# The Workshop Programme

## *Methodologies and Resources for Processing Spatial Language*

**31 May 2008**

9.00 Welcome & Introduction

9.10 - 9.50 Invited Talk:
The Long Road from Spatial Language to Geospatial Information, and the Even Longer Road Back: The Role of Ontological Heterogeneity
*John Bateman (University of Bremen)*

9.50 - 10.30 Placename Ambiguity Resolution
*Geoffrey Andogah, Gosse Bouma, John Nerbonne, Elwin Koster (University of Groningen)*

BREAK

11.00 - 11:40 Annotating Natural Language Geographic References
*Inderjeet Mani, Janet Hitzeman and Cheryl Clark (MITRE)*

11:40 - 12:20 Spatial Entities are Temporal Entities Too: The Case of Motion Verbs
*Nicholas Asher, Philippe Muller, and Mauro Gaio (IRIT/CNRS)*

12:20 - 12.40 Reusable Grammatical Resources for Spatial Language Processing (with **demo)**
*Robert J. Ross (University of Bremen)*

12.40 - 13.00 Realtime Path Descriptions Grounded with GPS Tracks: A Preliminary Report (with **demo)**
*Nate Blaylock and James Allen (IHMC/University of Florida)*

LUNCH

2:30 - 3:10 Building a Parallel Spatio-Temporal Data-Text Corpus for Summary Generation
*Ross Turner, Somayajulu Sripada, Ehud Reiter and Ian P Davy (University of Aberdeen and AMI)*

3:10 - 3:50 A Field Based Representation for Vague Areas Defined by Spatial Prepositions
*Mark M. Hall and Christopher B. Jones (Cardiff University)*

3:50 - 4:00 Concluding Remarks/Discussion

# Workshop Organisers

Graham Katz *(Georgetown)*
Inderjeet Mani *(MITRE)*
Thora Tenbrink *(Bremen)*

# Workshop Programme Committee

| | |
|---|---|
| Nicholas Asher | *IRIT/CNRS* |
| Janet Hitzeman | *MITRE* |
| Alexander Klippel | *Penn State* |
| Andras Kornai | *MetaCarta* |
| Jochen Leidner | *Edinburgh* |
| Amit Mukerjee | *IIT Kanpur* |
| James Pustejovsky | *Brandeis* |
| Ehud Reiter | *Aberdeen* |
| Frank Schilder | *Thomson* |
| Nicola Stokes | *Melbourne* |
| Andrea Tyler | *Georgetown* |
| Peter Viechnicki | *Board on Geographic Names* |
| Laure Vieu | *IRIT/CNRS* |
| Stephan Winter | *Melbourne* |

# Table of Contents

# Author Index

# The long road from spatial language to geospatial information, and the even longer road back: the role of ontological heterogeneity

**John A. Bateman**
Faculty of Linguistics and Literary Sciences & SFB/TR8 Spatial Cognition Research Center
University of Bremen, Bremen, Germany

In this talk I present from an ontological perspective some of our ongoing work on the situated interpretation of natural language concerning space. Since we are working on descriptions of space, scenes, navigation, way-finding assistance and tasks that need to be carried out in spatially rich environments, a broad range of spatial language and, more importantly, spatial language usage needs to be covered. Our long term goal is to construct systems which are capable of dealing with the full range of flexibility observed in natural langage concerning space rather than adopting artificial restrictions for the purposes of particular applications or tasks. This involves specifications of at least the following components:

- broad coverage analysis and generation components (not necessarily identical) that map between forms and semantic representations and which provide particular detail concerning meanings and forms concerned with space,
- a linguistic semantic organisation that spells out in detail the spatial commitment of linguistic utterances,
- a set of non-linguistic spatial accounts that provide the final targets of analysis and sources of generation.

Some aspects of our approach to the first of these components are addressed in more detail in the presentation by Ross in this workshop; I will focus therefore particularly on the second and third components, and some of the possible relations between them.

Our treatments have been particularly concerned with capturing spatial relations: that is, the semantics of utterances where distinct entities are brought into a spatial configuration in particular, more or less specifically specified relations holding between them. These relations can involve both static relations of relative location, distance, overlap, containment, etc. and dynamic relations where entities move with respect to each other. In this area there has been considerable work from non-linguistic perspectives, ranging from metric accounts operating in terms of precise measurement and geometry to qualitative accounts involving spatial qualitative calculi of various kinds (cf. Cohn & Hazariki 2001, Bateman & Farrar 2004). In addition, there are accounts of the various *kinds* of entities that can be play spatial roles: these also range from everyday objects that may take up the role of landmarks (‘behind the church’) to rich classifications of geographic entities, either types in their own right (‘mountain’, ‘lake’, ‘river’, . . . ), or individuals (‘Mount Everest’, ‘Marrakesh’, . . . ). Moreover, each of these exhibits interesting cross-cultural variability (cf. Mark et al. 2003, Mark & Turk 2003).

A reocurring problem that we have identified with respect to previous approaches (and even to much ongoing work) to the interpretation of spatial language is a marked underestimation of just what language is doing. When formalisation begins outside of linguistics, attempting to pin down non-linguistic characterisations of space, it is tempting to assume that language’s contribution to the problem will be relatively small. Here, it is salutory to note the experiences of Bennett and Agarwal concerning their own attempt to formalise the linguistic notion of 'place' starting from a non-linguistic characterisation:

> "When we began this work, we believed we could proceed directly to formulate a general logical theory of the concept of place. However, we soon found that the huge variety of different ways in which place enters language made it impossible to achieve a simple theory that covered all these modes. Thus we were driven to a detailed analysis of the many linguistic expressions of place concepts and their semantic content." (Bennett & Agarwal 2007)

Carrying out such a task demands a detailed linguistic analysis in its own right, and here we can of course draw on an already considerable literature

The conclusion that we draw from such analyses, which I will briefly review, is that it is necessary to cleanly separate the linguistic semantics of space from the non-linguistic, situation-specific interpretation of space. In particular, we consider *language itself* to contribute an ontology-like organisation, or construal, of the spatial world. This ontology-like organisation is adopted as an *additional layer* of ontological information the formalises the 'semantic commitments' entered into by any linguistic construction. The spatial configurations thus captured contain precisely the degree of formalisation required to explain the linguistic options taken up without overcommitting in terms of the physical or conceptual spatial situations that may be compatible with those commitments. The result is then what we may term a linguistically-motivated ontology specifically tailored to the requirements of spatial language and which, as a consequence, is also particularly well suited for natural language processing (cf. Bateman et al. 2008).

The principal theoretical motivation for assuming disjoint levels is that constraints from language (particularly linguistic expressions and grammar) and constraints from other levels of representation—in our case here, spatial semantics considered independently of language—often do not align. The semantics underlying linguistic usage tend to cross-cut and redefine distinctions that have been motivated solely in terms of space. I will illustrate this fact with respect to three perspectives on modelling spatial language: (a) the linguistic phenomena of spatial language use, (b) the formalisation of spatial language interpretation, and (c) the computational instantiation of processing schemes for natural language involving space. We will see that from all domains there is striking converging evidence that it is crucial to pull apart the relative contributions of spatial language, particularly spatial semantics, and domain or task characterisations of space.

Once this separation has been made, the next set of issues concerns how best specifications of linguistic semantics can be brought into suitable relations with non-linguistic specifications. To address this complex of problems, we are applying research that we are pursuing within formal ontological engineering that is committed to notions of ontological *structuring*, *modularity* and *heterogeneity* (cf. Bateman, Borgo, Lüttich, Masolo & Mossakowski 2007, Bateman, Tenbrink & Farrar 2007). In the talk, I will introduce and explain these notions and something of their formal background and go on to discuss how this can serve as a flexible bridge with existing and ongoing standardisation efforts for spatial annotation, spatial reasoning, and natural language components capable of dealing intelligently with spatial tasks.

## Acknowledgements

I gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center on Spatial Cognition (SFB/TR 8) for the work reported in this paper, as well as the substantial contributions of my co-workers in all the areas addressed.

## References

Bateman, J. A., Borgo, S., Lüttich, K., Masolo, C. & Mossakowski, T. (2007), 'Ontological Modularity and Spatial Diversity', *Spatial Cognition and Computation* **7**(1), 97–128.

Bateman, J. & Farrar, S. (2004), Spatial ontology baseline, SFB/TR8 internal report I1-[OntoSpace]: D2, Collaborative Research Center for Spatial Cognition, University of Bremen, Germany.

Bateman, J., Hois, J., Ross, R., Tenbrink, T. & Farrar, S. (2008), The Generalized Upper Model 3.0:

Documentation, SFB/TR8 internal report, Collaborative Research Center for Spatial Cognition, University of Bremen, Germany.

Bateman, J., Tenbrink, T. & Farrar, S. (2007), 'The Role of Conceptual and Linguistic Ontologies in Discourse', *Discourse Processes* **44**(3), 175–213. Special Issue on Dialogue Modelling: Computational and Empirical Approaches.

Bennett, B. & Agarwal, P. (2007), Semantic Categories Underlying the Meaning of 'Place', *in* S. Winter, M. Duckham, L. Kulik & B. Kuipers, eds, 'Spatial Information Theory, 8th International Conference, COSIT 2007, Melbourne, Australia, September 19-23, 2007, Proceedings', Springer, pp. 78–95.

Cohn, A. & Hazariki, S. (2001), 'Qualitative spatial representation and reasoning: an overview', *Fundamenta Informaticae* **43**, 2–32.

Mark, D., Kuhn, W., Smith, B. & Turk, A. (2003), Ontology, natural language, and information systems: implications of cross-linguistic studies of geographic terms, *in* M. Gould, R. Laurini & S. Coulondre, eds, 'AGILE 2003: 6th AGILE conference on geographic information science', Lyon, France, pp. 45–50.

Mark, D. & Turk, A. (2003), Landscape categories in Yindjibarndi: ontology, environment, and language, *in* W. Kuhn, M. Worboys & S. Timpf, eds, 'Spatial Information Theory: foundations of geographic information science', number 2825 *in* 'Lecture Notes in Computer Science', Springer-Verlag, pp. 31–49.

# Geographical Scope Resolution

**Geoffrey Andogah, Gosse Bouma, John Nerbonne, Elwin Koster**

Alfa Informatica, University of Groningen
Broerstraat 5, P.O. Box 72, 9700 AB Groningen, The Netherlands
{g.andogah,g.bouma,j.nerbonne,e.a.koster}@rug.nl

**Abstract**

It is common for placenames to reference other named entities (e.g., names of people, names of organizations, etc.) and to be used as vocabulary words (e.g., city of Split). Apart from reference ambiguity, placenames are faced with the problem of referent ambiguity (i.e., a placename referring to multiple places). Many places are also referred to by multiple names (e.g., Netherlands vs. Holland). In this paper we describe an approach to place ambiguity resolution in text, i.e., place reference resolution, resolution of a document's geographical scope and placename referent resolution. The approach is composed of three components: (1) geographical tagger, (2) geographical scope resolver and (3) placename referent resolver.

## 1. Introduction

Placenames are highly ambiguous as they reference other named entities (e.g., names of people, names of organizations, etc.) and are commonly used as language vocabulary words (e.g., city of Split). Apart from reference ambiguity, placenames are faced with the problem of referent ambiguity (i.e., a placename referring to multiple places). Many places are also referenced by multiple names (e.g., Netherlands vs. Holland).

Before proceeding further, a brief definition of some terminology is necessary:

**Place reference recognition and classification (PRRC):** The process of recognizing names in text and classifying them as place names as opposed to names of other entities.

**Place referent ambiguity resolution (PARR):** The process of assigning a place name identified in text to a single non-ambiguous place on the surface of the earth by means of a reference coordinate system such as longitudes and latitudes.

**Geographic scope resolution (GSR):** The process of assigning a geographical region or area to a document for which the document is geographically relevant.

We describe an approach to place ambiguity resolution in text consisting of three components: (1) a geographical tagger, (2) a geographical scope resolver, and (3) a placename referent resolver. The last two components were built inhouse while the first component is off-the-shelf software. Figure 1 shows the overall system architecture where the slanted boxes with dashed line boundaries are system outputs at various stages of processing.

Non-ambiguous geographical information (e.g., geographical scopes and placename referents) could improve the performance of standard information retrieval (IR) systems where the answer to the user's information need is geographically restricted (e.g., retrieving documents about "cities along river Nile") (Mandl et al., 2007). Placenames, geographic scopes (geo-scopes) and placename referents are used in query processing, document retrieval, document ranking and document visualization (Martins et al., 2006; Andogah and Bouma, 2007; Cardoso et al., 2007; Graupmann and Schenkel, 2006; Fu et al., 2005; Larson et al., 2006). The GSR approach reported in this paper exploits placename frequency of occurrence, geographical adjectives, place type (e.g., city), place importance (e.g., based-on population size and place type), and vertical (transitive parent/child) and horizontal (adjacency) relationships among places. On the other hand PRAR exploits geo-scopes assigned to documents, place type, place classification, place population and frequency information (e.g., counts of types of non-ambiguous places). Our GSR is implemented using a standard information retrieval (IR) library whilst our PRAR component is composed of simple heuristics. As mentioned before, the geographical tagger used is an off-the-shelf software[1] component pre-trained to mark place names, organization names and person names in text.

Our system is innovative in a few ways: (1) the GSR uses unresolved place names to resolve geographical scopes of

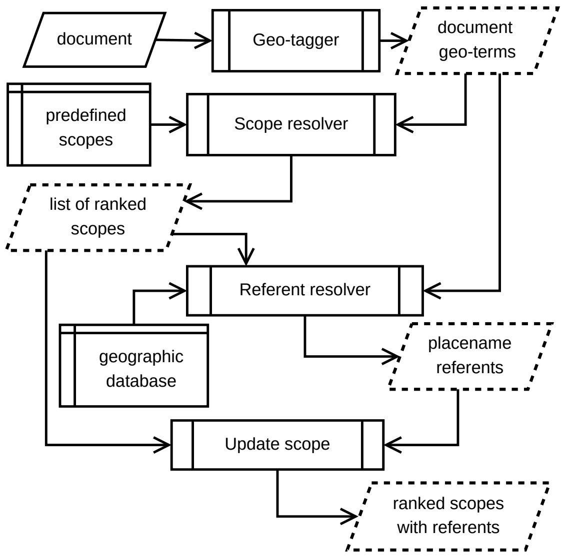


Figure 1: Placename ambiguity resolution system architecture.

[1] http://alias-i.com/lingpipe/

documents, (2) the GSR is implemented using a standard IR library, (3) the PRAR uses an elaborate range of geographical scopes assigned to a document as a basis to perform referent resolution and (4) the PRAR also makes extensive use of place types and classification to resolve among competing candidate places.

## 2. Geographic Scope Resolver

The geo-scope resolution approach discussed in this paper is based on Assumption 1.

**Assumption 1** *Places of the same type or under the same administrative jurisdiction or near/adjacent to each other are more likely to be mentioned in a given discourse. For example, a discourse mentioning The Netherlands is most likely to mention places of the type country (e.g., Spain, Uganda) or places under the jurisdiction of The Netherlands (e.g., Amsterdam, Rotterdam) or places adjacent to The Netherlands (e.g., Belgium, Germany).*

To implement the assumption, six groups of geo-scope are pre-defined at administrative (i.e., continent, country, province) and directional (i.e., at continent, country, province) levels. Province is used in a broader sense to mean first order administrative division of a country. The pre-defined geo-scopes are indexed and searched using the Apache Lucene IR library.

### 2.1. Apache Lucene

Lucene's default similarity measure is derived from the vector space model (VSM). The VSM is a classic document and query modeling technique in IR systems. In VSM both the document and query are viewed as vectors ( i.e., terms obtained from document and query texts with associated weights) in a multi-dimensional space (Lee et al., 1997).
The Lucene similarity score formula combines several factors to determine the document score for a query (Gospodnetic and Hatcher, 2005):

$$Sim(q,d) = \sum_{t\,in\,q} tf(t\,in\,d)\,.\,idf(t)\,.\,bst\,.\,lN(t.field\,in\,d) \quad (1)$$

where, $tf(t\,in\,d)$ is the term frequency factor for term $t$ in document $d$, $idf(t)$ is the inverse document frequency of term $t$, $bst$ is the field boost set during indexing and $lN(t.field\,in\,d)$ is the normalization value of a field given the number of terms in the field. In our implementation we leverage Lucene's capability to query on multiple fields and query term boosting.

### 2.2. Geographical knowledge

The Geonames.org[2] database is used as the basis of our geographical knowledge. It contains over eight million geographical names and consists of 6.5 million unique features including 2.2 million populated places and 1.8 million alternate names. All the features are categorized into one of nine feature classes and further subcategorized into one of 645 feature codes. We used features of the class administrative division (A) and populated place (P) to define geo-scopes.

| Feature class | No. features | Unique names |
|---|---|---|
| All classes | 6,603,579 | 4,230,969 |
| Class A & P | 2,564,814 | 1,640,422 |
| Class P | 2,393,808 | 1,565,458 |
| Class A | 171,006 | 144,684 |

Table 1: Geonames.org feature class A & P statistics.

| Name type | No. features | Unique names |
|---|---|---|
| Standard | 6,603,579 | 4,230,969 |
| Alternative (EN) | 1,237,759 | 1,735,528 |

Table 2: Geonames.org standard and alternate names statistics.

Tables 1 & 2 respectively show feature class and name statistics. Standard names are the feature names in the main Geonames.org database whilst alternative names consists of English name alternatives. Standard names have a one-to-many relationship with geographical features whilst alternative names stand in a many-to-one relationship with geographical features. Alternative names provide many surface forms of the name (e.g., Netherlands, the Netherlands, etc.). On the other hand standard names are more broad and may include feature specific qualifiers (e.g., Kingdom of the Netherlands, etc.). It is easier to find document placenames matching alternative names than standard names since people commonly use the shorter forms of placenames in documents.

### 2.3. Defining Geo-scopes

In this paper geo-scopes are limited to: (1) continent (CT) e.g., Europe, (2) continent directional (CD) as defined by the UN-statistics division[3] e.g., Western Europe, (3) country (PC) e.g., Netherlands, (4) country directional (PD) e.g., north-east-of Netherlands, (5) province (AM) e.g., Groningen and (6) province directional (AD) e.g., north-of Groningen. For directionally oriented scopes at country and province levels, the regions are divided into nine sections: north, north-east, east, south-east, south, south-west, west, north-west, and central.

#### 2.3.1. Continent and continent-directional scopes

Continent and continent-directional scopes consists of the following constituents: continent, countries, country-capitals (LC), provinces, provincial-capitals (LA) and cities with over 49,999 inhabitants. Table 3 shows the distribution of scopes, locations and names at continent and continent-directional level. The average ratio of name-to-location within the scopes is 4.68. There are 7 continent scopes compared to 24 continent-directional scopes.

#### 2.3.2. Country and country-directional scopes

Each country scope is defined by its child constituents, parent continent and adjacent countries. And each country-directional scope is defined by its child constituents and parent country. The following make up country and

[2]http://www.geonames.org

[3]http://unstats.un.org/unsd/default.htm

country-directional child constituents: country, country-capital, provinces, provincial-capitals, counties and cities with over 9,999 inhabitants. Distribution of scopes, locations and names at country and country-directional level is depicted in Table 3. The average ratio of name-to-location within the scopes is 1.73. There are 190 country scopes compared to 1089 country-directional scopes.

#### 2.3.3. Province and province-directional scopes

Each province scope is defined by its child constituents, parent country, and adjacent provinces. And each province-directional scope is defined by its child constituents and parent province. Province and province-directional consist of the following child constituents: province, provincial capitals, country-capitals, counties and all populated places. Table 3 shows the distribution of scopes, locations and names at province and province-directional level. The average ratio of name-to-location within the scopes is 1.02. There are 4,749 province scopes compared to 20,761 province-directional scopes.

| Scope | No. scopes | No. places | No. names |
|---|---|---|---|
| CT | 7 | 13,226 | 61,939 |
| CD | 24 | 13,226 | 61,990 |
| PC | 190 | 105,576 | 182,442 |
| PD | 1,089 | 105,569 | 182,442 |
| AM | 4,749 | 2,311,244 | 2,354,716 |
| AD | 20,761 | 2,005,682 | 2,068,732 |

Table 3: Geographic scope statistics. [see Section 2.3. for scope abbreviations.]

### 2.4. Storing Geo-scopes in Lucene Index

Each geo-scope group (e.g., continent scope) is stored in a separate index. Lucene provides the capability to query across multiple indexes. Ten Lucene fields are defined to store geo-scope data in the index: (1) scope-id (ID), (2) names of the scope (SNM), (3) names of capitals and populated places (i.e., cities, towns & villages) with large population (CNM), (4) names of primary administrative units (PAN), (5) names of secondary administrative units (SAN), (6) names of primary cities, towns and villages (PCN), (7) names of secondary cities, towns and villages (SCN), (8) names of adjacent regions of the same type (ASN), (9) names of parent regions (PRN) and (10) names of relatively smaller child places (CPN). The type of a place (e.g., capital city, provincial capital) and population size is used to group places within a scope category. For example to populate CNM field; cities, towns and villages with over 500.000 inhabitants are considered in country scope while the threshold is lowered to 100.000 inhabitants in province scope. Table 4 shows an example Lucene index data for the scope Europe. A complete geo-scope data storage layout inside the Lucene index is shown in Table 5.

| Field | Data |
|---|---|
| ID | EU |
| SNM | Europe, EU, Europa, etc. |
| CNM | – |
| PAN | Netherlands, Germany, Belgium, etc. |
| SAN | Groningen, Sachsen, Antwerp, etc. |
| PCN | Amsterdam, Berlin, Brussels, etc. |
| SCN | Utrecht, Hamburg, Antwerp, etc. |
| ASN | Africa, Asia, North America |
| PRN | Earth |
| CPN | Delft, Tournai, Unna, etc. |

Table 4: Example Lucene Index for scope Europe. [see Section 2.4. for acronym explanation.]

