz et al. (2007), we do not rank towns always higher than countries, or countries always higher than provinces, or a larger town higher than a smaller one. Understandably, this is the only way to attach ranks automatically for all geographic names in a gazetteer.

As we use just a selection, adding new names one by one, we can afford to attach ranks another way.

When the intended audience of texts published in Western Europe is confronted with a geographic name *Dover*, they are inclined to associate it with the town of that name in the United Kingdom, and not with the capital of the state of Delaware, USA, although the latter is the larger one. Also, the country of *Luxemburg* has a better ranking than its capital with the same name, and the same with respect to the Walloon province with that name, even in a Belgian text. A rule of thumb is that referents in neighbouring countries are preferred over referents in further away countries. Also the relations between the countries involved may play a role (export relations, former colonies, etc.).

The highest rank available is `1`, and several concepts sharing the same name may occasionally have the same (low) rank. In such a case two or more referents will be provided as value. The same holds mutatis mutandis for temporal referents.

In MiniSTEx, the steps taken for disambiguation are roughly the following (after each step elements are disambiguated (if possible), unless the results are contradictionary, in which case the next step is applied):

1. identification of unambiguous spatiotemporal elements;

2. identification of all general spatiotemporal (broad sense) expressions (*stad* (town), *land* (country), *noordelijk halfrond* (northern hemisphere), *christelijk* (christian), *burgemeester* (mayor), . . . ;

3. confrontation of these with ambiguous referents, first at the level of the constituent (*de stad Antwerpen*) (the city of Antwerp)), later at that of the sentence and the paragraph;

4. selection of readings coming with the same `background`, cf. section 5.6.2;

5. identification of the `division` value, cf. table 5.4, of the referents solved unambiguously;

6. select the reading with the best rank.

[34] The choice between 'last' or 'next' is guided by the tense of the verb. But note that a plain *Monday* never will be taken to refer to a Monday several weeks ago or in the future. It has to be `last` Monday.

5.6.1 One sense per text

When looking at a set of names (in a list, or in a window of *x* terms) the types and values of the unambiguous names will steer the interpretation of the ambiguous ones:

(1) {Antwerpen,Leuven,Utrecht,Groningen}

(2) {Antwerpen,Vlaams-Brabant,Utrecht,Groningen}

Just by the fact that *Leuven* is a town, whereas *Vlaams-Brabant* is a province, the other three names in the set are towns, resp. provinces as well, although they are not in the same country.

5.6.2 Additional world knowledge

The intended audience is expected to have some additional world knowledge.[35] Not only in order to be able to handle concepts like *World War II* when these occur without further details, but also for disambiguation purposes. The reader should for example be able to deduce the correct referent for *Antwerpen* (town or province) in expressions like the following:

(3) de burgemeester van Antwerpen
the mayor of Antwerp

(4) de gouverneur van Antwerpen
the governor of Antwerp

In (3) the town of *Antwerpen* is meant, as a province does not have a mayor, whereas a town does, and in (4) just the other way round.

5.7 Conclusion

For automatic spatiotemporal annotation, and especially disambiguation, of a text it turns out to be important to know the intended audience of that specific text. One needs to know *when, where* and *in which context* (which newspaper, website, . . .) a text appeared. That way, the spatiotemporal knowledge a reader (and a system) needs in order to understand the text can be derived.

Of course, one does not always have all these details. But, except for English and other global languages (like French and Spanish), the language used already gives a clue. Furthermore, the geographical scope as used by Ding et al. (2000) will provide some details, also for English texts. But in case more data are available, one should consider using them.

We just started adding data for English, as the system was originally designed for Dutch[36] only. Up till now we are using one large database for both languages. Whether it is to be split is still researched.

[35] Relevant for the specific part of the world under consideration.

[36] Both the variants as spoken in the Netherlands and Flanders.

In the future general spatial concepts will be added, which are more complex than the geospatial ones.

References

Dale, R. and Reiter, E.(1996), The role of the Gricean maxims in the generation of referring expressions, *in* B. D. Eugenio and N. L. Green (eds), *Working Notes: AAAI Spring Symposium on Computational Implicature: Computational Approaches to Interpreting and Generating Conversational Implicature*, American Association for Artificial Intelligence, Menlo Park, California, pp. 16–20.

Ding, J., Gravano, L. and Shivakumar, N.(2000), Computing geographical scopes of web resources, *26th International Conference on Very Large Databases, VLDB 2000*, Cairo, Egypt.

Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson, G.(2005), *TIDES 2005 Standard for the Annotation of Temporal Expressions*.

Grice, H.(1975), Logic and conversation, *in* P. Cole and J. Morgan (eds), *Speech Acts*, Vol. 3 of *Syntax and Semantic*, Academic Press, New York, pp. 43–58.

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

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

Schuurman, I.(2007), *MiniSTEx Protocol, version 0.2*, Centre of Computational Linguistics, K.U.Leuven. KULeuven 2007.

Schuurman, I. and Monachesi, P.(2006), The contours of a semantic annotation scheme for Dutch, *Proceedings of CLIN 2005*.

Volz, R., Kleb, J. and Mueller, W.(2007), Towards ontology-based disambiguation of geographical identifiers, *WWW2007*, Banff, Canada.