### 2.5. Resolving document scopes

The general idea is to assign each document to geo-scopes in the Lucene index. This basically involves three steps: (1) extracting place names, place types and geographical adjectives from the document using the geographical tagger, (2) submitting extracted geographical information to query the Lucene index of pre-defined geo-scopes, and (3) returning a ranked list of geo-scopes for the document. To effectly resolve a document's geo-scope with the approach reported in this paper, query formulation is crucial. The following features are considered in our query formulation strategy: (1) perceived importance of Lucene field (2) type of place, (3) importance of place determined by population and (4) the number of occurrences of place name in a document. The importance of assigning different weights to fields comes into play when the same place takes different roles in different scopes e.g., in the hierarchy `Groningen` ↦ `Netherlands` ↦ `Europe` ↦ `Earth`, `Groningen` is a primary administrative unit in `Netherlands` while a secondary administrative unit within `Europe`. That is, `Groningen` carries more importance within the scope `the Netherlands` in comparison to the scope within `Europe`. Importance is assigned to Lucene fields in the following order (i.e., descending order of importance): SNM ↦ CNM ↦ PCN ↦ PAN ↦ SCN ↦ SAN ↦ PRN ↦ CPN ↦ ASN. And weights are assigned to types of places according to the following order (i.e., descending order of importance): CT ↦ PC ↦ LC ↦ LA ↦ AM ↦ A2. Other cities are assigned weights according to their population size.

The aforementioned features are factored into our query formulation strategy as query term boost factor using Equation 2:

$$QueryGeoTermBoostFactor = tf * FWT * GWT \quad (2)$$

where $tf$ is the place name frequency count in the document, $FWT$ is the weight of the Lucene field being queried against and $GWT$ is place type or importance weight. Besides query formulation we pay attention to how the index is searched. Each geographical term in the query is analyzed to determine which field or fields to query against (e.g., Netherlands is submitted to search the field values of SNM and PAN as the Netherlands can be the name of scope Netherlands or the name of a primary administrative unit in scope Europe). Table 6 depicts feature weights implemented in our query formulation strategy. Geographical

| Scopes ↦ | CT | CD | PC | PD | AM | AD |
|---|---|---|---|---|---|---|
| ID | CT-ID | CD-ID | PC-ID | PD-ID | AM-ID | AD-ID |
| SNM | CT | | PC | | AM | |
| CNM | | | LC,P500 | LC,P500 | LA,LC,P150[a] | LA,P150 |
| PAN | PC | PC | AM | AM | A2 | A2 |
| SAN | AM | AM | A2[b] | A2 | | |
| PCN | LC,P500[c] | LC,P500 | LA,P100 | LA,P100 | P50 | P50 |
| SCN | LA,P100[d] | LA,P100 | P50 | P50 | P5[e],P10 | P5,P10 |
| ASN | CT | | PC | | AM | |
| PRN | EH[f] | CT | CT | PC | PC | AM |
| CPN | P50[g] | P50 | P10[h] | P10 | P0[i] | P0 |

[a]P150: Population centers (population $\geq$ 100,000).
[b]A2: Second order administrative division of a country.
[c]P500: Population centers (population $\geq$ 500,000).
[d]P100: Population centers (100,000 $\leq$ population $<$ 500,000).
[e]P5: Population centers (5,000 $\leq$ population $<$ 10,000).
[f]EH: Earth.
[g]P50: Population centers (50,000 $\leq$ population $<$ 100,000).
[h]P10: Population centers (10,000 $\leq$ population $<$ 50,000).
[i]P0: Population centers (population $<$ 5,000).

Table 5: Geo-scope data layout in Lucene index. [see Section 2.4. for explanations of acronyms.]

| Field | FWT | Type/Population | GWT |
|---|---|---|---|
| ID | - | CT | 10.0 |
| SNM | 10.0 | Country | 9.0 |
| CNM | 9.0 | Province | 2.5 |
| PAN | 5.0 | County | 1.5 |
| SAN | 3.0 | CountryCapital | 9.0 |
| PCN | 8.0 | ProvinceCapital | 7.0 |
| SCN | 5.0 | people $\geq$ 1M | 9.0 |
| ASN | 1.5 | 0.5M $\leq$ people $<$ 1M | 8.0 |
| PRN | 2.0 | 0.1M $\leq$ people $<$ 0.5M | 7.0 |
| CPN | 2.0 | 50K $\leq$ people $<$ 100K | 6.0 |
| | | 10K $\leq$ people $<$ 50K | 5.0 |
| | | 5K $\leq$ people $<$ 10K | 2.0 |
| | | people $<$ 5K | 1.0 |

Table 6: Field and place type weights. [see Section 2.4. for explanations of acronyms.]

adjectives, like placenames are highly ambiguous – seeing the geographical adjective `French` in a document does not necessarily refer to things explicitly connected to the nation of France (e.g., `French` in a document may refer to a subject in school or a type of cooking). Nevertheless, if used judicially, geographical adjectives can provide useful information to geographically resolve document scopes. We map query geographical adjectives (e.g., Dutch) and placename abbreviations (e.g., UK) to their corresponding country names (e.g., Dutch mapped-to Netherlands) and assign lower weights to them. We did not try to resolve geographical adjective ambiguities, instead we assume that the places the adjective is referring to are mentioned in the document and therefore, the geo-scope resolver will use the adjective to further reinforce scope resolution.

To illustrate our geo-scope resolution approach, consider a sample document containing the following placenames with their respective term frequency in brackets: `New York (1), Rwanda (4), France (1), Kigali (1)`[4]. Table 7 depicts how query geographical terms are analyzed per field at querying processing. Each geographical term is assigned a weight (in square brackets) according to Equation 2. The document is geographically resolved to ranked geo-scopes as: `Rwanda (0.082667), Eastern Africa (0.007700), Africa (0.004359), France (0.003444), United States (0.001750)`.

| Field | Query Formulation |
|---|---|
| ID | - |
| SNM | new york[25.0] rwanda[360.0] france[90.0] |
| CNM | kigali[81.0] |
| PAN | new york[12.5] rwanda[180.0] france[45.0] |
| SAN | new york[7.5] |
| PCN | new york[56.0] kigali[72.0] |
| SCN | new york[35.0] |
| ASN | new york[3.75] rwanda[54.0] france[13.5] |
| PRN | new york[5.0] rwanda[72.0] france[18.0] |
| CPN | new york[12.0] |

Table 7: Example query formulation for per field querying. [see Section 2.4. for explanations of acronyms.]

## 3. Placename Referent Resolver (PRR)

The placename referent resolver is a component that performs the PRAR task. PRR is fed the output of the (GeoSR) geographical scope resolver (i.e., a list of ranked document geo-scopes) and the output of the geographical tagger (i.e.,

[4]New York (State or City), Rwanda (Country), France (Country), Kigali (Country capital)

a list of place names extracted from the document) (see Figure 1). Figure 2 shows the algorithm to realize PRAR.

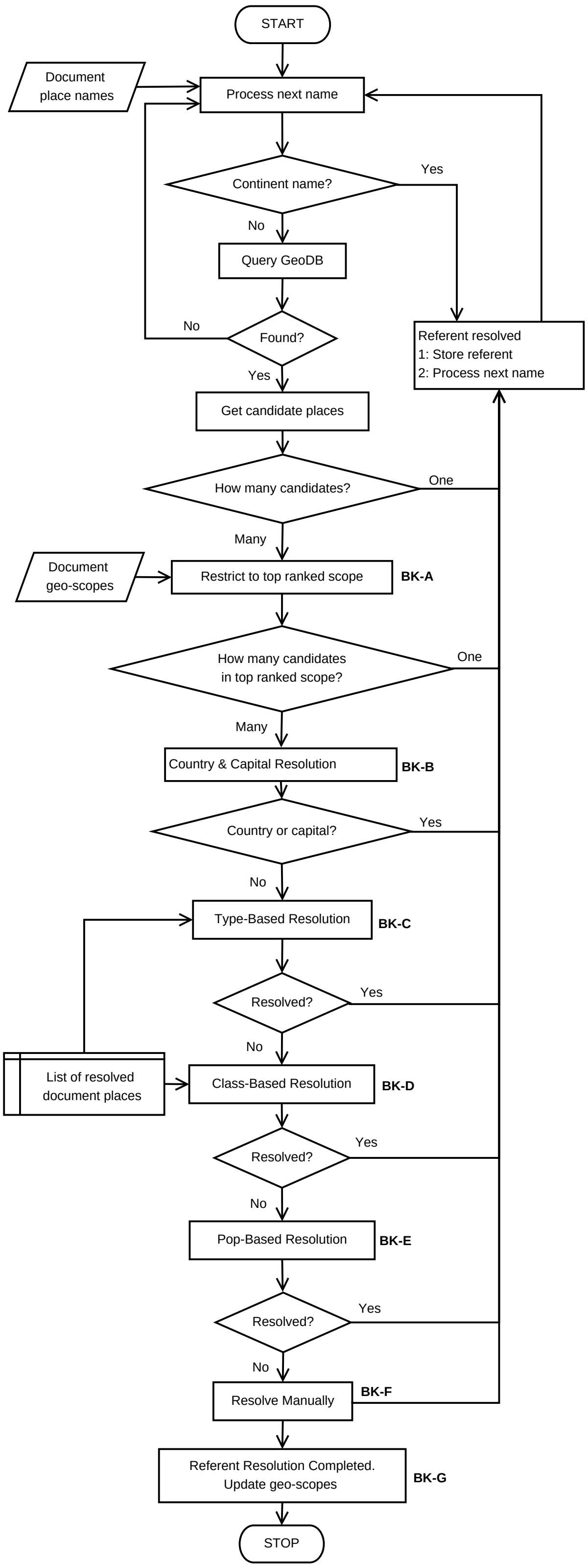


Figure 2: Referent ambiguity resolution algorithm.

Here we describe the functionality of the main processing blocks shown in Figure 2. The algorithm starts by assigning continent place names to continents. It then extracts candidate places for place names other than continent names from the geographical database (GeoDB). Place names with a single candidate place are resolved to these places while place names with multiple candidate places are passed to lower processing blocks starting with the scope restriction block. For illustration purposes, we use a sample document containing the following place names: `Sarajevo, Bosnia, Bihac, Tuzla, Britain, London`.

**Scope restriction block (BK-A):** This module extends the country-level restriction as reported in Pauliquen et al (2006). It exploits an elaborate list of ranked geographical scopes assigned to a document. A place name with multiple candidate referents is assigned to a single top ranked document geo-scope. The other candidates belonging to lower ranked document geo-scopes are discarded. If a selected scope contains a single candidate, the candidate is marked as the place being referred to by the name. The main source of error when using scope restriction arises from an inherited GeoSR error. However, if a selected scope contains multiple candidates, it is passed to the next processing block i.e., country & capitals resolution (BK-B). Back to our example above, the place names are restricted to the following scopes (scopes are presented as `NAME:COUNTRY@PROVINCE[CANDIDATE IDs]`): `Sarajevo:BA@01[1], Bosnia:BA@00[2], Bihac:BA@01[3], Tuzla:BA@01[4,5,6], Britain:GB@00[7,8], London:GB@H9[9,10]`. `Sarajevo, Bosnia and Bihac` are non-ambiguously resolved through scope restriction because the assigned scopes contain one candidate place each. On the other hand, `Tuzla, Britain and London` remain ambiguous within selected scopes because they contain multiple candidate places.

**Country & capitals resolution (BK-B):** A place name's candidate place of type country (PC) or country-capital (LC) or provincial capital (LA) is selected as the place being referred to by the name. The order of preference is PC $\mapsto$ LC $\mapsto$ LA. If the ambiguity is not resolved at this stage, it is passed to the next processing block i.e., Type-based resolution (BK-C). Back to our example above; we select any candidates for `Tuzla, Britain and London` which are of type PC or LC or LA as the referent. This routine resolves `Britain and London` to places of type PC and LC respectively. `Tuzla` remains ambiguous within the selected scope.

**Type-based resolution (BK-C):** Type-based resolution exploits types of resolved places as the basis to resolve among competing candidate places. The commonly occurring types are preferred. The assumption is that places of a similar type are more likely to be mentioned in a discourse. The candidate place of type matching the commonly occurring type among the resolved places is selected as the place being referenced. Back to our example above; here is the list of already resolved referents with their types in curly brackets: `Sarajevo{PPLC}, Bosnia{PCLI}, Bihac{PPL}, Britain{PCLI}, London{PPLC}`. From this list there are two places of type `PPLC`, two places of type `PCLI` and one place of type `PPL`. The ambiguous `Tuzla:BA@01[4,5,6]` has three candidate places in scope BA@01. The types of these candidate places are (candidate ID in square bracket and type in curly bracket): [4]{PPL}, [5]{ADM2} and [6]{ADM3}. Candidate [4]'s type matches one of the types of resolved referents and therefore, is selected as the place referred to by name Tuzla.

**Class-based resolution (BK-D):** The class-based resolution procedure is similar to the type-based resolution routine. The class-based procedure exploits feature classification of resolved places as the basis to resolve among competing candidate places (see Sec. 2.2. for feature classification detail). Again the assumption is that places of a similar class are more likely to be mentioned in a discourse. The candidate place of class matching the most frequently occurring class among the resolved places is selected as the place referred to. Back to our example above; we will try to resolve among the three candidates of Tuzla in scope BA@01 employing the class-based procedure. Here is a list of resolved places with their corresponding class in curly brackets: Sarajevo{P}, Bosnia{A}, Bihac{P}, Britain{A}, London{P}. There are two places classified as A and three places classified as P. The three candidates of reference Tuzla are classified as (candidate ID in square bracket and classification in curly bracket): [4]{P}, [5]{A} & [6]{A}. Candidate [4]'s class matches the most frequently occurring class among the resolved places and therefore, is selected as the place referred to by name Tuzla.

**Pop-based resolution (BK-E) & manual resolution (BK-F):** Population based resolution (BK-E) selects the place with the largest population as the place being referred to. Manual resolution (BK-F) passes the task of resolving among competing places to the user. Manual resolution is called when the preceding automated procedures fail to resolve the ambiguity.

**Update geo-scopes (BK-G):** Here the list of a document's ranked geographical scopes is updated by including only the scopes containing resolved places and their ancestor geo-scopes. The remaining geo-scopes in the ranked list are discarded. From our example above, scope list update with respect to London and Britain will include: Europe, GB@00, GB@H9, GB@S.East, Northern Europe, GB@H9@S.East. The following scopes in the original ranked scope list are discarded: CA@East, CA@08, CA@08@S.East, CA@00 where GB and CA stand for Great Britain and Canada respectively. The scope Canada featured in the original scope list because of a place named London in Ontario, Canada.

# 4. Evaluation

Here we report on geographical scope resolver (GSR) evaluation. Because of time constraints and lack of test dataset, we were unable to fully evaluate placename referent resolver (PRR) for this paper. However, a preliminary test on 102 documents containing 195 ambiguous place names, our PRR resolved 181 (92.8%) of the place names correctly[5].

[5]A comprehensive evaluation of our placename referent resolver (PRR) will be reported in the PhD thesis in preparation.

## 4.1. GSR Evaluation

### 4.1.1. Dataset

We evaluated our implementation using the CoNLL-2003 Shared Task (Sang and Meulder, 2003) training and development set of 1162 documents for English. The CoNLL-2003 English dataset is derived from the Reuters English corpus (RCV1) (Rose et al., 2002). Of the 1162 documents, 1124 documents contain geographical terms (place names and geographical adjectives). These documents have geographical scopes at country levels assigned to them. Of 1124 documents 686 were assigned single scopes, 313 double, 90 triple and 35 four or more.

### 4.1.2. Results

Our system can assign geographical scopes up to six levels: continent, continent-directional, country, country-directional, province and province-directional. For this evaluation, we turned on the country level scope resolver for that is the scope level assigned to our test document collection. Our system resolves documents geographically to multiple scopes ranking them from the most significant to the least significant scope.

**Single Scoped Documents.** Of the 686 documents with single scope, our system assigned scopes correctly to 645 (94%) documents (that is, the scopes assigned to the 645 documents were ranked at position one).

**Two Scoped Documents.** Of the 313 documents with two scopes, our system assigned scopes correctly to 197 (62.94%) documents (that is, the scopes assigned to the 197 documents were ranked at the top two positions). The remaining 116 (37.06%) documents had one scope correctly assigned to them in the top two rank postions.

**Three Scope Documents.** Of the 90 documents with three scopes, our system assigned scopes correctly to 18 (20%) documents (that is, the scopes assigned to the 18 document were ranked at the top three positions). Of the remaining 72 documents, 48 (53.33%) documents were correctly assigned two scopes in the top three rank postions. The remaining 24 (26.67%) documents had one scope correctly assigned to them in the top three rank positions.

# 5. Conclusion

We described a complete placename ambiguity resolution system consisting of three components: a geographical tagger, a geographical scope resolver (GeoSR) and a placename referent resolver (PRR). The last two components are built in-house while the geographical tagger is an off-the-shelf software component.

The novelty in GeoSR is that it uses unresolved place names as opposed to resolved place names used in previous works (Amitay et al., 2004; Martins and Silva, 2005). This means that geographical scopes can be computed independent of geographic name resolution, and thus does not suffer from mistakes in placename resolution. Also the GeoSR is implemented using a standard IR library exploiting a number of features, namely, placename frequency of occurrence, geographical adjectives, place type, population, vertical (transitive parent/child relation) and horizontal (adjacency relation) relationship among places. The GeoSR achieved a promising result on a subset of the Reuters English corpus (RCV1) dataset comparable with (Amitay et al., 2004; Martins and Silva, 2005): single scoped documents (96%) and two scoped documents (62.94%). However, the system performance for a three or more scoped documents is very poor (20%).

The novelty in PRR is that it uses an elaborate list of ranked geographical scope as the basis to resolve place ambiguity. The PRR also makes extensive use of place types and classification to resolve among competing candidate places. However, we are unable to evaluate PRR because of time constraints and lack of test dataset.

Lastly, there is an urgent need for freely available datasets to evaluate referent and scope resolution approaches. The datasets should consist of various genres, e.g., news articles and webpages. Leidner's work on toponomy resolution is a step in the right direction (Leidner, 2007).

## 6. Acknowledgements

This work is supported by NUFFIC within the framework of The Netherlands Programme for the Institutional Strengthening of Post-secondary Training Education and Capacity (NPT) under the project titled "Building a sustainable ICT training capacity in the public universities in Uganda".

## 7. References

Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. Web-a-Where: Geotagging Web content. In *Proceedings of SIGIR-04, the 27th conference on research and development in information retrieval*, pages 273–280. ACM Press.

Geoffrey Andogah and Gosse Bouma. 2007. University of Groningen at GeoCLEF 2007. In *Working Notes for CLEF 2007*, Budapest, Hungary.

Nuno Cardoso, David Cruz, Marcirio Chaves, and Mário J. Silva. 2007. The University of Lisbon at GeoCLEF 2007. In *Working Notes for CLEF 2007*, Budapest, Hungary.

Gaihua Fu, Christopher B. Jones, and Alia I. Abdelmonty. 2005. Ontology-based Spatial Query Expansion in Information Retrieval. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, volume 3761/2005, pages 1466–1482. Springer (Lecture Notes in Computer Science LNCS).

Otis Gospodnetic and Eric Hatcher. 2005. *Lucene in Action*. Manning Publications Co., 206 Bruce Park Avenue, Greenwich, CT 06830.

Jens Graupmann and Ralf Schenkel. 2006. GeoSphere-Search: Context-Aware Geographic Web Search. In *Proceedings of the 3rd Workshop on Geographic Information Retrieval held at The 29th Annual International ACM SIGIR Conference*, Seattle, WA, USA.

Ray R. Larson, Fredric C. Gey, and Vivien Petras. 2006. Berkeley at GeoCLEF: Logistic Regression and Fusion for Geographic Information Retrieval. In *Accessing Multilingual Information Repositories*, volume 4022/2006, pages 963–976. Springer (Lecture Notes in Computer Science LNCS).

Dik L. Lee, Huei Chuang, and Kent Seamons. 1997. Document ranking and the vector space model. *IEEE Software*, pages 67–75.

Jochen Lothar Leidner. 2007. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. 2007. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview . In *Working Notes for CLEF 2007*.

Bruno Martins and Mario J. Silva. 2005. A graph-ranking algorithm for geo-referencing documents. In *Proceedings of ICDM-05, the 5Th IEEE International Conference on Data Mining*.

Bruno Martins, Mario J. Silva, Sergio Freitas, and Ana Paula Afonso. 2006. Handling locations in search engine queries. In *Workshop on Geographical Information Retrieval, SIGIR'06*, August.

Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fluart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, and Clive Best. 2006. Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation. In *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 53–58.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus Volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, volume 3, pages 827–833.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language independent named entity recognition. In *Walter Daelemans and Miles Osborne, Editors, Proceedings of CoNLL-2003*, pages 142–147.

# Annotating Natural Language Geographic References

**Inderjeet Mani, Janet Hitzeman and Cheryl Clark**

**The MITRE Corporation**

**202 Burlington Road**

**Bedford, MA 01730, USA**

E-mail: **{imani, hitz, cclark} at mitre dot org**

**Abstract**

SpatialML is an annotation scheme for marking up references to places in natural language. It covers both named and nominal references to places, grounding them where possible with geo-coordinates, including both relative and absolute locations, and characterizes relationships among places in terms of a region calculus. A freely available annotation editor has been developed for SpatialML, along with three annotated corpora, including a corpus of annotated documents released by the Linguistic Data Consortium. Inter-annotator agreement on SpatialML extents is 77.0 F-measure on that corpus, but 92.3 F-measure on another (ProMED) corpus. The paper discusses a number of issues affecting inter-annotator agreement.

## 1. Introduction

The problem of understanding spatial references in natural language poses many interesting opportunities and representational challenges. Spatial references include both ‘absolute’ references (e.g., “Rome”, “Rochester, NY”, “southern Kerala district of Cudallah”), as well as relative references (“thirty miles north of Boston”, “an underpass beneath Pushkin Square”, “in the vicinity of Georgetown University"). We have developed an annotation scheme called SpatialML[1] that attempts to address these concerns.

The main goal of SpatialML is to mark places mentioned in text (indicated with PLACE tags) and map them to data from gazetteers and other databases. Semantic attributes such as country abbreviations, country subdivision and dependent area abbreviations, and geo-coordinates are used to help establish such a mapping. SpatialML uses LINK tags to express relations between places, such as inclusion between regions, and PATH tags to capture spatial trajectories for relative locations, involving a particular direction and/or distance. The SpatialML guidelines indicate language-specific rules for marking up SpatialML tags in English, as well as language-independent rules for marking up semantic attributes of tags. The guidelines also provide a handful of multilingual examples.

There are two critical aspects that make this approach especially attractive: (i) the annotation scheme is compatible with a variety of different standards (ii) most of the resources and tools used are freely available. For practical reasons, our focus is on geography and culturally-relevant landmarks, rather than other domains of spatial language. However, we expect that these guidelines could be adapted to other domains with some extensions without changing the fundamental framework.

[1]http://sourceforge.net/projects/spatialml

Any representation scheme for spatial language has to decide what sorts of entities should be tagged as PLACEs. Natural language allows buildings and other objects to be coerced into places, as in “I visited the Eiffel Tower”. We are ontologically permissive in our annotation, annotating such entities as PLACEs. Natural language also abounds in metonymic references to places, as in “Italy has announced its withdrawal”. While the annotation scheme does not attempt to distinguish metonymic from non-metonymic references, it does record the fact that the country of Italy is mentioned. The representation scheme must also deal with spatial relationships of interest; to represent topological relations, we adopt a modified version of a region calculus.

We discuss the annotation scheme in Section 2, followed, in Section 3, by an account of the standards that the scheme is compatible with. In Section 4, we illustrate the annotation editor, and describe the annotated corpora, while Section 5 discusses inter-annotator agreement and associated annotation challenges. Section 6 concludes. For a description of automatic tagging of SpatialML, see (Mani et al. 2008).

## 2. SpatialML Annotation Approach

In order to make SpatialML easy to annotate by people without considerable training, the annotation scheme is kept fairly simple, with straightforward rules for what to mark and with a relatively “flat” annotation scheme. Here is an example:

*<PLACE id=“4” type=“PPL” country=“TW” form=“NAM” latLong=“22◦37’N 120◦21’E”>***Fengshan***</PLACE>*

Here the place is marked as being a named place, and in addition, key gazetteer-related attributes are filled in, including latitude and longitude, and the country code for Taiwan.

SpatialML also tags nominal references to places. In this example, we see that the mention has been tagged as a nominal reference to an entity that is a facility.

***a*** *<PLACE id="1" type="FAC" form="NOM">**building**</PLACE>*

Now, let us consider a location which is expressed relative to another. The idea here is that the relative location's offset as described in the text are captured in the tags. The PATH tag expresses a relation between a source PLACE and a target PLACE, qualified by distance and direction attributes.

***a*** *<PLACE id="1" type="FAC" form="NOM">**building**</PLACE>*
*<SIGNAL id="2">**5 miles**</SIGNAL>*
*<SIGNAL id="3">**east**</SIGNAL>* ***of***
*<PLACE id="4" type="PPL" country="TW" form="NAM" latLong="22◦37'N 120◦21'E">**Fengshan**</PLACE>*
*<PATH id="5" source="4" destination="1" distance="5:mi" direction="E" signals="2 3"/>*

| | |
|---|---|
| WATER | River, stream, ocean, sea, lake, canal, aqueduct, geyser, etc. |
| CELESTIAL | Sun, Moon, Jupiter, Gemini, etc. |
| CIVIL | Political Region or Administrative Area, usually sub-national, e.g. State, Province, certain instances of towns and cities. |
| CONTINENT | Denotes a continent, including ancient ones. |
| COUNTRY | Denotes a country, including ancient ones. |
| FAC | Facility, usually a catchall category for restaurants, churches, schools, ice-cream parlors, bowling alleys, you name it! |
| GRID | A grid reference indication of the location, e.g., MGRS (Military Grid Reference System) |
| LATLONG | A latitude/longitude indication of the location |
| MTN | Mountain |
| MTS | Range of mountains |
| POSTALCODE | Zip codes, postcodes, pin codes etc. |
| POSTBOX | P. O. Box segments of addresses |
| PPL | Populated Place (usually conceived of as a point), other than PPLA or PPLC |
| PPLA | Capital of a first-order administrative division, e.g., a state capital |
| PPLC | Capital of a country |
| RGN | Region other than Political/Administrative Region |
| ROAD | Street, road, highway, etc. |
| STATE | A first-order administrative division within a country, e.g., state, province, gubernia, territory, etc. |
| UTM | A Universal Transverse Mercator (UTM) format indication of the location |
| VEHICLE | Car, truck, train, etc. |

**Table 1: Types of Places Annotated**

We have opportunistically drawn the inventory of different PLACE *type*s (20 in all, shown in Table 1) from the much larger thesaurus (211 categories) of the Alexandria Digital Library (ADL)[2].

| LinkType | Example |
|---|---|
| IN (tangential and non-tangential proper parts) | [Paris], [Texas] |
| EC (extended connection) | the border between [Lebanon] and [Israel] |
| NR (near) | visited [Belmont], near [San Mateo] |
| DC (discrete connection) | the [well] outside the [house] |
| PO (partial overlap) | [Russia] and [Asia] |
| EQ (equality) | [Rochester] and [382044N 0874941W] |

**Table 2: Link Types**

The set of LINK types is derived from the Region Connection Calculus (RCC8) (Randell et al. 1992, Cohn et al. 1997). The LINK codes are shown in Table 2.
Here is an example involving LINKs. Both English and Chinese versions are shown.

*a [town] some [50 miles] [south] of [Salzburg] in the central [Austrian] [Alps]*
***a*** *<PLACE type="PPL" id=1 form="NOM" ctv="TOWN">**town**</PLACE>*
*<SIGNAL id=2>**50 miles**</SIGNAL>*
*<SIGNAL id=3>**south**</SIGNAL>* ***of***
*<PLACE id=4 type="PPLA" country="AT" form="NAM">**Salzburg**</PLACE>* ***in the central***
*<PLACE id=5 type="COUNTRY" country="AT" mod="C">**Austrian**</PLACE>*
*<PLACE id=6 type="MTS">**Alps**</PLACE>*
*<PATH id=7 distance="50:mi" direction="S" source= 4 destination=1 signals="2 3"/>*
*<LINK id=8 source=1 target=6 linkType="IN"/>*
*<LINK id=9 source=6 target=5 linkType="IN"/>*

*我居住在一个离中[奥地利] [阿尔卑斯] [萨尔茨堡] [以南]大约[50 英哩] 的 [镇子]里。*

*我居住在一个离中*
*<PLACE id =1 type="COUNTRY" country="AU" mod="C">**奥地利**</PLACE>*
*<PLACE id =2 type="MTS">**阿尔卑斯** </PLACE>*
*<PLACE id=3 type="PPLA" country="AT" form="NAM">**萨尔茨堡**</PLACE>*
*<SIGNAL id=4>**以南**</SIGNAL> 大约*
*<SIGNAL id=5>**50 英哩**</SIGNAL> 的*
*<PLACE type="PPL" id=6 form="NOM"*

[2] http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/top.htm

*ctv=“TOWN”>镇子</PLACE>里。*
*<PATH id=7 distance=”50:mi” direction=S source=3 destination=6 signals=“2 3”/>*
*<LINK id=8 source=1 target=6 linkType=“IN”/>*

Syntactically, SpatialML tries to keep the tag extents as small as possible, to make annotation easier. Pre-modifiers such as adjectives, determiners, etc. are NOT included in the extent unless they are part of a proper name. For example, for “the river Thames,” only “Thames” is marked, but, for the proper names “River Thames” and “the Netherlands,” the entire phrase is marked. There is no need for tag embedding, since we have non-consuming tags (LINK and PATH) to express relationships between PLACEs. Adjectival forms of proper names (“U.S.,” “Brazilian”) are, however, tagged in order to allow one to link expressions such as “Georgian” to “capital” in the phrase “the Georgian capital”.

Deictic references such as “here” are not tagged. Non-referring expressions, such as “town” and “city” in “a small town is better to live in than a big city.” aren’t tagged. Also, “city” in “the city of Baton Rouge” is not tagged; the use of such a modifier is simply to indicate a property of the PLACE. In contrast, when “city” does refer, as in “John lives in the city” where “the city,” in context, must be interpreted as referring, for example, to Baton Rouge, it is tagged as a place and given the coordinates, etc., of Baton Rouge.

## 3. Standards Compatibility

SpatialML leverages ISO (ISO-3166-1 for countries and ISO-3166-2 for provinces), as well as various proposed standards towards the goal of making the scheme compatible with existing and future corpora.

The SpatialML guidelines are compatible with existing guidelines for spatial annotation and existing corpora within the Automatic Content Extraction[3] (ACE) research program. In particular, we exploit the English Annotation Guidelines for Entities (Version 5.6.6 2006.08.01), specifically the GPE, Location, and Facility entity tags and the Physical relation tags, all of which are mapped to SpatialML tags. In comparison with ACE, SpatialML attempts to use a classification scheme that’s closer to information represented in gazetteers, thereby making the grounding of spatial locations in terms of geo-coordinates easier. SpatialML also doesn’t concern itself with referential subtleties like metonymy; the latter has proven to be difficult for humans to annotate. Finally, SpatialML addresses relative locations involving distances and topological relations that ACE ignores. ACE ‘GPE’, ‘Location’, and ‘Facility’ Entity types are representable in SpatialML, as are ACE ‘Near’ Relations. SpatialML, unlike ACE, is a ‘flat’ annotation scheme. Instead of grouping mentions into classes (called “entities” in ACE), SpatialML simply annotates mentions of places.

We also borrow ideas from the Toponym Resolution Markup Language of Leidner (2006), the research of Schilder et al. (2004) and the annotation scheme in Garbin and Mani (2005).

The SpatialML annotation scheme can be integrated with the Geography Markup Language[4] (GML) defined by the Open Geospatial Consortium (OGC). Mappings have also been implemented from SpatialML to Google Earth’s Keyhole Markup Language (KML), and from the output of a commercial geo-tagging tool, MetaCarta, to SpatialML.

## 4. Annotation Environment and Corpora

We have annotated documents in SpatialML using the freely available Callisto[5] annotation editor (Figure 1) which includes the SpatialML task extension.

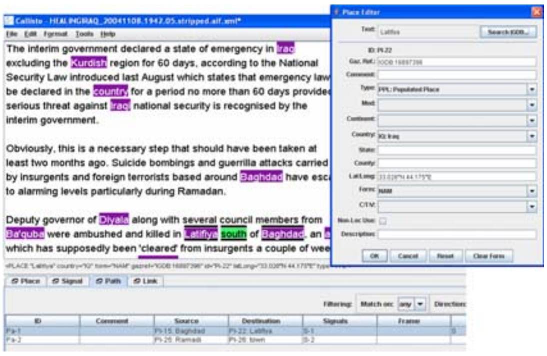

**FIGURE 1: Callisto Editing Session**

The gazetteer used is the Integrated Gazetteer Database (IGDB) (Mardis and Burger 2005) (Sundheim et al. 2006). IGDB integrates together place name data from a number of different resources, including NGA GeoNames[6], USGS GNIS[7], Tipster, WordNet, and a few others. It contains about 6.5 million entries. The ADL Gazetteer Protocol[8] is used to access IGDB.

Three corpora have been annotated in SpatialML. The first consists of 428 ACE documents, originally from the University of Pennsylvania’s Linguistics Data Consortium (LDC), has been annotated in SpatialML. This corpus, drawn mainly from broadcast conversation, broadcast news, news magazine, newsgroups, and weblogs, contains 6338 PLACE tags, of which 4,783 are named PLACEs with geo-coordinates. This *ACE SpatialML Corpus* (ASC) has been re-released to the LDC, and is available to LDC members from the LDC Catalog as LDC2008T03[9].

The second corpus consists of 100 documents from ProMED[10], an email reporting system for monitoring emerging diseases provided by the International Society for Infectious Diseases. This corpus yielded 995 PLACE tags.

[3]http://projects.ldc.upenn.edu/ace/annotation/2005Tasks.html

[4]http://www.opengis.net/gml/
[5]http://callisto.mitre.org
[6]http://gnswww.nga.mil/geonames/GNS/index.jsp
[7]http://geonames.usgs.gov/pls/gnispublic
[8]http://www.alexandria.ucsb.edu/downloads/gazprotocol/
[9] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T03
[10]http://www.promedmail.org

The third is a corpus of 121 news releases spidered from the U.S. Immigration and Customs Enforcement (ICE) web site[11]. This corpus provides 3,477 PLACE tags.

As a result of the ProMED annotation, we decided that a "non-loc" feature was needed to address location complements (*outside Australia*; *all continents except Europe*). This feature has been added to the SpatialML Guidelines version 2.2. We also discovered some limits to the expressiveness of SpatialML. Sets of places (*the Americas*) and complex modification (*subtropical and temperate regions of …*) aren't handled as yet. Topological relations between a pair of locations (*the border between Lebanon and Israel*) can be represented in the LINK representation, but 3-way relationships (*meeting of the frontiers between Peru, Colombia and Brazil*) cannot be represented.

## 5. Inter-Annotator Agreement

Inter-annotator agreement on SpatialML PLACE tags in the ASC corpus is 77.0 F-measure.

Disagreements stemmed from two sources: application of guidelines and use of tools. The guideline application problems included an annotator failing to mark discourse-dependent references like "the state", as well as specific references like "area" (to be marked as a REGION), incorrectly marking generic phrases like "areas" or "cities", among others.

The disagreement due to tool use has to do with one version of Callisto lacking the ability to carry out inexact string matches for text mentions of places against IGDB entries, including adjectival forms of names (e.g., "Rwandan"), different transliterations (e.g., "Nisarah" vs. "Nisara"), in addition to various alternative ways of looking up a name ("New York, State of" vs. "New York"). Computing agreement on disambiguation in the ASC is underway.

| Attribute | P | R | F |
|---|---|---|---|
| Extent | 89.32 | 95.4 | 92.3 |
| Form | 100 | 99.14 | 99.56 |
| LatLong | 96.51 | 57.22 | 71.85 |
| Gazref | 70.44 | 57.17 | 63.11 |

**Table 3: Inter-annotator agreement on ProMED**

Table 3 shows the agreement on SpatialML attributes for ProMED.

The agreement on extent is much higher than on the ASC, for two reasons. First, it was carried out much later in the project, with later versions of the tools as well as guidelines. Second, both annotators were expert linguistic annotators, whereas in the first study only one was (her annotations were used for the ASC).

The lower agreement on LatLong is due to different versions of Callisto being used in the study, giving rise to the tool use issues mentioned above. The higher agreement on LatLong compared to Gazref (i.e., IGDB gazetteer id) is a result of not being able to find an entry with a geo-coordinate in IGDB, using the Web instead, or else finding an alternative (redundant) entry in IGDB. These observations re-emphasize the need to take both guidelines and tool training into account during annotation.

It is worth pointing out that the level of agreement on disambiguation in turn depends on the size of the gazetteer. Large gazetteers increase the degree of ambiguity; for example, there are 1420 matches for the name "La Esperanza" in IGDB. A study by (Garbin and Mani 2005) on 6.5 million words of news text found that two-thirds of the place name mentions that were ambiguous in the USGS GNIS gazetteer were 'bare' place names that lacked any disambiguating information in the containing text sentence.

As with any expressive annotation scheme, given certain natural language constructions, the annotator may need to choose between different types of formal annotations that may or may not perfectly fit the natural language. The phrase "Pacific coast of Australia" is annotated as two places, "Pacific" of type WATER, with a modifier "BR" indicating a border, and with "Australia" being of type COUNTRY (or CONTINENT, depending on context). The former PLACE is linked to the latter with a LINK type of EC. An alternative would be to represent "coast" as a PLACE of type RGN, with a modifier "W" indicating west, and having the same EC relation to Australia. A weakness of the present annotation methodology in NLP in general is that it isn't directly tied to inferential processing; the latter could help provide additional criteria for such choices[12].

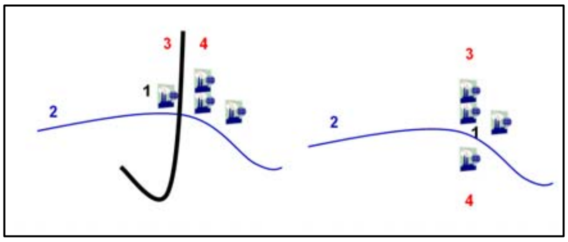


**FIGURE 2: Spatial Interpretations**

Our guidelines include the stern injunction that "the annotator is not to use specialized knowledge that is not part of commonsense knowledge that everyone is expected to have". The annotator must rely solely on the information in the text and in the gazetteer in order to keep the annotation more representative of general geospatial knowledge, and therefore more consistent with the work of other annotators. This guideline is sometimes

[11] http://www.ice.gov/

[12] In other words, *no representation without utilization*, to paraphrase an old AI slogan.

hard to enforce. Consider the phrase "at the [factory $_1$] spanning the [Winooski River $_2$] in [Essex Junction $_3$] and [Williston $_4$]". The text here is compatible with many different scenarios. One annotator marked the pairs of PLACEs {2, 3}, {2, 4}, {1, 3}, and {1, 4}, each LINKed by the relation PO, with in addition the pair {1, 2} being linked by EC, interpreting the relations as in the left hand side of Figure 2 (where the thick arc is the border between Essex Junction and Williston).

However, a second annotator knowledgeable about the area pointed out that the factory (an IBM factory allegedly responsible for some serious environmental hazards) sits on both sides of the Winooski river -- the latter forms the boundary between Essex Junction and Williston. (To make matters worse, the factory's address is in Essex Junction, but part of the facility is actually in Williston). This interpretation corresponds to the right hand side of Figure 2. Both annotators were over-specific in their interpretations; a preferred annotation would not commit to the relations between the river, the factory, and the two villages.

## 6. Conclusion

Current work on SpatialML is focused on further inter-annotator studies and annotation of additional corpora, including multilingual data. In joint work with Brandeis University, we will also be integrating SpatialML with TimeML (Pustejovsky et al. 2005) and the Suggested Upper Merged Ontology[13] (SUMO). The automatic taggers have been integrated with tools such as Google Earth to provide for a text-to-map capability for those expressions that can be grounded. PATH expressions (as in the case of "a building five miles east of Fengshan") result in lines being drawn between the source and target PLACEs. Research is underway to determine appropriate fudge factors to compute the actual orientation and length of such lines from their natural language descriptions. A few primitives have been introduced to represent orientation relations expressed in language ("top", "bottom", etc.). The extent to which these can be extended cross-linguistically remains to be seen.

## 7. References


Cohn, A. G., Bennett, B., Gooday, J. and Gotts, N. M. (1997). Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. GeoInformatica, 1, 275–316.

Garbin, E. and Mani, I. (2005). Disambiguating Toponyms in News. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 363–370.

Leidner, J. L. (2006). Toponym Resolution: A First Large-Scale Comparative Evaluation. Research Report EDI-INF-RR-0839.

Mani, I., Hitzeman, J., Richer, J., Harris, D. R., Quimby, R. and Wellner, B. (2008). SpatialML: Annotation Scheme, Corpora, and Tools. Proceedings of LREC'2008.

Mardis, S., and Burger, J. (2005). Design for an Integrated Gazetteer Database: Technical Description and User Guide for a Gazetteer to Support Natural Language Processing Applications. MITRE TECHNICAL REPORT, MTR 05B0000085. http://www.mitre.org/work/tech_papers/tech_papers_06/06_0375/index.html

Pustejovsky, James, Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G. and Mani, I. (2005). *The Specification Language TimeML*. In I. Mani, J. Pustejovsky, and R. Gaizauskas, (eds.), The Language of Time: A Reader, 545-557, Oxford University Press.

Randell, D. A., Cui, Z. and Cohn, A. G. (1992). A Spatial Logic Based on Regions and Connection, Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning, Morgan Kaufmann, San Mateo, pp. 165–176.

Schilder, F., Versley, Y. and Habel, C. (2004). Extracting Spatial Information: Grounding, Classifying and Linking Spatial Expressions. Workshop on Geographic Information Retrieval at the 27th ACM SIGIR conference, Sheffield, England, UK.

Sundheim, B., Mardis, S. and Burger, J. (2006). *Gazetteer Linkage to WordNet*. The Third International WordNet Conference, South Jeju Island, Korea. http://nlpweb.kaist.ac.kr/gwc/pdf2006/7.pdf


[13] http://www.ontologyportal.org/

# Spatial entities are temporal entities too: the case of motion verbs

**Nicholas Asher, Philippe Muller, and Mauro Gaio**

## Abstract

We argue that spatio-temporal primitives are crucial in giving a full view of the spatial and temporal structure of texts. We think that temporal and spatial structure are projections of a more complex and more complete spatio-temporal structure. We will make our case based on the an analysis of movement verbs, showing how they contribute in an important way to both temporal and spatial structure within discourse. Our analysis of movement verbs is based on a detailed lexical semantic analysis of a wide class of verbs in French. We give some ideas for how this lexical semantics when coupled with an analysis of how clauses involving these expressions are related to each other within a discourse using rhetorical relations can aid in determining the spatio-temporal structure of the text. We apply our approach to descriptions of climbing cliffs as well as descriptions of walking tours in the Pyrenees and descriptions of itineraries in Toulouse. We think that this provides sufficient justification for including movement verbs and spatio-temporal information in general within the specification of a SpatialML or rather its fusion with TimeML.

## 1 Introduction

Many texts are full of spatial information, descriptions of itineraries, trajectories and locations. This spatial information is, however, very often bound together with temporal information, in particular through the descriptions of movement both at the lexical, clausal and discourse levels. The fact that this is so is made particularly clear in certain texts, which include descriptions of itineraries. We argue that spatio-temporal primitives are crucial in giving a full view of the spatial and temporal structure of texts. We think that temporal and spatial structure are projections of a more complex and more complete spatio-temporal structure. We will make our case based on the an analysis of movement verbs, showing how they contribute in an important way to both temporal and spatial structure within discourse. We think that this provides sufficient justification for including movement verbs and spatio-temporal information in general within the specification of a SpatialML or rather its fusion with TimeML (Pustejovsky et al., 2005).

Our analysis of movement verbs is based on a detailed lexical semantic analysis of a wide class of verbs in French. We give some ideas for how this lexical semantics when coupled with an analysis of how clauses involving these expressions are related to each other within a discourse using rhetorical relations can aid in determining the spatio-temporal structure of the text. We have assembled a small corpus involving descriptions of climbing cliffs as well as descriptions of walking tours in the Pyrenees and descriptions of itineraries in Toulouse (Prvot, 2004), from which we will draw certain illustrative examples to support our thesis.

One of our principal aims is to add to the annotations proposed for space parameters or primitive s relevant to encoding motion. In SpatialML[1], the spatial information that is encoded is almost all static, except for the PATH elements.

The vast majority of motion verbs, for instance, indicate a spatial trajectory through time: if we want to know for instance the position of an object at a certain time given the information within a particular text, we often have to know what motions it has undergone. Consider the following example

(1) Laisser la voiture au parking de Sinsat et prendre le sentier du rocher école. Continuer après le secteur ”de la dalle”, vers le secteur ”du lac” qui surplombe l’Ariege. (*Escalade en Haute Ariege*,

[1]SpatialML: Annotation Scheme for Marking Spatial Expressions in Natural Language, March 30, 2007, Version 1.0, http://kent.dl.sourceforge.net/sourceforge/spatialml/SpatialML-1.0-March30-2007.pdf

Thierry et Colette Pouxviel, Publications Sicre, 1993)

This text involves a sequence of ”instructions” of how to get to different sectors of a climbing cliff. *Laisser* is a verb that implies a leaving of the parking lot, while *prendre* is a verb that tells us the direction of the itinerary from the partking lot. It does so by locating the movement along or within another location, a path, *le sentier du rocher école*. The instructions following are sequenced together to give a narrative of how one proceeds through time and space to various sections of the cliff. This is a typical description of an itinerary and one couldn't begin to separate out purely spatial information from the temporal information. For instance, there are no directions given, no distances given. There is a path that one follows. But that is all that is needed to figure out the geographical site of the climbs. This is just one instance of why we think that an annotation scheme for texts and a conception of spatial information within a text should not separate spatial from temporal information. In particular, temporal information can often organize spatial information, as for instance in a description of a walk-through of an apartment, or of an itinerary to a climbing spot.

## 2 Previous work

In previous work, we have worked on the spatiotemporal information encoded in verbs (Asher and Sablayrolles, 1995; Muller and Sarda, 1997)), as well as on how discourse structure conveys spatio-temporal information through the use of discourse relations (Asher et al., 1995; Prvot, 2004). We have based our work on formal investigations of topological information encoded in prepositions (Vieu, 1991) and verbs (Muller, 2002), for which we were able to provide a complete axiomatization (Asher and Vieu, 1995). Geometrical information was a lot harder to axiomatize; most extant attempts in the AI field to provide axiomatizations of geometrical information fail to preclude completely unintuitive interpretations within the natural numbers (see the thesis of (Donnelly, 2001) for some telling examples of how badly various proposed axiomatizations have fared in capturing the intended model $R^3$).

An assumption underlying this work was that lexical semantics as well as discourse information provided spatio-temporal information and needed to be integrated to provide a correct analysis of spatio-temporal structure in text (Asher et al., 1995). Recently, we have begun to annotate corpora for discourse structure in a large scale effort to examine empirically the effects of discourse structure in a variety of domains (anaphora resolution, temporal structure of texts, evaluating opinions in texts, *inter alia*). We can add spatio temporal structure to that list of effects that we would like to study. We think that these texts amply support the idea of encoding spatio temporal information, in particular the information encoded in movement verbs, in any attempt to get at the spatial information expressed in the text.[2]

## 3 Our corpus

Our corpus includes climbing guide texts, texts on ski randonnée outings and mountain biking guide books to various areas in the Pyrénées. We also have access to a number of descriptions by famous and not so famous authors of their journeys through the Pyrenees from the Mediathèque in Pau used in GIS project described in (Lousteau et al., 2008). The climbing guides have short to medium descriptions of situations of cliffs and the climbs or boulder problems on them. They contain some straightforward geographical information and well-known towns or location. They are also usually well laid out with subsections that give rise to discourse structure that can be easily captured in automatic fashion. In this structure each subsection elaborates on its parent. Each subsection includes graphics or text and usually important spatial information. For example, a typical climbing guide presents a site Arabaux by first giving its geographical location and then goes on to describe the various sectors of the cliffs. It begins in the following way:

(2) A 3km au nord-est de Foix, le petit village d'Arabaux est dominé par plusieurs barres calcaires juxtaposées. Celles-ci proposent dans leurs parties centrales un fabuleux potentiel de blocs. Dans les années 70 et 80, plus de 150 passages existaient. Il s'agissait souvent de blocs hauts pratiqués en moulinette. Les plus beaux ont ensuite été équipés... (Jean Denis Achard *Escalades en Ariège: Le Plantaurel*, Lavelanet: Noisetier, 2000)

It is not completely straightforward to isolate the spatial information. Here the first sentence uses a frame adverbial to situate the village of Arabaux, which is dominated by ”several limestone cliffs”. An anaphor then links the cliffs ”in their central parts” to the ”fabulous potential of boulder problems” which the text then goes on to give a historical background to.

The next section gives directions to the site and the section 'Acces aux voies' directions to the different areas of the site. These directions resemble those in (1) and use movement verbs to provide information relevant to the spatial location of objects. The sections on each site give physical details of the different routes or problems in the site. Some guidebooks give comments on the difficulty or the type of climb, and some give important information

[2]This conclusion is also supported by the work in (Lousteau et al., 2008).

about how to climb or do the particular problem which may important spatial information (start to the left of the big boulder).

The description of itineraries using motion verbs and temporal adverbials having a spatio-temporal usage is common not only in the climbing guides but in the ski guides and others. Here is an extended example describing a complete itinerary to the top of a mountain, the Mont Rouch, a difficult ski tour in the Pyrenees.

(3) Suivre le sentier balisé (jaune) qui remonte la vallée de Leziou rive gauche dans la forêt. Apres une montée raide, on débouche sur le plateau de Leziou (1662m) direction sud. On peut remarquer sur la gauche une cabane de berger; continuer sud jusqu'a l'altitude de 1930m puis obliquer en direction des Clos de Dessus. Plateau idéal pour bivouaquer ou dormir à la ouvelle cabane 4 places située en contrebas du plateau. Continuer à se diriger plein est sur la rive droite du ruisseau, laisser les skis et gagner la crête orientée sud pour attendre l'arête frontièrequi mène au sommet. (Daniel Daubin, Michel Dedieu, *Cent Randonnées à Ski en Ariège, Andorre, Pyrénées Orientales*, Randonnées Pyrénéennes, 1992)

Other texts in our corpus are narratives of journeys taken through the Pyrenees. Much less structured, they nevertheless exhibit some of the same tendencies. They do not have a wealth of precise spatial information and often use temporal information to situate the journey. About 30% of the temporal adverbs in those texts have a spatio temporal use of the sort explored in (Vieu et al. 2005).

## 4 A word on semantic types

Prior to our semantics, we need to think a bit about ontology. Like SpatialML we think it important to make a distinction between places or locations (fixed elements in the terrestrial reference frame) and objects (elements that have a complex internal structure and typically move with respect to the terrestrial reference frame). (Asher, 2007) argues that a failure to keep the types of object and place distinct will lead to difficulties in formulating relations of inclusion for spatial prepositions like *dans* ((Vieu, 1991). It also appears that a failure to distinguish between objects and places will miss grammaticalizations of these categories. For instance, in Basque there are two genitive cases *-ko* and *-ren* and they have a quite interesting distribution, once one distinguishes between geographical locations and objects; locations in general easily take the genitive *-ko* but not *-ren*, whereas objects in general do the reverse (Aurnague, 2004).

(Asher, 2007) proposes a possible test for the distinction between an object and a location using the alternation in English *in* or *at* versus *inside*. One can easily say that one is at or in a location. One can also be *in* or *inside* a physical object. On the other hand, it is dispreferred to say that something is inside a location but quite alright to say that it is inside an object, if it's enclosed. Similarly, the relative pronoun *where* refers to locations rather than physical objects.

(4) a. *The worm is inside /in the apple. ??The apple is where the worm is.*
b. *John is in/?inside New York. New York is where John is.*
c. *The tractor is in/??inside the field. The field is where the tractor is.*

Given the conceptual and grammatical reasons for making the distinction between places and object, it's very surprising to note, as (Aurnague, 2004) does, that some lexical items appear to act both like objects and like places. Aurnague calls these "mixed entities". Mixed entities are things like buildings; they have a complex internal structure like other movable objects but which are also fixed elements with respect to the terrestrial reference frame. Thus (Aurnague, 2004) distinguishes the following:

- places: *valley, field, river, mountain, hill* ...
- objects: *apple, glass, chair, car* ...
- "mixed entities": *house, church, town hall* ....

Using the grammatical clues given in Basque, Aurnague suggests that a mixed entity noun (as *castle* in the example below) functions both as an object and a place.

(5) *Gazteluko paretak harriz eginak dira, haren dorre zaharra aldiz egur eta buztinez.* ('The walls of the castle are made of stone, its old tower however (is made) of wood and clay').

If we attend to the distribution of *in* and *inside* in English with objects that have an inside and use that as a key to distinguishing between the way their objects are typed, then it appears that houses, as well as trains and kitchen drawers can be understood not only as physical objects but as locations as well. Cities can be understood as locations but also as many other things—political organizations and even physical objects as well

(6) a. *The checkbook is inside the drawer. The drawer is where the checkbook is.*
b. *I'm inside a train where there are some very comfortable seats.*
c. *John must be inside the house where there are some very expensive paintings.*
d. *There are some beautiful paintings inside the house where John resides.*

A careful study of ontological categories relative to the spatial domain reveals both a distinction between locations, objects and mixed entities and various means for shifting from one type to the other. More important, for our purposes, is to single out among spatiotemporal entities the usual temporal enttities (eventualities, dates and times) and to also signal that many eventualities typically have a spatial as well as a temporal dimension. We will focus on a particular subtype of eventuality conveying important spatiotemporal information, movement eventualities, in the next section

## 5 Mouvements and the verbs that express them

Our lexical semantics taken from (Asher and Sablayrolles, 1995) and (Muller and Sarda, 1997) comprises an exhaustive list of transitive and intransitive motion verbs in French (about 400 in all), which we classify into:

- change of location verbs which are arranged into 10 general types according to the type of motion involved.
- change of position verbs (within a given location) (.e.g, *circuler, parcourir, sillonner* (*circulate*)
- inertial change of position verbs (within a given location) (e.g., *courir*, *danser*, (*run, dance*) etc.[3]
- change of posture verbs (*s'assoir*, *se lever* (*sit down, stand up*).

Some distance information is also encoded to describe the motion. In addition to the temporal prepositions that can also have a spatiotemporal use, we draw on the classification 189 prepositions in French having an almost exclusive spatial use organized into 16 general types. With each general type we associate a particular feature structure that specifies the verb in terms of its "polarity," the relation it evokes within the background mereotopological framework developed by (Vieu, 1991) and extended to space-time by (Muller, 2002), and whether or not it is a telic verb. Here for instance is an example of a transitive verb from the initial, telic internal verb class:

$$\begin{bmatrix} \textbf{quitter} & \\ \text{Event\_Str} & \begin{bmatrix} \text{event} & \\ \text{target: 27} & \\ \text{landmark: 28} & \\ \text{process:} & \text{quitter'} \end{bmatrix} \\ \text{Mvt\_Str} & \begin{bmatrix} \text{polarity} & \text{initial} \\ \text{loc. rel.:} & \text{internal} \\ \text{telicity} & \text{true} \end{bmatrix} \end{bmatrix}$$

[3]It should be noted as we have done in previous work that these verbs in French behave quite differently from the way manner of motion verbs in English are described, e.g., by Beth Levin.

Transitive movement verbs in general make their subjects be the target. Initial transitive verbs specify the landmark as the *source* via their direct objects or internal arguments, while final transitive verbs specify the landmark as the *goal* or end location of the movement via their direct objects. The internal argument of a median verb specifies a path argument, a location at which the target is located throughout the movement. The feature *internal* describes a relation between the target and source or initial point of the movement that holds at the beginning of the displacement and no longer holds at the end of the event. Besides the internal relation, Muller and Sarda as well as Sablayrolles and Asher and Sablayrolles provide other relations describing topological relations between the target and the landmark; these topological relations are the basis of the taxonomy of transitive verbs made by Muller and Sarda. Besides these topological relations, Sablayrolles appeals to zones which incorporate some crude distance information into the lexical semantics of movement verbs. For example, according to Sablayrolles, the French verbs *arriver* and *entrer* are both final, telic, internal verbs, but *entrer* suggests that the displacement originates from somewhere not inside but close to the goal location, whereas *arriver* does not have this meaning. We would represent this information within particular spatio-temporal relations that incorporate geometrical as well as topological information.

Like SpatialML we want to include as features the other PATH attributes: Direction, Distance and Frame. These can also contribute to the specfication of a verb class. Verbs like *olbliquer* (*proceed at an angle from one's present direction*) would specify a value for the Direction attribute.

Coercions or cocompositions in the sense of (Pustejovsky, 1995) may turn certain verbs into movement verbs—for instance, when a location is given as their internal argument or external argument. For instance, the verb *suivre* (*follow*) when it takes a location or a sign to some location is clearly a movement verb. *Mener* (*lead*) is a movement verb when its subject is a location (a path).

## 6 How we would go about acquiring spatio temporal information automatically or semi-automatically

Our detailed lexical semantics is nice and we think that we need these features in any adequate coding of spatiotemporal information. From the standpoint of automatic construction of these feature structures, however, we think that we need to pay attention to how discourse structure interacts with lexical semantics. Below, we will sketch an approach to text annotation in which lexical information, compositional semantics as well as discourse provide information crucial to the automatic annotation of spatial temporal information. With a good syntactic parser, we can capture relatively reliably argument structure of motion verbs, as well as sentential spatial IP adverbs or frame adverbials which have been argued to be important discourse devices and which (Vieu et al., 2005) have shown how to integrate within a framework of formal discourse interpretation. We hope to use such a syntactic parse together with manual annotations to arrive at a corpus from which we can induce a discourse parser. Having tried already on open domain texts (DISCOR, NSF project IIS-0535154), we know that the problem of induction rears its head for us in the form of sparse data: we need to annotate a lot of texts to have a decent discourse parser. One hope we have is to use more symbolic means to compute discourse structures. This seems to be not feasible for the moment in open domain texts with a general discourse structure annotation scheme. But by restricting our attention to a certain type information like spatio-temporal information, we hope to be able to have a more tractable task.

Rhetorical structure is an important element in understanding the spatio temporal information conveyed by a text. Together with compositional semantics, it tells us how to integrate the information given by lexical elements. Discourse relations indicate how to string together bits of spatiotemporal information into trajectories. Take for instance (3). Practically each clause therein provides a displacement from one position to another, but they are linked in a narrative sequence. If we use the axioms of Asher and Lascarides (2003) for Narration, we can link these together to get a trajectory of the author to the top of the mountain.

To go into just a few details, we need to say a bit about our model of discourse structure. We first isolate the basic units with which we will associate a feature structure involving one or more targets, a source a path and a goal (these may be empty if the segment does not contain any spatiotemporal information). Discourse relations manipulate or help us link these feature structures together. Narration, for instance, tells us that the goal of the feature structure of its first constituent should be identified with the source of the feature structure of its second constituent.[4] There are also several types of Background relations. The relation of $S-$Background says that the eventuality described by the second argument spatio-temporally overlaps the location of the object denoted by the NP in the first constituent that the relative clause or modifier expressing the $S-$Background modifies. Thus, for the first two sentences of (3), we would get 4 basic segments, which we label here:

(3') [Suivre le sentier balisé (jaune)]$_1$ [qui remonte la vallée de Leziou rive gauche dans la forêt.]$_2$ [Apres une montée raide,]$_3$ [on débouche sur le plateau de Leziou (1662m) direction sud]$_4$.

We have for this part of the text: $S-$Background(1,2), Narration(1,3), Narration(3,4). Some of the parameters are not specified completely. Nevertheless, the combination of syntax, lexical semantics and discourse structure tells us quite a bit about the spatio-temporal structure of this text. Abstracting away from the specific details of the exact spatio-temporal relations conveyed by the prepositions here we will concentrate on verbs; the first and third verbs in the sequence specify a contact relation between target and landmark; the second specifies an internal relation relation while also conveying, along with the third verb, a certain directionality—up. The last specifies an external spatio-temporal relation between target and landmark. From this lexical information together with axioms about discourse structure, we can infer the following:

- Path$_1$ = le sentier balisé (jaune)
- $e_1$ and target spatially included within Path$_1$ at the time specified by eventuality $e_1$.
- Source$_2$ = Path$_1$ (temporally unrestricted)
- Target in contact with Path$_1$ during eventuality $e_1$.
- Source$_2$ included in Path$_2$ (temporally unrestricted).
- Path$_2$ = la vallée de Leziou rive gauche dans la forêt.
- Goal$_1$ = Source$_3$ at the temporal onset of $e_3$ (after $e_3$)
- Goal$_3$ = Source$_4$; target located there after $e_3$
- Goal$_4$ = le plateau de Leziou (1662m).
- Source$_4$ = Goal$_4$ (temporally unrestricted).
- target located at plateau de Leziou after $e_3$

We can thus follow the target's trajectory after each one of these events, and we can in principle answer queries like 'Where am I after I've climbed up the steep section?" Much more in terms of inference can be done here, given that we can link these surface features with the axiomatic mereo topology (Gerevini and Nebel, 2002; Wolter and Zakharyaschev, 2000; Yaman et al., 2004).

Beside S-Background and Narration, other discourse relations like Precondition, Explanation, Result and Elaboration, have spatiotemporal consequences. We ignore other so called structural relations, featured in many theories of discourse analysis (Asher and Lascarides 2003). We plan to study the spatio-temporal consequences of these relations.

[4] This is a slight simplification of the rules in (Asher and Lascarides, 2003).

As within TimeML, we think it important to take account of modals, disjunctions, conditionals and negations in processing spatio-temporal relations conveyed by a text. Negations will be treated eliminating spatio-temporal information in their scope. Conditionals, disjunctions and various modal operators affect the status of the information within their scope as well.

## 7 Conclusion

Our detailed work on the lexical semantics of motion verbs and prepositions leads us to believe that the annotation of spatiotemporal information is crucial to understanding the spatial information in a text. While very much in the programmatic stage, we feel that we are close to having the tools needed to induced much of this spatiotemporal structure automatically. But only time will tell whether the ideas sketched here will bear the fruit we hope they will.

## References


Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

N. Asher and P. Sablayrolles. 1995. A typology and discourse semantics for motion verbs and spatial PPs in French. *Journal of Semantics*, 12(1):163–209, June.

N. Asher and L. Vieu. 1995. Towards a geometry of common sense: a semantics and a complete axiomatisation of mereotopology. In *Proceedings of IJCAI95*.

Nicholas Asher, Michel Aurnague, Myriam Bras, Pierre Sablayrolles, and Laure Vieu. 1995. De l'espace-temps dans l'analyse du discours. *Semiotique*, 9:11–62.

N. Asher. 2007. Objects, locations and complex types. In M. Aurnague, M. Hickmann, and L. Vieu, editors, *Categorization of Spatial Entities in Language and Cognition*, pages 317–361. John Benjamins.

M. Aurnague. 2004. *Les structures de l'espace linguistique : regards croiss sur quelques constructions spatiales du basque et du franais*. Number 56 in Bibliothque de l'Information Grammaticale. Peeters.

M. Donnelly. 2001. *An Axiomatization of Common Sense Geometry*. Ph.D. thesis, University of Texas at Austin, Austin, TX.

Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, editors. 2004. *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004), Whistler, Canada, June 2-5, 2004*. AAAI Press.

A. Gerevini and B. Nebel. 2002. Qualitative spatio-temporal reasoning with rcc-8 and allen's interval calculus: Computational complexity. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'02)*.

P. Lousteau, T. Nodenot, and M. Gaio. 2008. Spatial decision support in the pedagogical area: Processing travel stories to discover itineraries hidden beneath the surface. In *11th AGILE International Conference on Geographic Information Science*, Girona, Spain.

P. Muller and L. Sarda. 1997. The semantics of french transitive movement verbs and the ontological nature of their objects. In *Proceedings of the International Colloquium of Cognitive Science (ICCS'97)*, Donostia-San Sebastian.

P. Muller. 2002. Topological spatio-temporal reasoning and representation. *Computational Intelligence*, 18(3):420–450.

Laurent Prvot. 2004. *Structures smantique et pragmatiques pour la modlisation de la cohrence dans des dialogues finaliss*. Ph.D. thesis, Universit Paul Sabatier.

James Pustejovsky, Robert Ingria, Roser Sauri, Jose Castano, Jessica Littman, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language TimeML. In I. Mani, J. Pustejovsky, and R Gaizauskas, editors, *The Language of Time: A Reader*. Oxford University Press.

J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

L. Vieu, M. Bras, N. Asher, and M. Aurnague. 2005. Locating adverbials in discourse. *Journal of French Language Studies*.

L. Vieu. 1991. *Sémantique des relations spatiales et inférences spatio-temporelles : une contribution à l'étude des structures formelles de l'espace en Langage Naturel*. Phd thesis, Université Paul Sabatier, IRIT, April.

F. Wolter and M. Zakharyaschev. 2000. Spatio-temporal representation and reasoning based on rcc-8. In Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman, editors, *Proceedings of the Conference on Principiles of Knowledge Representation and Reasoning (KR-00)*, pages 3–14, S.F., April 11–15. Morgan Kaufman Publishers.

Fusun Yaman, Dana S. Nau, and V. S. Subrahmanian. 2004. A logic of motion. In Dubois et al. (Dubois et al., 2004), pages 85–94.

# Reusable Grammatical Resources for Spatial Language Processing

**Robert J. Ross**

SFB/TR8 Spatial Cognition Research Centre
Bremen, Germany
robertr@informatik.uni-bremen.de

**Abstract**

While modularity and resource reuse are key to rapid development in the software engineering community, the emphasis on re-use has not been equalled in components which provide linguistic interfaces to spatial applications. As a step towards reducing this problem, we demonstrate open-source grammatical resources which we are developing for applications which require a clear and controllable spatial language interface. The grammars, based on Steeedman's Combinatorial Categorial Grammar (CCG), can be used for both language analysis and production in the German and English languages, and have been fused with a rich spatial semantics based on a well-defined linguistic ontology. We demonstrate the use of the grammars themselves through the publicly available OpenCCG tool, and also illustrate their use in an implemented dialogue based spatial language processing application for navigating agents.

## 1. Introduction

Within the software engineering community, modularity and reuse play a pivotal role in the rapid development of rich applications by reducing time to market by allowing developers to focus on application specific issues. As the prominence of spatial language applications grow, it is crucial that we apply the same re-use methodologies. One area where this is possible is in the grammatical resources which, in the broadest sense, provide the mapping between surface spatial language and logical form.

Growing consensus in the linguistics and language technology community regarding language resource organization means that we can replace ad-hoc language technology solutions with reusable grammatical resources. However, care must be taken to ensure that the semantics interfaces to such grammars are suited for reuse in a range of domains and are not overly application dependent. In this short paper we will describe grammatical resources which we have been developing to meet requirements of re-usability, tractability, accuracy, and compliance with a well-defined semantics which provides detailed spatial information without succumbing to application dependence.

## 2. Generic Grammar Interfaces

In many applied spatial language systems there exists an implicit assumption that the grammar interface is defined directly in terms of the same types used for application specific reasoning. While such a view is appealing for its simplicity, it unfortunately belies both the complexity of spatial language, spatial reasoning, and the relationships between the two. An alternative view is to subscribe strictly to a "two level semantics" view of knowledge representation within a spatial language system, within which the first level or "Linguistic Semantics" captures the direct surface meaning of spatial language, and a secondary application specific "Conceptual Semantics" captures the application's own spatial knowledge and reasoning process. A mapping then exists between the two, with the complexity of the mapping being a function of the particular domain knowledge organization.

While the linguistic and philosophical reasons for such a distinction have been discussed at length elsewhere [Farrar and Bateman, 2004], here we are particularly motivated by factors related to the practicalities of system design and the features of spatial language. Most practically, it can be virtually impossible to ensure that the organization of world knowledge used within a domain application maps to the organization of world knowledge assumed by grammars of language analysis or production. If either the grammars to be used or the systems's knowledge pre-exist, then merging them as part of the development process could be a futile endeavour; it is better instead to clearly distinguish the two concerns. This is particularly relevant to the development of language interfaces for spatial systems where the types of representation and reasoning used by applications such as robots or spatial information systems can be fundamentally different to the types assumed by the descriptions of space used in natural language.

Moreover, we believe that the adoption of a strict two-level semantics follows from current trends in the modularization of language processing systems - where it has been found useful to distinguish between different representation and reasoning logics [Asher and Lascarides, 2003, Dizikovska et al., 2007]. We also see the two-level approach to be particularly motivated by the spatial language systems, where, for example, critical spatial reasoning information such as perspective and frame of reference are typically unmarked or under-specified in surface language. While it is in principle possible for a single grammar to organize completely the transformation between the agent's knowledge and the surface form, intermediate steps of contextualization and language planning allow a cleaner modelling of the various information types and processes necessary to map between surface from and the types used for actual spatial reasoning within our systems.

In summary, to aid re-usability and modularity of resources, and to help cope with the particular differences between expressed surface language and spatial reasoning types, the grammars discussed in this paper adhere strictly to the two-level semantics account.

## 3. The Grammar

As grammatical formalism we have chosen Steedman [2000]'s Combinatorial Categorial Grammar (CCG) due to


Figure 1: A Demonstration Application for the Spatial Grammars. The application allows a human participant to play the role of route giver in directing a simulated robot around a schematized office environment. Communication is text based and entered and read through the bottom portion of the screen.

its semantic transparency in syntactic derivations, its support for incremental parsing; the ability to use the grammar for both efficient language analysis and production, as well as the good tool support available.

CCG's combinatorial mechanisms operate over categories which link form with grammatical and semantic features. This coupling of semantic and grammatical features at the lexical level allows CCG grammars to construct logical forms for a given utterance compositionally and in parallel to structural analysis. While many different logical and functional formalisms can be used for such construction, our adopted CCG framework, i.e., OpenCCG[1], makes us of a syntax-semantic interface based upon Hybrid Logic Dependency Semantics [Baldridge and Kruijff, 2002].

While HLDS gives a basic semantics framework for the grammar interface, we must populate that framework with appropriate types and roles for the description of spatial language. Following our desire to maintain a level of modularity suited to grammar re-use across applications, the grammars' semantics interfaces are defined in terms of the Bateman's Generalised Upper Model [Henschel and Bateman, 1994], a "Linguistic Ontology" or formal theory of the world whose categories are motivated based on the lexicogrammatical evidence of natural language.

The grammars described here have been built around the latest version of the Generalised Upper Model (GUM), which has been specifically extended to provide a comprehensive account of the natural language semantics of spatial language. Those extension, described in detail by Bateman et al. [2008], are rooted in the traditions of formal spatial language semantics (e.g., [Eschenbach, 1999]) and more descriptive accounts of spatial phenomena in language (e.g., [Levinson, 2003]), resulting in category types which are wholly motivated by the distinctions made by language in its construal of space. To illustrate the semantics interface, and hence the output of our grammar, we give a short locative expression below along with a simplified frame view of the semantics produced and accepted by our English grammar:

1. The river surrounds the town

```
(SL1 / SpatialLocating
  process   (b1/Being)
  locatum   (b2/River)
  placement (GL1/GeneralisedLocation
    hasSpatialModality (S1/Surrounding)
    relatum  (T1/Town)))
```

where `Being` is a sub-concept of the GUM `Process` category, `River` and `Town` are sub-concepts of the GUM `SimpleThing` category, and the categories `SpatialLocating`, `GeneralisedLocation`, and `Surrounding`, as well as all relations, are provided by GUM directly.

We have focused development efforts on two classes of spatial language which we see as having particular usefulness across a range of applications - namely: locative expressions, and motion processes. We have not considered language types used to describe more general physical properties of objects and sets such as shape or size attribution. The grammars themselves, one English and one German, and both under regular development, are available for free download, use, or extension[2]. Since OpenCCG provides both a language realizer as well as a language analysis tool, the grammars can be used for both language production and interpretation, but it should be noted that when enhanced control of features in production is required, we see the CCG grammars as complimenting dedicated production oriented grammatical formalisms, e.g., Bateman [1997].

[1] http://openccg.sourceforge.net/

[2] http://www.diaspace.uni-bremen.de/twiki/bin/view/DiaSpace

## 4. Demonstration Description

Our spatial grammars have been developed for applications which involve the negotiation of spatial information in both information-seeking and robot control applications. Figure 1 depicts our "route interpretation" system which will be used as a focal point for our grammars' demonstration. During the demonstration we will:

i discuss the availability of existing open-source tools for applying and testing these grammars;

ii illustrate a broader set of language examples to demonstrate the coverage of our grammars;

iii demonstrate libraries of functionality which facilitate grammar interface mapping within systems such as the demonstration application.

## 5. Application & Evaluation

In addition to being used in the demonstration application mentioned above, earlier version of our spatial language grammars and semantic resources have been used to interface with a robotic wheelchair which uses a substantively different model of spatial representation and reasoning to that used in our toy system [Mandel et al., 2006]. Moreover, our spatial language resources are also currently being applied to the development of a dialogic wayfinding assistance system being developed within our research group. Critically, we see such applications as being both validation of the re-usability of the application independent but non-trivial grammars described here, but also as a necessary source of feedback which allows application independent resources for spatial language to be more widely deployed.

Additionally, each of our grammars is being constructed and evaluated against test beds which include a range of spatial expressions taken from diverse sources including existing spatial language corpora, the coverage of the spatial semantics model described by Bateman et al. [2008], and the needs of our own test applications. While the coverage of our hand-crafted grammars is naturally significantly smaller than wide-coverage semantics producing grammars, e.g., Bos et al. [2004], we have found that the quality or accuracy of our grammars in terms of spatial language to be considerably more reliable than statistical parsers which neither provide a detailed account of spatial meaning, nor do they take semantic constraints into account in the parse process.

## 6. Outlook

We conclude that while our hand-crafted grammars provide modest coverage, the produced semantics are accurate, spatially rich, and we believe that the construction of such reusable resources represents an important starting point for re-use in spatial language interpretation.

## Acknowledgements

We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Centre on Spatial Cognition (SFB/TR 8) for the work reported in this paper. We wish to thank Prof. John Bateman and the anonymous reviewers for their helpful comments on an earlier version of this paper. Finally, we also extend our thanks to Niels Schütte and David Mautz for their work on the German grammar.

## References

Nicholas Asher and Alex Lascarides. *Logics of conversation*. Cambridge University Press, Cambridge, 2003.

Jason Baldridge and Geert-Jan Kruijff. Coupling CCG and hybrid logic dependency semantics. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Philadelphia, Pennsylvania, 2002.

John Bateman, Joana Hois, Robert J. Ross, and Thora Tenbrink. The Generalized Upper Model 3.0 - Space Components. Technical report, Collaborative Research Center for Spatial Cognition, University of Bremen, Germany, 2008.

John A. Bateman. Enabling technology for multilingual natural language generation: the KPML development environment. *Journal of Natural Language Engineering*, 3 (1):15–55, 1997.

Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. Wide-coverage semantic representations from a ccg parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, 2004.

Myroslava O. Dizikovska, James F. Allen, and Mary D. Swift. Linking semantic and knowledge represenations in a multi-domain dialogue system. *Journal of Logic and Computation*, 2007.

Carola Eschenbach. Geometric structures of frames of reference and natural language semantics. *Spatial Cognition and Computation*, 1(4):329–348, 1999.

Scott Farrar and John Bateman. Linguistic Ontology Baseline. SFB/TR8 internal report I1-[OntoSpace]: D2, Collaborative Research Center for Spatial Cognition, University of Bremen, Germany, 2004.

Renate Henschel and John A. Bateman. The merged upper model: a linguistic ontology for German and English. In *Proceedings of COLING '94*, volume II, pages 803–809, Kyoto, Japan, August 1994.

Stephen C. Levinson. *Space in language and cognition: explorations in cognitive diversity*. Cambridge University Press, Cambridge, 2003.

C. Mandel, U. Frese, and T. Röfer. Robot navigation based on the mapping of coarse qualitative route descriptions to route graphs. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, 2006.

Mark J. Steedman. *The syntactic process*. MIT Press, Cambridge, Massachusetts, 2000.

# Real-time Path Descriptions Grounded with GPS Tracks: a preliminary report

**Nate Blaylock and James Allen**

Institute for Human and Machine Cognition (IHMC)
Pensacola, Florida, USA
{blaylock,jallen}@ihmc.us

**Abstract**

In this paper, we present a novel method for data collection which produces aligned real-time speech recordings of path descriptions and the corresponding GPS track of that path in the real world. We give a preliminary report on the pilot corpus we have gathered using this method and our initial annotation plans for references to location, orientation and movement within the speech. Using the GPS track of the real path and a GIS database, we plan to annotate spatial references in the corpus with the "ground truth" of objects in the GIS database, or lat/lon coordinates corresponding to referrant object or location. This annotated data will provide a set of natural language descriptions of paths, locations, orientation and movement which can be used for training/testing algorithms for understanding spatial language situated in the real world and aided by a GIS database. We also describe an initial annotation tool we are building that uses Google Earth for visualizing and annotating the corpus.

## 1. Introduction

We are interested in building algorithms that understand natural language (NL) descriptions of spatial locations, orientation, movement and paths that are grounded in the real world. In particular, we are interested in algorithms that can ground these NL descriptions in real-world coordinates and entities by leveraging geographical information system (GIS) databases. Such algorithms would enable a number of applications, including automated geotagging of text and speech, robots that can follow human route instructions, and location pinpointing without the use of GPS.

To aide development of our NL understanding system, we are developing a corpus of natural language path descriptions (as recorded from humans in real-time while they travel the path) where each NL description of location, orientation, movement, and the overall path, are annotated with the actual entity (from a GIS database) or location (lat/lon) that is referred to by that description. This "ground truth" annotation will provide a set of NL descriptions that are marked with the "correct answer", on which we plan to use to test our own GIS-informed NL understanding algorithms.

In this paper, we present a novel method for speech data collection in which subjects describe their path in real time (i.e., while they are traveling it) and a GPS receiver simultaneously records their actual paths. These GPS tracks of the actual path can aide the annotator in determining what GIS entities/locations were meant by each spatial reference. To be clear, the GPS tracks are to help a human annotator with the task of creating the annotated "ground truth" test corpus. GPS information would not help the eventual NL understanding system (which would be meant to geotag spatial references using just the natural language and a GIS database).

Although corpora exist for studying NL path descriptions, we are not aware of any that are bundled with the corresponding GPS track for the paths. In addition, many corpora are not in domains where real-world GIS databases would be useful for NL understanding. For example, in the Map Task Corpus (Anderson et al., 1991), paths described were drawn on 2-D maps of a fictitious world with relatively few landmarks and no streets. The MARCO corpus (MacMahon et al., 2006) describes paths through a 3-D virtual world of indoor corridors. The IBL corpus (Bugmann et al., 2004) contains path descriptions of robot movement through a miniature (fictitious) town model. None of these are directly applicable to GIS databases since each is in a fictitious environment and, with the exception of Map Task, movement on each is on a small scale. The smallest objects in our GIS database (as we will outline below) are at the scale of buildings—thus the scale of the path needs to be on the order of hundreds of meters so that multiple GIS objects might be referenced.

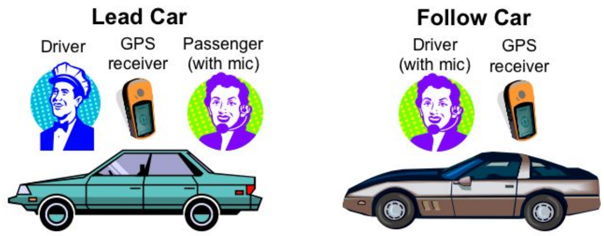


Figure 1: Experiment Setup

In the remaining sections of this paper, we first describe the pilot corpus and methodology. We then outline our plans for annotating it using a GIS database and Google Earth. Finally, we conclude and discuss future work and directions.

## 2. Pilot Corpus and Methodology

In this section, we describe our initial pilot data collection. We will first describe the general setup of the experiment and then the data we have collected thus far.

### 2.1. Setup

The setup of a typical data session is shown in Figure 1. Each session consisted of a lead car and a follow car. The driver of the lead car was instructed to drive wherever he wanted for an approximate amount of time (around 15 minutes). The driver of the follow car was instructed to follow the lead car. One person in the lead car (usually a passenger) and one person in the follow car (usually the driver) were given close-speaking headset microphones and instructed to describe, during the ride, where the lead car was going, as if they were speaking to someone in a remote location who was trying to follow the car on a map. The speakers were also instructed to try to be verbose, and that they did not need to restrict themselves to street names—they could use businesses, landmarks, or whatever was natural. Both speakers' speech was recorded during the session. In addition, a GPS receiver was placed in each car and the GPS track was recorded at a high sampling rate (usually once per second).

The goal of this setup was to yield interesting data for our research goals mentioned above. The paths taken are in the real world and because the movement is in cars, it is on a scale that makes frequent references to entities in the GIS database. Additionally, the recorded GPS tracks provide us with the ground truth about the path taken against which we can compare the results of path description understanding algorithms. Finally, having the recordings of two speakers in two separate cars gives us essentially two different descriptions of the same path at the same time.

### 2.2. Data

In this pilot study, we conducted seven data collection sessions in the downtown Pensacola area with IHMC staff and associated researchers. This yielded 13 audio recordings[1] of seven paths along with corresponding GPS tracks. The average session length was 19 minutes.

The data is still in the process of being transcribed, but the following gives some examples of the contents:

- *...and we're going under the I-110 overpass I believe and the Civic Center is on the right side on the corner of Alcaniz and East Gregory Street where we are going to be taking a left turn...*
- *... he's going to turn left right here by the UWF Small Business Development Center heading toward Gulf Power ...*
- *... we've stopped at a red light at Tarragona Street okay we're going now across Tarragona passing the Music House ...*
- *... we're at the intersection of East Gregory and 9th near a restaurant called Carrabas I think and a Shell station just a little south of the railway crossing ...*

## 3. Planned Annotation

In this section we discuss our initial ideas for annotating the corpus. These ideas will be refined as annotation proceeds. With a GPS track and access to a GIS database, we can annotate the corpus with "ground truth" as to the spatial references which are made as well as the events in the path which are described.

[1]In one session, there was no speaker in the lead car.

We have created an initial visualization tool in Google Earth (shown in Figure 2) which allows an annotator to see the lead car's path and replay its progress along the track in real time. This is more than a video, however, as the annotator can freely zoom and scroll in Google Earth during the track replay. This, synchronized with the speech track, shows the annotator where the car was located when certain utterances were made, giving the annotator a much better context.

The GPS track gives us ground truth for the path described, but it does not give ground truth for most spatial location references used in describing the path (such as businesses, streets, bodies of water, etc.) These we will annotate with objects in a GIS database, instead of just lat/lon coordinates. We have obtained access to TerraFly (Rishe et al., 2005), an industrial-strength GIS system developed by the High-Performance Database Research Center/NASA Regional Applications Center at Florida International University (FIU). TerraFly contains a large aggregation of GIS data from major distributors, including NavTeq and Tiger streets and roads, 12 million U.S. businesses through Yellow Pages, and property lines for the state of Florida. Property lines are useful, as they enable a more accurate placement (geocoding) of the location of street addresses (e.g., from the Yellow Pages). We plan to hierarchically bracket and annotate each spatial location reference as well as movement and orientation events in the corpus. Spatial entities will be annotated with the corresponding object in the TerraFly database.

Figure 3 shows an example annotation for the first utterance shown in the examples above. Here, each segment is assigned a GIS_ID corresponding to the object in the TerraFly database. For example, it is obvious from the GPS track that the phrase "the Civic Center" corresponds to the "Pensacola Civic Center", which is a business in the TerraFly database. The annotation of this phrase is shown in Figure 3a. Similarly, the streets "Alcaniz" and "East Gregory" are found and annotated with corresponding TerraFly objects.

Composite objects not found in the database (such as the intersection in Figure 3e) are built up out of smaller objects (the streets). The two events in this utterance, passing the Civic Center and then turning left, are annotated as shown in Figures 3b and 3f, respectively. Each makes reference to the previous spatial locations as arguments.

## 4. Conclusion

We have presented a methodology and pilot data collection which aligns real-time speech recordings of path descriptions with a GPS track of that path. We also presented an initial tool in Google Earth for replaying that track in real time and our initial plans for data annotation using Google Earth and a GIS database.

For future work, we plan to collect and annotate more data using a similar methodology (as informed by our pilot experience). In addition, we plan to gather data by replaying these tracks using our Google Earth tool described above and having subjects describe the path as it unfolds from an aerial view. In addition to giving us more descriptions of the same path, we believe this will provide interesting com-

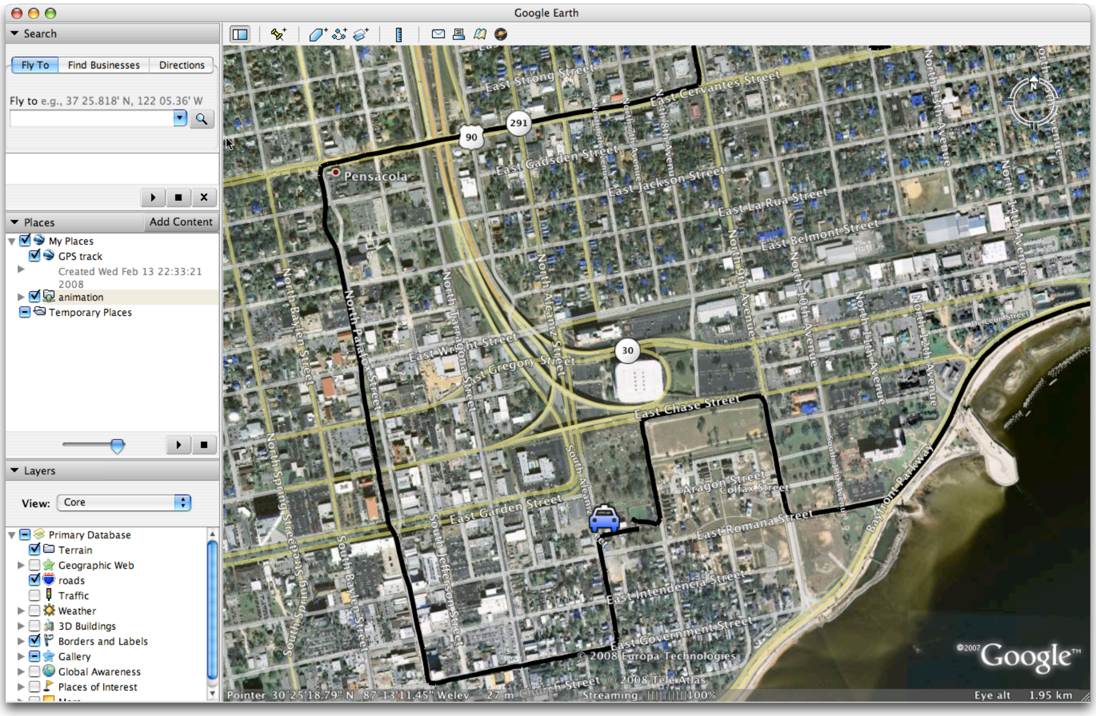


Figure 2: Screenshot of an overall path (in black) and the lead car's "current" position in Google Earth

[1]

| **Business** | |
|---|---|
| `GIS_ID` | B8392 |
| `name` | "Civic Center" |

(a) "the Civic Center"

| **PassLandmark** | |
|---|---|
| `landmark` | [1] |
| `on` | right |

(b) "the Civic Center is on the right side"

[2]

| **Street** | |
|---|---|
| `GIS_ID` | S9323 |
| `name` | "North Alcaniz St." |

(c) "Alcaniz"

[3]

| **Street** | |
|---|---|
| `GIS_ID` | S7723 |
| `name` | "East Gregory St." |

(d) "East Gregory Street"

[4]

| **StreetIntersection** | |
|---|---|
| `street1` | [2] |
| `street2` | [3] |

(e) "the corner of Alcaniz and East Gregory Street"

| **TurnAtIntersection** | |
|---|---|
| `intersection` | [4] |
| `direction` | left |

(f) "the corner of Alcaniz and East Gregory Street where we are going to be taking a left turn"

Figure 3: Example Hierarchical Annotation of the Entities and Events in the Utterance: *"the Civic Center is on the right side on the corner of Alcaniz and East Gregory Street where we are going to be taking a left turn"*

parison data as to how paths are described from a ground view (e.g., from the follow car) versus from an aerial view (e.g., from a following helicopter or UAV).

## 5. References

A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34.

Guido Bugmann, Ewan Klein, Stanislao Lauria, and Theocharis Kyriacou. 2004. Corpus-based robotics: A route instruction example. In *Proceedings of Intelligent Autonomous Systems (IAS-8)*, Amsterdam.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, pages 1475–1482, Boston.

N. Rishe, M. Gutierrez, A. Selivonenko, and S. Graham. 2005. TerraFly: A tool for visualizing and dispensing geospatial data. *Imaging Notes*, 20(2):22–23.

# Building a Parallel Spatio-Temporal Data-Text Corpus for Summary Generation

**Ross Turner*, Somayajulu Sripada*, Ehud Reiter* and Ian P Davy****

*
Dept of Computing Science,
University of Aberdeen,
{r.turner,yaji.sripada,e.reiter}@abdn.ac.uk

**
Aerospace and Marine Intl. (UK)
Banchory, Aberdeenshire, UK,
idavy@weather3000.com

**Abstract**

We describe a corpus of naturally occurring road ice weather forecasts and the associated weather prediction data they are based upon. We also show how observations from an analysis of this corpus have been applied to build a prototype Natural Language Generation (NLG) system for producing road ice forecasts. While this corpus occurs in a narrow domain, it has much wider applicability due to the nature of its spatial descriptions, whose primary communicative goal is to describe the interaction between meteorological parameters and geographic features.

## 1. Introduction

Road ice weather forecasts are an essential aid for road engineers to base their salting and gritting application decisions during the winter months. They are important because consistent unnecessary treatment of a road can be as hazardous as leaving it untreated. Winter road maintenance operations also present a significant cost to local councils in the UK, where a single treatment of a road network can run into tens of thousands of pounds. Modern weather forecasting is driven by Numerical Weather Prediction (NWP) Models. Recent advances in technology have seen road ice models become increasingly localised as meteorological parameters can be measured at very fine grained spatial intervals along a road surface. One side affect of this is that the resultant output of the model becomes increasingly complex for a human expert to analyse and describe in a short textual summary.

Much previous work e.g. (Reiter et al., 2005; Boyd, 1998; Coch, 1998; Goldberg et al., 1994) has shown NLG Systems are particularly well suited for producing such textual reports; therefore, we have been developing RoadSafe, a NLG application for automatically generating road ice forecasts, in collaboration with a local weather forecasting company Aerospace and Marine International (AMI) UK. As part of the knowledge acquisition process for developing this application, we have built a parallel data-text corpus of naturally occurring road ice forecasts and associated input data, described in Section 2.. While extensively used in Machine Translation, there are few examples of using parallel corpora for NLG. (Snyder and Barzilay, 2007) describe an algorithm for automatically aligning textual descriptions to their corresponding database entries. While the Knowledge Acquisition (KA) studies carried out during the design of the SumTime Mousam weather forecast generator made extensive use of a parallel data-text corpus (Sripada et al., 2003).

The main aspect of RoadSafe that differs from other weather forecast generators, is its explicit handling of spatial data. We have been analysing the corpus to understand the process of summarising spatial data as well as understanding both the linguistic and non-linguistic requirements of the system. To this end we have annotated the corpus as described in Section 3.. In Section 4. we describe the corpus analysis process. Section 5. describes how the knowledge acquired from the corpus has guided the design of the RoadSafe summary generator. Finally we summarise our conclusions in Section 6..

## 2. Corpus Description

As stated in the previous section, the RoadSafe corpus is a parallel Data-Text corpus consisting of the output data from a NWP Road Ice Model[1] and the corresponding Road Ice Weather forecast. The corpus was collected between March 2006 and January 2008 for two local councils in England, UK, during their routine winter road maintenance operations, which last between October the 1st and April the 30th each year. The corpus consists of 431 data-text pairs with a of total of 29,857 words.

The model data and texts, described in more detail in Sections 2.1. and 2.2., comprise part of a road ice forecasting service provided by AMI UK. The aim of this service is to provide road engineers with continuous access to up to the minute weather information using various modes of presentation, such as graphs, graphics, tables and text. This information is delivered via a secure website to each council to help them plan how to grit and salt their local road network. To this end, we have developed a system to generate tabular summaries of the road ice model data in HTML format, shown in figure 1, which expert meteorologists at AMI UK can insert manually written textual wind and weather forecast statements into, via an online interface. The inserted text from the HTML files along with a file containing the raw Road Ice Model data, issued daily for both councils, form the basis of the corpus.

### 2.1. Model Data

The NWP data generated by the Road Ice Model is a large spatio-temporal data set (in order of Megabytes depending on the size of the area the model is being run for). The data contains predicted measurements of 9 Meteorological parameters (such as road surface temperature and wind

[1] The AMI UK GRIP model

Issued by Aerospace & Marine International:- Valid from noon on 26/12/2006

| 24 Hour Forecast | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Confidence/Comments | This initial analysis is carried out by the advanced RoadSafe system and is computer | | | | | | | | |
| All Routes | Min RST | Time <= 0c | Ice | Hoar Frost | Snow | Fog | MaxGusts | Rain | TS |
| Worst/Best | -1.1 /1.4 | 21:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/13 | NO /NO | NO |
| Wind (mph) | Light south to south-easterlies for the duration of the forecast period. Winds may become more moderate late morning on higher ground, but remaining southerly. (GT) | | | | | | | | |
| Weather | A mainly cloudy night, with foggy patches across much of the forecast area. Higher ground above the low cloud level could see temperatures drop below freezing during the late evening, with most western parts of the forecast area dropping below freezing by the morning. Urban areas are expected to remain marginal throughout the night. (GT) | | | | | | | | |
| Route | All routes summary worst/best | | | | | | | | |
| 1 | 0.4/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 2 | 0.7/2.0 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/10 | NO /NO | NO |
| 3 | 0.5/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 4 | 0.4/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/12 | NO /NO | NO |
| 5 | 0.7/1.9 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 6 | 0.7/2.1 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 7 | 0.9/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 8 | 0.8/2.1 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 9 | 1.4/2.1 | NA/NA | NO /NO | NO/NO | NO/NO | FOG/FOG | 13/9 | NO /NO | NO |
| 10 | 0.8/1.9 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 11 | 0.3/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 12 | -0.8 /1.5 | 22:40 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 13 | 0.1/1.6 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 14 | 0.7/1.7 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 15 | 0.4/1.7 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/11 | NO /NO | NO |
| 16 | 0.3/1.9 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 17 | 0.4/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/9 | NO /NO | NO |
| 18 | 1.2/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/12 | NO /NO | NO |
| 19 | -0.6 /1.2 | 00:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FREEZING | 15/13 | NO /NO | NO |
| 20 | -0.9 /1.7 | 22:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 21 | -1.1 /1.4 | 21:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/13 | NO /NO | NO |
| 22 | -0.3 /1.6 | 03:20 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 23 | 0.3/1.4 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 24 | 0.0 /0.7 | 08:20 /NA | NO /NO | NO/NO | NO/NO | FREEZING/FREEZING | 15/13 | NO /NO | NO |
| 26 | -0.9 /1.2 | 21:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 27 | -0.5 /1.5 | 01:20 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 28 | -0.3 /1.6 | 00:40 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 13/9 | NO /NO | NO |
| 29 | 0.3/1.3 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 30 | -0.2 /1.4 | 04:00 /NA | YES /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 31 | 0.6/1.6 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/11 | NO /NO | NO |
| 32 | 0.3/1.6 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 13/10 | NO /NO | NO |
| 33 | 0.2/1.2 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |
| 34 | 0.4/1.8 | NA/NA | NO /NO | NO/NO | NO/NO | FREEZING/FOG | 15/12 | NO /NO | NO |

Route no. key:-

Green:- RST > 1.5 or RST > 1 and all routes dry.

Amber:- RST 0.5 - 1 or RST 0.5 - 1.5 if road wet or light sleet/snow present.

Red:- Any ice,hoar frost,heavy snow accumulations or RST <= 0.5 or RST < 1 with moderate snow.

Figure 1: Example Human Written Corpus Text and System Generated Table

speed) for several thousand geographical locations along a local council road network. Each geographical location is a point, indexed by a unique id, which ties it to a particular route[2], and a latitude longitude coordinate. For each point, each parameter is calculated at 20 minute intervals throughout a 24 hour forecast period. This means that there are 9 time series consisting of 72 time points associated with each point. An example of a small subset of the raw model data is given in Figure 2.

### 2.2. Forecast Texts

The forecast texts are written by 7 different expert meteorologists employed at AMI UK and consist of two paragraphs: one describing the wind conditions over the forecast area for the 24hr period between midday and midday the following day, and another describing the weather conditions. These paragraphs are generally short: typically around 1 sentence for wind descriptions and around 3 sentences for weather descriptions. The purpose of the texts are to complement the very fine grained details of the Road Ice Model presented in the tabular data with a more general weather overview. Essentially, the table is intended to present the user with worst/best conditions for specific points on each route, while the texts convey higher level information about the general area using spatial information not contained within the model. The table is also used by the experts as one of the resources to analyse the data; along with maps and graphs, and satellite imagery of the local area.

Each text contain a number of spatial, temporal and spatio-temporal descriptions describing the various weather conditions. In particular, the spatial descriptions can vary between texts depending on the road network being forecast

[2]The route number each point belongs to is indicated by the 7th and 8th digits in the ID number. For example, each point in the example in Figure 2 belongs to Route 1

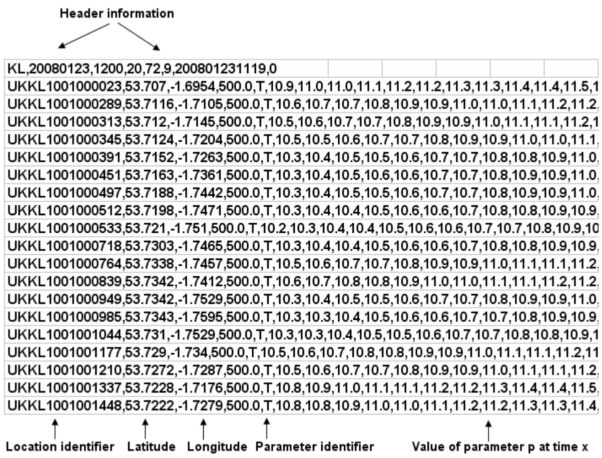


Figure 2: Raw Road Ice Model Data

for. The two road networks the forecast texts in the corpus describe belong to Hampshire and Kirklees councils in England, UK. These two areas of the UK have quite different geographies. The area covered by Kirklees council is in northern England and land locked, with some relatively high ground (above 500m) in the south west of the region. Hampshire is a much larger area, is a relatively flat county and is situated on the south coast of the UK.

## 3. Annotation Scheme

Our interest in using this corpus is to understand the requirements for building a natural language generation system to automatically produce the textual summaries from the input data. Therefore, our annotation scheme was designed to understand the content of the forecasts in terms of its underlying data; in particular, identification of message types, along with the events in the data they convey and their arguments.

Analysing a corpus (in tandem with other KA activities) is a fundamental part of deriving content selection rules when developing a NLG system for a given domain. (Callaway and Moore, 2007) describe a methodology for analysing a tutoring corpus to inform a content planning model of an Intelligent Tutoring System, while (Williams and Reiter, 2005) describe another methodology for selecting content for a system that generates personalised feedback reports. Our domain is quite different to the ones described in those papers. Therefore, the process to extract message types we used in our domain was more similar in spirit to the procedure outlined in (Geldof, 2003), which is applied to route descriptions. This procedure initially segments the corpus into message types, which are further refined on the basis of their occurring arguments.

A message in a road ice forecast typically involves an event, or series of events, that describe the state of a particular parameter or parameters. For example from Figure 1, ‘Winds may become more moderate late morning on higher ground, but remaining southerly’ or ‘Urban areas are expected to remain marginal throughout the night’. An example of an annotated message is shown in Figure 3.

Events normally consist of a verb, describing the general trend or change in state of a parameter, along with an optional spatial phrase and/or temporal phrase describing the location and time of the event. E.g. ‘Winds may become more moderate late morning on higher ground’ in Figure 1. Events can also describe the general qualitative state of a parameter such as the two events that make up the message in Figure 4.

```
<MESSAGE ID="4612" MAINPARAMETER="wind">
 <EVENT ID="7706">
  <W>Winds</W>
  <W>may</W>
  <TREND ID="10380" TREND="increase">
    <VERB ID="13985">
      <W>become</W>
    </VERB>
    <W>more</W>
    <W>moderate</W>
  </TREND>
  <TEMPORALPHRASE ID="2167" TIMEPERIOD="morning">
   <W>late</W>
   <W>morning</W>
  </TEMPORALPHRASE>
  <SPATIALPHRASE AREA="part" ID="3864"
   PRIMITIVE="altitude" TYPE="geofeature">
   <W>on</W>
   <W>higher</W>
   <W>ground,</W>
  </SPATIALPHRASE>
 </EVENT>
 <W>but</W>
 <EVENT ID="7707">
  <TREND ID="13987" TREND="constant">
   <VERB ID="13986">
    <W>remaining</W>
   </VERB>
   <W>southerly.</W>
  </TREND>
 <W>(GT)</W>
 </EVENT>
</MESSAGE>
```

Figure 3: Annotation of the second sentence in the wind statement paragraph in Figure 1

```
<MESSAGE ID="1080">
 <EVENT ID="7708">
  <W>A</W>
  <W>mainly</W>
  <PARAMETER ID="2168" TYPE="cloudcover">
   <W>cloudy</W>
  </PARAMETER>
  <TEMPORALPHRASE ID="2169" TIMEPERIOD="aftermidnight">
   <W>night,</W>
  </TEMPORALPHRASE>
 </EVENT>
 <W>with</W>
 <EVENT ID="7709">
  <PARAMETER ID="2170" TYPE="fog">
   <W>foggy</W>
   <W>patches</W>
  <SPATIALPHRASE AREA="part" ID="4999" TYPE="other">
   <W>across</W>
   <W>much</W>
   <W>of</W>
   <W>the</W>
   <W>forecast</W>
   <W>area.</W>
  </SPATIALPHRASE>
  </EVENT>
 </PARAMETER>
</MESSAGE>
```

Figure 4: Annotation of the first sentence in the weather statement paragraph in Figure 1

Another important aspect of the forecasts and an important requirement of our system, are spatial descriptions. The

corpus contains many such descriptions, which generally do not consist of references to geographical landmarks such as towns or monuments. Instead they are vague, such as 'Most higher level and rural roads'. This is mainly due to the fact that when writing forecasts, forecasters try to avoid use of more specific spatial descriptions unless the pattern in the data is very clear cut, as weather system boundaries are inherently fuzzy. They also try to avoid ambiguity when they may not be aware of more provincial terminology used by road engineers and other users of the forecasts. This is challenging for an annotation scheme in this context as the spatial properties (i.e. altitude,urbanicity) used in such descriptions are not explicit in the input data, unlike temporal phrases whose value can be mapped to actual time values in the data (e.g. 0900 - 1140 for morning). Furthermore, vague concepts such as higher are open to interpretation and are dependent upon how the referent object is characterised (for example, the altitude resolution used).
From our other knowledge acquisition (KA) studies we found, as a first (and one-off) step when writing forecasts, forecasters build frames of reference to salient geographical features in the forecast region. This essentially involves familiarising themselves with parts of the region that may have an affect on the general weather pattern in the area, such as areas of high ground. This was accounted for in our annotation scheme, as spatial phrases were annotated for the spatial reference frame used. After some initial iterations spatial descriptions were classified into combinations of 4 main Frames of Reference:

- Altitude e.g. 'possible gale force gusts on higher ground'
- Direction
  - Absolute e.g. 'minimum temperatures around 5-6 degrees in the more Northern Routes'
  - Motion e.g. 'A band of rain moving across from the west'
- Population e.g. 'many urban routes will drop to be critical but remain above zero'
- Coastal Proximity - e.g. 'a few showers at first mainly along the coast'

In this context, a frame of reference is a particular perspective in which the domain can be observed; moreover, it is a set of related geographical features which allows the domain to be categorised into meaningful sub areas. How the domain is categorised is dependent upon the observer, but any classification should provide the ability to describe the location of domain entities in terms of its interaction with the underlying geography. For example, the population frame of reference is the set of town boundaries contained within the domain and its complement, classifying the domain into two sub areas: urban and rural. Altitude partitions the domain based on elevation at a chosen resolution, while coastal proximity characterises areas in terms of a specified distance from the coast, providing a binary classification: inland and coastal. Direction is based upon partitioning the domain into fixed compass bearings.

# 4. Corpus Anaylsis

Our initial methodology for analysing the corpus was to align individual words and phrases in the texts to specific data points in the input data files as described in (Reiter and Sripada, 2003). This method allows the meaning of meteorological terms to be 'grounded' in the data. However, this proved to be particularly difficult due to the sheer size of the input data and brevity of the texts. Essentially the data/text compression ratio accomplished by the texts is too high to reliably identify which part of the data set a particular phrase describes. Therefore, we had to rely on the data only as a means of comparison during the development of our system. We built a parser to parse the annotated corpus into its constituent parts. As mentioned in the previous section, our analysis described in Sections 4.1. and 4.2., concentrated on understanding the structure of message types and spatial descriptions.

## 4.1. Message Types

**Structure** Our annotation scheme described in the previous section, initially segmented the corpus into messages using full stops as a boundary. Rather than refining the messages further we concentrated on identifying the different types of events in the data that make up the sentence. Thus, in our system a message is a sentence, with event descriptions as clauses of that sentence. Event descriptions are characterised as predicate arguments structures similar to how Message types are defined by (Geldof, 2003). Table 1 shows an example of a corpus text split into events with their relevant arguments. The corpus contains a total of 1749 messages and 3598 events. The mean number of messages per forecast is 4.03 with a standard deviation of 1.165. The mean number of events per message is 2.08 with a standard deviation of 0.59.

**Types and Arguments** Event description predicates can take any number of the arguments outlined in Table 1. The only mandatory argument is the parameter, where no Area or Timeperiod is provided the event description is taken to apply to the whole forecast area and forecast period respectively. Event descriptions may also depict a trend in a parameter, such as a fall in temperature or rise in wind speed, or a parameter remaining around a particular threshold for a substantial period of time. The FrameofReference argument describes the type of location phrase used and can assume the value of any the frames or reference described in Section 3. along with a combination of these or other.
Event descriptions can be split into two very general types: global and local, as denoted by the area attribute. A value of whole indicates the event is global and applies to the whole forecast area, whereas a value of part indicates a local event is being described. After manual inspection of the event texts and their arguments it was possible for event descriptions to be further broken down into 4 types:

1. **Overview(OV)** - Describe a general weather pattern in broad terms. Normally a forecaster has used some kind of domain knowledge to perform some simple interpretation of the data, e.g. 'rain will fall over most routes' or 'Very marginal night'.

| No. | County | Date | Statement | Trend | Timeperiod | Area | FrameofReference | Parameters | MainVerb |
|---|---|---|---|---|---|---|---|---|---|
| 1.1 | KIRKLEES | 26/12/2006 | WINDSTATEMENT | constant | forecastperiod | null | null | wind | |
| 2.1 | KIRKLEES | 26/12/2006 | WINDSTATEMENT | increase | morning | part | altitude | wind | become |
| 2.2 | KIRKLEES | 26/12/2006 | WINDSTATEMENT | constant | null | null | null | wind | remaining |
| 3.1 | KIRKLEES | 26/12/2006 | WEATHERSTATEMENT | null | aftermidnight | null | null | cloudcover | |
| 3.2 | KIRKLEES | 26/12/2006 | WEATHERSTATEMENT | null | null | part | other | fog | |
| 4.1 | KIRKLEES | 26/12/2006 | WEATHERSTATEMENT | decrease | evening | part | altitude | rst | drop |
| 4.2 | KIRKLEES | 26/12/2006 | WEATHERSTATEMENT | decrease | morning | part | absolute | rst | dropping |
| 5.1 | KIRKLEES | 16/11/2006 | WEATHERSTATEMENT | constant | aftermidnight | part | population | rst | remain |

Table 1: Parsed Events Table for the Corpus Text in Figure 1

| No. | Text | Event Predicate |
|---|---|---|
| 1.1 | Light south to south-easterlies for the duration of the forecast period. | TS(Constant(Light,SSE),ForecastPeriod,Wind) |
| 2.1 | Winds may become more moderate late morning on higher ground | ST(Increase(Moderate),Morning,Part,Altitude,Wind) |
| 2.2 | remaining southerly. | TS(constant(S),Wind) |
| 3.1 | A mainly cloudy night | OV(aftermidnight,CloudCover) |
| 3.2 | foggy patches across much of the forecast area | OV(part,other,Fog) |
| 4.1 | Higher ground above the low cloud level could see temperatures drop below freezing during the late evening | NST(decrease(Subzero),evening,part,altitude,RST) |
| 4.2 | most western parts of the forecast area dropping below freezing by the morning | ST(decrease(Subzero),morning,part,absolute,RST) |
| 5.1 | Urban areas are expected to remain marginal throughout the night | NST(constant(Marginal),aftermidnight,part,population,RST) |

Table 2: Corresponding Text Entries and Event Predicates for Table 1

2. **TimeSeries(TS)** - Describe the global state of a parameter during a particular time period. These descriptions are normally derived from a forecaster inspecting time series graphs of the input data, e.g.'Temperatures will drop away quickly into the night' and always.

3. **Stationary(ST)** - Describe events at specific time points in the data. These are normally the first appearance of a particular value of a parameter or condition, e.g. '35-40 gusts 55 mph in exposed places by 0600' or it's disappearance, e.g. 'showers mostly dying out by midnight'.

4. **Non-Stationary(NST)** - Describe local weather conditions developing over a period of time and is typically spatio-temporal, e.g. 'patchy rain spreading from the northwest around midnight'.

The results of this process for Table 1 is shown Table 2.

### 4.2. Spatial Descriptions

Our corpus contains a total of 857 spatial descriptions. Of that total, 662 refer to sub areas of the spatial domain, i.e. they do not refer to the whole area. We found a substantial number of descriptions that were entirely vague; such as 'in most areas','in many places' and 'on most roads' which we classified under other as they are simply expressing proportions which can be inferred from the generated table. We also found that many spatial descriptions often involved using combinations of frames of reference such as 'high ground in the south west', similar to a map overlay operation in a Geographic information system (GIS). Table 3 shows the proportions of spatial descriptions in the corpus based on frames of reference used. As the corpus is unevenly distributed in terms of the counties the texts describe (63% Kirklees and 37% Hampshire), the distribution is skewed towards altitude due to the fact that the dominant geographical feature affecting the weather in Kirklees is an elevated area to the soutwest of the region.

| Dir. | Pop. | Alt. | Coastal Prox. | Other |
|---|---|---|---|---|
| 19% | 5% | 34% | 7% | 35% |

Table 3: Proportions of spatial descriptions by Frame of Reference used

Perhaps most interestingly, we found that certain frames of reference were used more frequently to describe certain parameters or certain events than others. For example, population is used almost exclusively (all but 14% of descriptions using the Population Frame of Reference) in descriptions of road surface temperature; while changes in precipitation type, i.e. rain turning to snow, are mainly described using altitude. These observations agree with our other KA studies we have carried out with the meteorologists at AMI where we have found that the spatial descriptions they make in the forecasts are expressing causality, i.e. the effect that a geographical feature has on a parameter, rather than being purely locative. For example a spatial description such as 'rain turning to snow in the north' may be geographically accurate, but a more useful spatial description would be 'rain turning to snow on higher ground' which also explains the cause of the meteorological event being described. Therefore, the observation of a link between Population and road surface temprature can be explained by the fact that urban roads tend to be warmer than rural roads due to their more frequent use, the population effect and their tendency to be at lower elevations; whereas changes in precipitation type are more commonly seen on higher ground where the air temperature is generally lower. The graph in Figure 5 shows the distribution of spatial de-

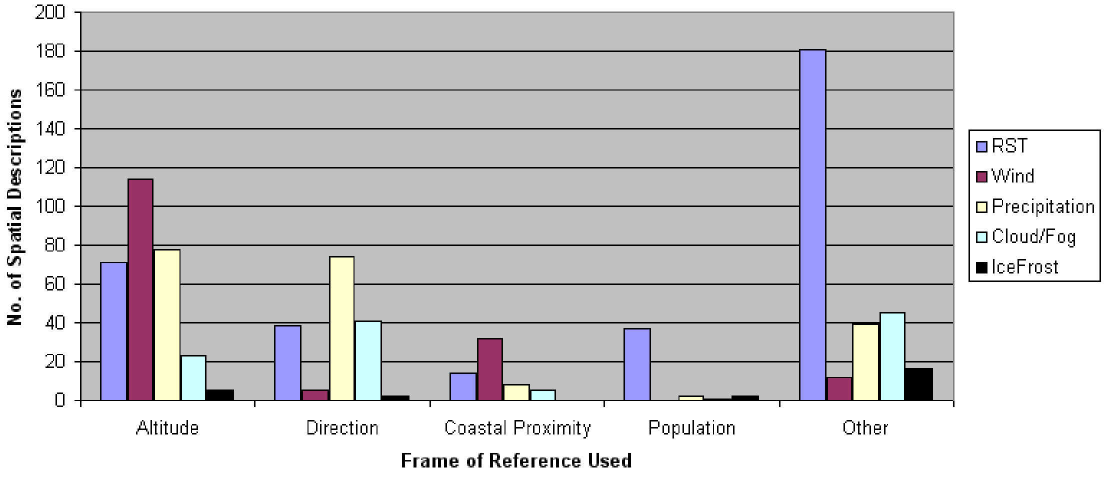


Figure 5: Spatial Description Frequencies by Frame of Reference Used and Parameter Described

scriptions in the corpus. The large number of spatial descriptions describing road surface temperature being classified under other can be accounted for by the fact that descriptions such as ‘on most roads’ and ‘all roads’ were included in this category. While the graph shows a clear preference for describing the location of events affecting wind in terms of altitude, the affect of the dominant geographical features affecting the weather pattern in a region becomes more clear when the distribution is viewed between the two counties. 93% of all spatial descriptions describing wind in Kirklees use the Altitude Frame of Reference as opposed to 30% in Hampshire which incorporates a coastline. Here the location of wind events are mainly described in terms of their proximity to the coast (54%).

From a linguistic perspective, the spatial descriptions in the corpus are relatively simple syntactically. They are normally a single prepositional phrase - ‘on higher ground’ or concatenation of prepositional phrases - ‘in some places in the north’. As would be expected, vague quantifiers such as all,some,most,many and few are fairly commonly used (27% of all phrases). These are used where a forecaster is referring to the proportion of points contained within the bounding area of the event.

The number of variations in lexical choices for referring to values within a reference frame vary between each individual frame in the corpus. Direction values, as would be expected, are referred to using the standard points of the compass; coastal proximity, is also referred to using the standard distinctions between inland and coastal; within population, while urban and rural distinctions are mainly used, urban areas are also sometimes described using near synonyms such as suburb and residential as well as occasionally by proper names. The greatest variation is seen within references to altitudes, where the same elevated area can be referred using a modified noun phrase such as high ground, common nouns such as hills or moors, or its actual height value such as 300m.

## 5. Implications for Summary Generation

For building a prototype road ice forecast generator, it was clear from observing how the input data mapped to the output texts that a number of extensions to the architecture for data-to-text systems proposed in (Reiter, 2007), were required to facilitate the handling of spatial data. We describe these extensions in Section 5.1. and highlight how observations from the corpus analysis influenced Content Section in our system in Section 5.2.. Finally, Figure 8 provides an example of a weather statement generated by the RoadSafe system.

### 5.1. General System Architecture

As mentioned in the previous section, our system extends the general architecture outlined in (Reiter, 2007). The extended architecture used in our RoadSafe prototype, the details of which are to be published elsewhere, contains 3 extensions; a Geographic Characterisation stage, a Spatial Reasoning module and a Spatial Database, which we describe next.

**Geographic Characterisation** Characterisation is defined as finding compact descriptions of data (Miller and Han, 2001). As the only spatial information contained within the input data to our system described in 2.1., is a latitude-longitude coordinate pair, it is necessary to characterise the data in terms of the frames of reference we have identified in the output text. We treat this as a one-off preprocessing step for each new forecast area in the initial data analysis stage of our architecture. This process is a form of data enrichment (Miller and Han, 2001) performed by combining the data with other external spatial data sources. Essentially each point in the input data set is assigned additional spatial properties as shown in Figure 6.

**Spatial Database** The spatial database acts as a store for the external spatial data information used in the Geographic Characterisation step. Frames of reference are stored as thematic layers in the database. For example, the altitude frame of reference is stored as the set of all elevation contour lines at a given resolution for the area, while the population frame of reference is the set of all town boundary

| ID | Latitude | Longitude |
|---|---|---|
| UKKL1001000023 | 53.707 | -1.6954 |
| UKKL1001000289 | 53.7116 | -1.7105 |
| UKKL1001000313 | 53.712 | -1.7145 |
| UKKL1001000345 | 53.7124 | -1.7204 |
| UKKL1001000391 | 53.7152 | -1.7263 |
| UKKL1001000451 | 53.7163 | -1.7361 |
| UKKL1001000497 | 53.7188 | -1.7442 |
| UKKL1001000512 | 53.7198 | -1.7471 |
| UKKL1001000533 | 53.721 | -1.751 |

| ID | Latitude | Longitude | Height | Direction | UrbanRural |
|---|---|---|---|---|---|
| UKKL1001000023 | 53.707 | -1.6954 | 0 | NORTHNORTHEAST | Dewsbury |
| UKKL1001000289 | 53.7116 | -1.7105 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000313 | 53.712 | -1.7145 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000345 | 53.7124 | -1.7204 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000391 | 53.7152 | -1.7263 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000451 | 53.7163 | -1.7361 | 100 | NORTHNORTHEAST | Cleckheaton |
| UKKL1001000497 | 53.7188 | -1.7442 | 100 | NORTHNORTHEAST | RURAL |
| UKKL1001000512 | 53.7198 | -1.7471 | 100 | NORTHNORTHEAST | RURAL |
| UKKL1001000533 | 53.721 | -1.751 | 100 | NORTHNORTHEAST | RURAL |

Figure 6: Geographic Characterisation of input data

polygons and it's complement within the area. The spatial database also allows the system to perform topological queries on the stored data.

**Spatial Reasoning module** The Spatial Reasoning module acts as layer between the main system and the spatial database. It performs two main functions: the first is to perform the geographic characterisation of the input data, the second is to provide functionality for the rest of the system to perform higher level spatial queries. These queries can range from combining frames of reference to adding location information to system events. Together with the Spatial database, the Spatial Reasoner acts in a similar way to a limited GIS system.

### 5.2. Content Selection

Our spatio-temporal analysis method, described in more detail in (Turner et al., 2007), explicitly takes into account the geographic characterisation of the forecast region. The data is clustered according to the frames of reference we have identified in our corpus providing results that can be easily mapped to spatial descriptions. The clustering method is also density based, applying proportions to each cluster that provide a simple mechanism for the system to include vague quantifiers in the generated spatial descriptions, such as those described in Section 4.2..

As the spatial descriptions generated by our system should express the effect of geographic features on weather conditions, our system implements a preference ordering over the way it selects the frame of reference it uses in the description. This is dependent on the type of event and parameter being described and is implemented based upon our corpus analysis described in Section 4.2.. An example of the preference ordering for describing road surface temperature is shown in Figure 7. This approach has similarities to the one described in (Kelleher and Kruijff, 2006), as it takes the context into account when considering which properties to use in the resulting description. Lexical choice is done simply by choosing the option for the reference frame value with the highest frequency in the corpus. The only exception to this is altitudes, which are described using the actual height value as requested by experts.

1. Altitude
2. Population
3. Coastal Proximity
4. Direction

Figure 7: Preference Ordering for Road Surface Temperature Events

*Road surface temperatures will fall below zero on most routes by evening. Wintry precipitation will affect some routes during the afternoon and evening clearing for a time by evening, falling as snow flurries in some places above 400M at first. Snow clearing by 21:40. Road surface temperatures will fall slowly during the mid afternoon and evening, reaching zero in some places above 400M by 18:00. Ice and hoar frost will affect most routes during the evening and tonight. Hoar frost turning heavy by evening except in areas below 100M. Rain will affect all routes during tonight and tomorrow morning turning heavy tomorrow morning except in far southern and north western areas.*

Figure 8: Example Weather Statement Generated by RoadSafe

## 6. Future Work and Conclusions

We have described a parallel data-text corpus of spatio-temporal NWP data and its corresponding output texts. The corpus was built as part of the KA studies carried out during the development of RoadSafe, a prototype NLG system that automatically produce road ice forecasts. The system is currently installed at AMI's premises and generating draft forecast texts which are being post edited by forecasters before being released to clients. The purpose of this evaluation is to further improve the quality of the texts in preparation for a full scale user evaluation in which we plan to compare the quality of the generated texts with human written ones.

As the input data in the corpus is large and complex, aligning words and phrases in the texts to actual data points was not possible. An analysis of the spatial descriptions in the corpus found that it contained no definite descriptions and almost no reference to named landmarks. Instead of named landmarks, spatial descriptions describe the location of events in the data in terms of frames of reference, which are sets of related geographical features that affect the general weather pattern. We found that this is because the purpose of the spatial descriptions is to communicate the effect of the geography on the data rather than be purely locative. Our observations led us to identify 3 additional modules necessary to incorporate into the architecture of our NLG system: geographic characterisation, spatial reasoning and a spatial database.

## Acknowledgments

Many thanks to our collaborators at Aerospace and Marine International UK, especially Keith Thomson and the other Meteorologists, for their helpful feedback and comments.

The RoadSafe project is supported jointly by Aerospace and Marine International UK, and the UK Engineering and Physical Sciences Research Council (EPSRC), under a CASE PhD studentship.

## 7. References

S. Boyd. 1998. Trend: a system for generating intelligent descriptions of time-series data. In *IEEE International Conference on Intelligent Processing Systems (ICIPS1998)*.

Charles B. Callaway and Johanna D. Moore. 2007. Determining tutorial remediation strategies from a corpus of human-human tutoring dialogues. In *Proceedings of the 11th European Workshop on Natural Language Generation*, Schloss Dagstuhl, Germany, June.

J. Coch. 1998. Multimeteo: multilingual production of weather forecasts. *ELRA Newsletter*, 3(2).

S. Geldof. 2003. Corpus analysis for nlg. In Reiter E., Horacek H., and van Deemter K., editors, *9th European Workshop on NLG*, pages 31–38.

E. Goldberg, N. Driedger, and R. Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 1041–1048.

Harvey J. Miller and Jiawei Han. 2001. Geographic data mining and knowledge discovery: An overview. In *Geographic Data Mining and Knowledge Discovery*, chapter 1, pages 1–32. Taylor & Francis.

Ehud Reiter and Somayajulu Sripada. 2003. Learning the meaning and usage of time phrases from a parallel text-data corpus. In *HLT-NAACL 2003 Workshop: Learning Word Meaning from Non-Linguistic Data*, pages 78–85, Edmonton, Alberta, Canada.

E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy. 2005. Choosing words in computer-generated weather forecasts. In *Artificial Intelligence*, volume 67, pages 137–169.

E. Reiter. 2007. An architecture for data-to-text systems. In *ENLG07*, pages 97–104.

Benjamin Snyder and Regina Barzilay. 2007. Database-text alignment via structured multilabel classification. In Manuela M. Veloso, editor, *IJCAI*, pages 1713–1718.

S. Sripada, E. Reiter, J. Hunter, and J. Yu. 2003. Exploiting a parallel text-data corpus. In *Proceedings of Corpus Linguistics*, pages 734–743, Lancaster, UK.

R. Turner, S. Sripada, E. Reiter, and I. Davy. 2007. Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data. In *Applications and Innovations in Intelligent Systems XV*, pages 75–88.

S. Williams and E. Reiter. 2005. Deriving content selection rules from a corpus of non-naturally occurring documents for a novel nlg application. In *Corpus Linguistics 2005 Workshop on Using Corpora for Natural Language Generation*.

# A field based representation for vague areas defined by spatial prepositions

**Mark M. Hall, Christopher B. Jones**

School of Computer Science
Cardiff University
Cardiff CF24 3AA, Wales, United Kingdom
[M.M.Hall, C.B.Jones]@cs.cardiff.ac.uk

**Abstract**

Natural language is one of the primary means of communicating spatial information, but existing geographic information retrieval (GIR) facilities are weak in this respect. One of the major challenges in automated interpetation of spatial natural language is how to model the regions implied by spatial expressions. This paper presents a field-based model for representing the vague regions defined by spatial language, including a method for defining field values from existing spatial language data sources. To interface this new vague field model with existing GIR systems and methods, an algorithm for extracting crisp boundaries from the field representation is also presented.

## 1. Introduction

Natural language is one of the primary means of communicating spatial information, but existing computing facilities for accessing geographically referenced information have limited capability for understanding spatial language except within the context of highly structured user interfaces. The quality of web search tools for accessing geographical information would be much improved if these systems could intelligently process queries that contained spatial prepositions such as near, in front and north when used in combination with named places.

A number of hurdles need to be overcome to succeed in intelligent interpretation of spatial natural language. A major challenge is to model spatial language, especially the spatial relations primarily defined by spatial prepositions. The regions that are referred to by the spatial relations have vague boundaries, and their extent is context dependent, varying with the types of phenomena involved and with the events taking place. Part of the vagueness arises from the fact that for a given situation many factors influence whether a particular spatial relation applies and there are no standard definitions for the use of the different prepositions that may apply. It is also the case that for most spatial prepositions there will be no agreement between users on exactly what is the form and extent of the space referred to. Indeed a single individual may not have a precise view of the region referred to.

This paper presents a field-based model for vague spatial relations and also a method for creating a crisp boundary from the field for integration with existing GIS methods. To avoid errors introduced by mixing knowledge from different domains within language, only spatial relations as used in image captions will be used.

Section 2 provides background information on spatial language and models, section 3 and 4 present the field based model and crisping algorithm and section 5 contains concluding remarks and an outlook to future work.

## 2. Representing vague areas

When dealing with spatial phenomena, the need to handle vagueness is unavoidable. A number of sources exist for this vagueness

- Multivariate classification - vagueness due to multiple non-independant classification criteria;
- Multiperson disagreement - vagueness due to different interpretations by people, see (Montello et al., 2003) and (Robinson, 2000);
- Natural vagueness - vagueness due to the sorites paradox as illustrated by (Fisher, 2000);
- Precision - vagueness due to representational and scale factors.

For a more detailed classification of vagueness in the context of geographical information, see (Evans and Waters, 2007).

### 2.1. Spatial models

Initial spatial models represented space in a crisp way. (Egenhofer, 1989) used a set-theoretic approach to represent the topological relations between two crisp regions. A second early topological model is the Region Connection Calculus (RCC) defined by (Randell et al., 1992), based on the connectivity relation $C(x, y)$ between two regions. Both models work well for the fiat and bona-fide boundaries as defined by (Smith and Varzi, 1997), but many of the spatial regions people encounter in practice are not crisp.

To deal with this vagueness inherent in human geography a number of extensions to the crisp models and also new vague models were defined. (Cohn and Gotts, 1996) created the *egg-yolk* model which is an extension of the RCC model. The *egg-yolk* model introduces a broad border region which defines the area that partially but not fully belongs to the core region. To what degree points within this border region are elements of the core region is not defined, only that they are no longer fully part of the core region and not yet part of the area outside. This provides a notion of vagueness while retaining the simplicity of reasoning with crisp regions. With a similar approach (Clementini and Felice, 1996) extend the 9-intersection model to deal with such broad-boundary regions. The two models differ in what relations they support, but are of similar expressivity. There are other approaches that also lead to a broad-boundary model (Schneider, 1996) (Bennet, 2001)

(Kettani and Moulin, 1999) (Kulik, 2001), the attraction of this approach being that they allow the basic modelling of vagueness without the complexity of having to consider the inherent characteristics of the vagueness or its relationship to real-world phenomena.

To model the details of how the vagueness works, fuzzy sets have been proposed as a solution. Fuzzy sets were introduced by (Zadeh, 1965) and are an extension of classical set theory. Instead of only providing a boolean member/non-member definition of a set, fuzzy sets provide a membership function $\mu : X \rightarrow [0, 1]$, where 0 is classical not-member and 1 is classical member-of. In GIS (Schneider, 1999) provides a definition for fuzzy points, lines and regions. This definition is extended in (Schneider, 2000) to a full algebra for fuzzy regions. Fuzzy sets are harder to handle than crisp sets, but (Schneider, 1999) shows how fuzzy sets can be reduced to a (possibly infinite) set of $\alpha$-cuts. An $\alpha$-cut is a crisping of a fuzzy set at an arbitrary value (the $\alpha$ value), where only those elements with membership values higher than $\alpha$ are part of the cut, thus deriving a classical crisp set from the original fuzzy set. The advantage of this approach is that all the existing work on crisp regions can be applied to such a set of $\alpha$-cuts.

When using fuzzy models to represent real world phenomena, the hardest problem is the definition of the membership function. (Robinson, 2003) gives an overview of different methods for defining the membership function. The approach taken by most is to use one of the standard membership functions to approximate the actual membership function. The properties of these standard functions are well understood and they are easy to handle. (Schockaert et al., 2008) follow this path in their work on modelling phrases such as "within walking distance". They use a standard trapezoidal membership function to approximate the data they mined from the web. Similarily (Mukerjee et al., 2000) use human input to modify the shape of a standard ellipsoid field representing the extent of a spatial relation.

### 2.2. Spatial language

Apart from maps and other graphical representations of space, the primary means for exchanging spatial information is natural language. The issues arising from spatial language and its use in accessing geographic information were raised by (Frank and Mark, 1991). The primary elements used in spatial language are count nouns referring to objects and spatial prepositions defining the spatial relations between the objects (Landau and Jackendoff, 1993). In image caption spatial language the role of the objects is taken by toponyms referring to places that are linked via the spatial prepositions. This paper focuses only on representing and handling the vague areas introduced by spatial prepositions and not on the representations of the places referred to by toponyms.

Spatial relations relate at least two objects to each other as in "A pond north of Stackpole". In this paper the object that acts as the reference object, "Stackpole" in this case, will be referred to as the *ground*, while the referred object "pond" will be called the *figure*. In the kind of spatial relations used in image captions, the *figure* object usually describes the content of the image, while the *ground* refers to a toponym in the close vicinity.

(Landau and Jackendoff, 1993) showed that the number of spatial prepositions is very small compared to the number of names for shapes and locations. Thus in order to be able to describe all possible configurations, they must be very flexible with respect to the situations they can be applied to. As (Herskovits, 1985) illustrates that means that it is very hard to cleanly define how and when they can be applied, using only a simple relations based approach.

Various models have been proposed to represent and reason with spatial relations derived from natural language. (Frank, 1996) describes an algebra for reasoning on the cardinal directions. The algebra can deal with a four direction (N, E, S, W) or an eight direction (N, NE, E, SE, S, SW, W, NW) model and allows answering queries of the kind "A is north of B and B is east of C. What is the relation between A and C?". (Goyal and Egenhofer, 2000) provide a similar eight direction model, extending the Frank model by allowing the regions involved in the cardinal relation to cover more than one direction. Other qualitative models for the cardinal directions can be found in (Ligozat, 1998) and in (Kulik et al., 2002) who introduce a ranking method for objects involved in a cardinal direction. (Hernández, 1991) describes a model for qualitatively representing spatial relations in an indoor context.

These models take a purely qualitative view of the spatial relations, and do not address the issue of geometric extent implied by the spatial relations. For modelling the quantitative aspects, fuzzy sets are advantageous as they avoid having to create a crisp approximation of the representation too early in the analysis process (Altman, 1994). (Robinson, 2000) uses a fuzzy approach to model nearness at the inter-town scale. Users are presented with a set of questions "Is town A close to town B?", with both towns shown on a map and the users answer *yes* or *no*. This builds a representation of nearness for one user, which can then be combined for multiple users to produce a general fuzzy representation of nearness.

One issue that is not addressed in this paper is how strongly language and culture influence spatial representation and reasoning. (Mark et al., 2007) and (Levinson, 2003) make a very strong case for the influence of space on language and vice-versa. Contrasting that (Xiao and Liu, 2007) and (Ragni et al., 2007) found that for non-linguistic tasks focusing on latitude estimates and topological classification, no significant differences between different cultures exist. While the methods presented in this paper are applicable to any language, the data presented here is taken from UK English image captions and thus only directly applicable to UK English image caption spatial language.

## 3. A field-based model for spatial relations

As (Couclelis, 1992) illustrates, geographic space can be seen from a vector or object, or from a raster or field perspective. While currently the object view dominates in GIS software, most of the data handled by these systems are actually more field-like in nature. This is especially true for the spatial relations this paper focuses on, which describe regions in which the membership to the spatial relation varies across the whole region. (Nishida et al., 1987) describe a field model for placing objects in a spatial scene, but how empirical data would work into the field is unclear. For representing spatial relations such as "north of", a field based approach has one strong advantage over traditional crisp and *egg-yolk* models, in that it can accurately represent the level of membership at each point and does not only provide a rough approximation. The fuzzy models described earlier are much closer to a field representation, but the field model does offer an advantage. The hardest problem for fuzzy models is determining the membership function. Also it is difficult to combine multiple membership functions into one final result, in those cases where multiple factors are relevant to the membership. A field based model avoids both problems, as the field values can be taken directly from the data source, without having to be fit into a functional model first.

The field model for spatial relations is represented as an $n \times n$ matrix, with the ground for the spatial relation usually located at $(\frac{n}{2}, \frac{n}{2})$, although that is not necessary and in the examples presented, the ground is placed further towards the bottom of the field to improve image clarity. Each field cell holds a membership value that is defined on the interval $[0, 1]$, with 0 signifying no membership and 1 complete membership to the spatial relation. Depending on the spatial relation used the size of each raster cell can be varied, but for the spatial relations described here, a cell size of 50x50m has been determined to be the best compromise between spatial resolution and computational complexity.

Using this model a number of spatial relations have been modelled, with the results for "north of" presented here. The raw data that the field is based on, were taken from the Geograph project. The Geograph[1] project aims to cover each square kilometre of the UK with a representative photograph. These photographs contain a caption and location information from GPS units. Since the aim is to create representative photographs, the captions chosen tend to be spatial in nature, describing the location of the photograph and not only its contents. As such the project represents an ideal source of spatial linguistic information. The Geograph project has provided a database dump of roughly 350,000 records, containing image captions and location information, but not the actual images themselves.

The first step in constructing the field is extracting uses of the desired spatial relation from the Geograph captions. GATE[2] is used for part of speech tagging and the identification of spatial relations. Identifying toponyms in captions is not an easy task, but based on an analysis of the image captions, a simple metric was devised. Any word that starts with an uppercase letter is assumed to be a candidate toponym. Combined with a list of excluded words such as "A" and "The", this metric provides good results.

The tagged captions are then matched against patterns of the form "<spatial relation> <toponym>", in this case "north of <toponym>". The hypothesis employed was that the GPS coordinates of the image and the location of the toponym matched by the pattern formed one valid use for the spatial relation. As each spatial relation appears multiple times it is possible to build up a set of valid uses, which then feed into the field representation.

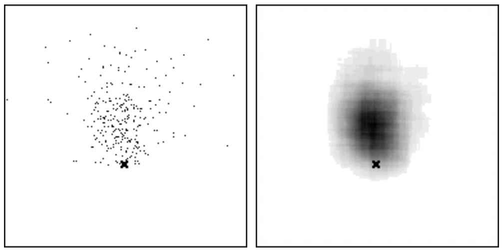

Figure 1: Field model representation: Initial point cloud, smoothed field ($\times$ marks the ground toponym location)

The toponyms are geocoded using the Geonames.org service, which returns a point representation of the centre of the toponym location. No toponym disambiguation was performed, except for only accepting exact toponym matches. As the distances involved in the spatial language of image captions tend to be short (most less than 5km), an incorrect disambiguation is immediately clear as a statistical outlier. For each of the patterns the GPS co-ordinates of the image and the location of the toponym are combined to calculate the angle and distance of the image location from the ground toponym.

These distance/angle pairs are then plotted onto the field, relative to the ground toponym. As this method combines distance and angle data from multiple captions, it is necessary to guarantee that the scale involved in all captions is the same. The area "north of" a point of interest such as a church will have a different scale to that "north of" a town or village. The Geonames.org service in addition to the location of the toponym also provides information on the toponym's type and in the data presented here only toponyms of the type *populated place* were used. In combination with the fact that when locating images only very local information is used, this gives a high confidence in the data. Figure 1 shows the plot of the spatial relation "north of <toponym>". The plotted field is then smoothed using a 30x30 cell rectangular kernel and the resulting values normalised to the [0, 1] range, so that the crisping algorithm can be applied.

## 4. Crisping the field model

The field model provides a very powerful representation for spatial relations, but for using the results in other GI systems or applying existing crisp methods, a vector based representation is required. This crisping makes it possible to use a vague representation of "north of" as the input into a crisp spatial query in current GI systems. The crispings should always have meta-data associated with them, that document that the crisp representation is just one possible crisping, and not a normative result for the spatial relation. In fuzzy models $\alpha$-cuts (Klir and Yuan, 1995) or centre of area methods (Power et al., 2001) (Palanciogla and Beard, 2001) are employed to crisp the fuzzy representation. While the $\alpha$-cut method could also be applicable to the field model, this paper presents an active contour based crisping algorithm. The advantage of an active contour crisping algorithm is that further constraints and influences

[1] http://www.geograph.org.uk

[2] http://gate.ac.uk

can easily be integrated into the crisping algorithm. Examples of such constraints might be hard boundaries such as shorelines or mountains, or softer influences like the influence of road-connectivity on the shape of the relation, or the conflicting influence of other spatial relations that could be used to describe the location.

Active contour models were first introduced by (Kass et al., 1988) for finding boundaries in image data. They are defined as energy minimising functions, consisting of an internal energy that is responsible for maintaining the active contour's shape and an external energy that represents the data to be modelled (equation 1). In image processing the internal energy is usually defined so as to maintain an even spacing between the control points and also to smoothen the angles at each point (Lam and Yan, 1994). The external energy is then defined by the image being processed. Frequently the energy source is the gradient of the image, the active contour is then attracted to boundaries in the image where the gradient is steepest (Lam and Yan, 1994). These kinds of active contours are frequently employed in medical image feature extraction (Shang et al., 2008).

$$E(s) = E_{int}(s) + E_{ext}(s) \quad (1)$$

Active contours have also been used in the GIS field, (Burghardt, 2005) and (Steiniger and Meier, 2004) use them for line smoothing in map generalisation applications. For the line smoothing used in map generalisation, no external energy is needed. The active contour is initialised with the points of the original line and the smoothing is defined solely by the internal energy. An external energy is only applied if the active contour should also maintain a distance from certain points, as in the situation when the active contour overlaps with another line in the map. Then proximity of the control points on the second line acts as the external energy pushing the active contour away from the second line.

The key difference between using active contours for crisping the field representation and the previously listed problems, is that the crisping problem lacks a clearly defined border to which the active contour could be attracted. To counter this a third energy has been introduced, which forces the active contour to contract towards the centre of the spatial relation field. The active contour now consists of three energies (equation 2). $E_{int}(s)$ maintains the active contour's shape, $E_{relation}(s)$ is the external energy defined by the spatial relation field and $E_{contract}(s)$ specifies the contraction energy.

$$E(s) = \alpha \cdot E_{int}(s) + \beta \cdot E_{relation}(s) + \gamma \cdot E_{contract}(s) \quad (2)$$

A greedy algorithm has been designed, that iteratively finds a solution to the crisping problem. For computational reasons, instead of working directly on the scalar values, a vector representation has been chosen. On each iteration $E_{int}(s)$ is calculated for each control point by first determining the vector from the predecessor control point to the successor control point. The mid-point of this vector is then calculated and the internal energy is defined as the vector from the current control point to that mid-point. This maintains a smooth curvature and an even spread between the control points.

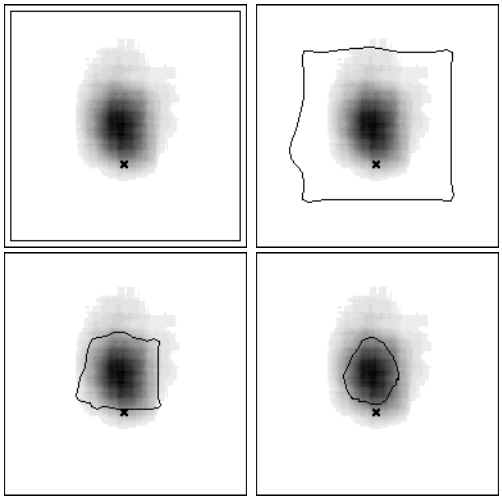

Figure 2: Four states in the active contour process for "north of <toponym>": Initial position, intermediate shape 1 and 2, final result ($\times$ marks the ground toponym location)

$E_{relation}(s)$ is calculated from the original field data by applying the gradient operator to it, creating a vector field, the gradient flow field. For each point in the original scalar field the gradient operator finds the neighbouring point with the largest difference in the scalar value. This defines the direction of the vector and the length is then calculated from the difference in the scalar values. $E_{contract}(s)$ is defined directly as a vector field in which all vectors point to the area within the relation field with the highest membership value and are of equal length, providing a constant contraction energy.

Each of these energies has a weight associated with it, the manipulation of which modifies the final contour form. The $\alpha$ weight modifies the internal energy of the active contour. Increasing the value creates a stiffer active contour, while lowering it allows sharp corners to appear. The contraction weight $\gamma$ controls how far the active contour contracts. A high value leads to a smaller final result, while a smaller value produces a larger active contour. The weight $\beta$ on the relation energy acts as a balance between the two other weights. For example increasing the internal energy weight $\alpha$ will also lead to a slightly stronger contraction. In order to maintain the same level of contraction, $\beta$ is increased, maintaining the amount of contraction, but now with a more rigid active contour. The results shown in figure 2 used weights of $\alpha = 70$, $\beta = 255$, $\gamma = 110$.

After weighting each energy, the vectors are combined and the control point is then moved one cell in the direction of the total energy vector. The length of the total energy vector is not taken into account. As the control point is immediately updated, it changes its influence on the internal energy of the next control point. This means that the algorithm will find locally optimal solutions for each control point, but the results will not necessarily be globally optimal. Determining when to terminate the algorithm is a hard problem.

Multiple termination criteria are under consideration, but currently a hard limit of 400 iterations is implemented.

When applying the algorithm to an image caption, after the algorithm terminates, the co-ordinates for the crisp boundary are calculated from the active contour control points based on a mapping of the ground location in the field to the actual location of the toponym from the image caption.

### 4.1. Evaluation

As the crisping algorithm is essentially arbitrary, it is necessary to provide a confidence value for it. This confidence value is not a measure of whether the generated shape is correct or not, but describes how confident the algorithm is that the resulting polygon is acceptable to a majority of people. The confidence function $C(s,t) \rightarrow [0,1]$ is defined on the active contour s and a test set of valid uses of the spatial relation r. These valid uses can either be automatically calculated from an existing data set such as the Geograph data, or elicited directly from users using other methods.

$$C(s,t) = 1 - \left|\frac{\frac{count(t)}{2} - inside(s,t)}{\frac{count(t)}{2}}\right| \quad (3)$$

In the confidence function the confidence is highest at the point where half the points in the test set are outside and half inside the active contour. The confidence decreases as the active contour covers either less or more of the test set. The rationale behind this is that the confidence in the active contour result is highest when the number of test points inside and outside the active contour are the same. In this situation the number of people who would say that the relation extends further is in balance with the number of people who would say that the relation does not extend that far, increasing the likelihood that both groups agree that the result is an acceptable approximation of what they believe to be true. The final active contour shown in the fourth image of figure 2 has a confidence value $C = 0.96$.

## 5. Conclusion and future work

This paper presented a field-based model for representing vague areas defined by natural language spatial relations. In order to interface the model with current crisp-region based models, a crisping algorithm based on active contours is described. The model and crisping algorithm are applied to the domain of spatial language in image captions and the field creation and crisping is demonstrated on data generated with image captions taken from the Geograph project. The major advantage of the field model over current broad boundary and fuzzy models is that it allows the precise modelling of vague regions, while avoiding the complexity of fuzzy representations. As few current GI systems support vague regions, the crisping algorithm makes it possible to easily integrate the field model into existing systems and methods.

Due to the nature of the data that forms the basis for this work, the results are restricted to the context of image captioning and the scale of populated places. Future work will focus on extending the model to further contexts, spatial relations and scales. This will include dealing with differing spatial reference frames. The focus will also be on acquiring spatial relation extents directly from people and on how different languages influence these extents.


## Acknowledgements

We would like to gratefully acknowledge contributors to Geograph British Isles (see http://www.geograph.org.uk/credits/2007-02-24, whose work is made available under the following Creative Commons Attribution-ShareAlike 2.5 Licence (http://creativecommons.org/licenses/by-sa/2.5/).

This material is based upon work supported by the European Community in the TRIPOD (FP6 cr n045335) project.


## 6. References


D. Altman. 1994. Fuzzy set theoretic approaches for handling imprecision in spatial analysis. *International Journal of Geographical Information Science*, 8(3):271–289.

B. Bennet. 2001. Application of supervaluation semantics to vaguely defined spatial concepts. In *Spatial Information Theory. Foundations of Geographic Information Science : International Conference, COSIT 2001 Morro Bay, CA, USA, September 19-23, 2001. Proceedings*, pages 108–123.

D. Burghardt. 2005. Controlled line smoothing by snakes. *GeoInformatica*, 9(3):237–252.

E. Clementini and P. Di Felice. 1996. An algebraic model for spatial objects with indeterminate boundaries. In *Geographic Objects with Indeterminate Boundaries*, pages 155–169. Taylor and Francis.

A.G. Cohn and N.M. Gotts. 1996. The 'egg-yolk' representation of regions with indeterminate boundaries. In *Proceedings, GISDATA Specialist Meeting on Geographical Objects with Undetermined Boundaries*, pages 171–187. Francis Taylor.

H. Couclelis, 1992. *People manipulate objects (but cultivate fields): Beyond the raster-vector debate in GIS*, volume 639/1992 of *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, pages 65–77. Springer Berlin / Heidelberg.

M.J. Egenhofer. 1989. A formal definition of binary topological relationships. In *3rd Internationl Conference, FODO 1989 on Foundations of Data Organizationand Algorithms*, pages 457–472. Springer-Verlag New York.

A.J. Evans and T. Waters. 2007. Mapping vernacular geography: web-based gis tools for capturing 'fuzzy' or 'vague' entities. *International Journal of Technology, Policy and MAnagement*, 7(2):134–150.

P. Fisher. 2000. Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, 113(1):7–18.

A.U. Frank and D.M. Mark, 1991. *Language Issues for Geographical Information Systems*, pages 147–163. Geographical Information Systems: Principles and Applications. London: Longman.

A.U. Frank. 1996. Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Science*, 10(3):269–290.

R. Goyal and M. J. Egenhofer. 2000. Cardinal directions between extended spatial objects. *IEEE Transactions on Knowledge and Data Engineering*.

Daniel Hernández. 1991. Relative representation of spatial knowledge: The 2-d case. In D.M. Mark and A.U. Frank, editors, *Cognitive and Linguistic Aspects of Geographic Space*, pages 373–385. Kluwer Academic Publishers.

A. Herskovits. 1985. Semantics and pragmatics of locative expressions. *Cognitive Science: A Multidisciplinary Journal*, 9(3):341–378.

M. Kass, A. Witkin, and D. Terzopoulos. 1988. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331.

D. Kettani and B. Moulin. 1999. A spatial model based on the notions of spatial conceptual map and of object's influence areas. In *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science: International Conference COSIT'99*, volume 1661/1999 of *Lecture Notes in Computer Science*, pages 401–416. Springer Berlin / Heidelberg.

G.J. Klir and B. Yuan. 1995. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice-Hall.

L. Kulik, C. Eschenbach, C. Habel, and H.R. Schmidtke. 2002. A graded approach to directions between extended objects. In *Proceedings of the Second International Conference on Geographic Information Science*, volume 2478, pages 119–131. Springer Berlin / Heidelberg.

L. Kulik. 2001. A geometric theory of vague boundaries based on supervaluation. In *Spatial Information Theory. Foundations of Geographic Information Science : International Conference, COSIT 2001*, pages 44–59. Springer Berlin / Heidelberg.

K.-M. Lam and H. Yan. 1994. Fast greedy algorithm for active contours. *Electronics Letters*, 30(1):21–23.

B. Landau and R. Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238.

S.C. Levinson. 2003. *Space in language and cognition: Explorations in cognitive diversity*. Cambridge: CUP.

G. É. Ligozat. 1998. Reasoning about cardinal directions. *Journal of Visual Languages and Computing*, 9(1):23–44.

D.M. Mark, A.G. Turk, and D. Stea. 2007. Progress on yindjibarndi ethnophysiography. In *Spatial Information Theory, 8th International Conference, COSIT 2007, Melbourne, Australia, September 19-23, 2007, Proceedings*, pages 1–19.

D.R. Montello, M.F. Goodchild, J. Gottsegen, and P. Fohl. 2003. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(3):185–104.

A. Mukerjee, K. Gupta, S. Nautiyal, M.P. Singh, and N. Mishra. 2000. Conceptual description of visual scenes from linguistic models. *Image and Vision Computing*, 18(2):173–187.

T. Nishida, A. Yamada, and S. Doshita. 1987. Figuring out most plausible interpretation from spatial constraints. In *Spatial Reasoning and Multi-sensor Fusion: Proceedings of the 1987 Workshop*, pages 158–167.

H.M. Palanciogla and K. Beard. 2001. Modeling moving objects and their movements using fuzzy logic approach. In *Proceedings of the American Society for Photogrammetry and Remote Sensing*, page CD ROM.

C. Power, A. Simms, and R. White. 2001. Hierarchical fuzzy pattern matching for the regional comparison of land use maps. *International Journal of Geographical Science*, 15:77–100.

M. Ragni, B. Tseden, and M. Knauff. 2007. Cross-cultural similarities in topological reasoning. In *Spatial Information Theory, 8th International Conference, COSIT 2007, Melbourne, Australia, September 19-23, 2007, Proceedings*, pages 32–46.

D.A. Randell, Z. Cui, and A. Cohn. 1992. A spatial logic based on regions and connection. In *KR'92. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Converence*, pages 165–176. Morgan Kaufmann.

V.B. Robinson. 2000. Individual and multipersonal fuzzy spatial relations acquired using human–machine interaction. *Fuzzy Sets and Systems*, 113(1):133–145.

V.B. Robinson. 2003. A perspective on the fundamentals of fuzzy sets and their use in geographic information systems. *Transactions in GIS*, 7(1):3–30.

M. Schneider. 1996. Modelling spatial objects with undetermined boundaries using the realm/rose approach. In *Geographic Objects with Indeterminate Boundaries*, volume 2, pages 141–152. Taylor and Francis.

Markus Schneider. 1999. Uncertainty management for spatial data in databases: Fuzzy spatial data types. *Lecture Notes in Computer Science*, 1651:330–351.

M. Schneider. 2000. Finite resolution crisp and fuzzy spatial objects. In *International Symposium on Spatial Data Handling*, pages 3–17.

S. Schockaert, M. de Cock, and E.E. Kerre. 2008. Location approximation for local search services using natural language hints. *International Journal of Geographical Information Science*, 22(3):315–336.

Y. Shang, X. Yang, L. Zhu, R. Deklerck, and E. Nyssen. 2008. Region competition based active contour for medical object extraction. *Computerized Medical Imaging and Graphics*, 32(2):109–117.

B. Smith and A.C. Varzi. 1997. Fiat and bona fide boundaries: Towards an ontology of spatially extended objects. In *Spatial Information Theory A Theoretical Basis for GIS (COSIT'97)*, volume 1329 of *Lecture Notes in Computer Science*, pages 103–119.

S. Steiniger and S. Meier. 2004. Snakes: a technique for line smoothing and displacement in map generalisation. In *ICA workshop on generalisation and multiple representation*.

D. Xiao and Y. Liu. 2007. Study of cultural impacts on location judgments in eastern china. In *Spatial Information Theory, 8th International Conference, COSIT 2007, Melbourne, Australia, September 19-23, 2007, Proceedings*, pages 20–31.

L.A. Zadeh. 1965. Fuzzy sets. *Information and Control*, 8:338–353.