Trailfinder: A Case Study in Extracting Spatial Information

Using Deep Language Processing

*Lars Hellan, Dorothee Beermann, Jon Atle Gulla, *Atle Prange*

NTNU, Trondheim, Norway *Businesscape, Trondheim, Norway

Abstract

The present paper reports on an end-to-end application using a deep processing grammar to extract spatial and temporal information of prepositional and adverbial expressions from running text. The extraction process is based on the full understanding of the input text. It is represented in a formalism standard for unification-based grammars and with a language-independent vocabulary as far as spatiotemporal information is concerned. The latter feature in principle allows portability of the extraction algorithm across languages and applications, as long as the domain is kept constant.

The present application is called 'Trailfinder', and supports web-queries about information concerning mountain hikes. A standard hike-description is parsed by an HPSG-based grammar augmented by Minimal Recursion Semantics ('MRS'; (Copestake 2002)). To represent domain-specific meaning concerning location and direction, we enrich MRS structures with feature-based interlingua specifications. Utilizing the 'Heart of Gold' (HoG)[1] technology developed as part of the Deep Thought project[2], and conversion algorithms employing XML sheets, these specifications are mapped to the query interface language.

1 Introduction

One of the problems in arriving at a theoretically satisfactory semantic account of prepositions is their well known polysemy. The Reader's Digest Great Encyclopedic Dictionary for example lists 13 different meanings for the word *behind* - 5 for the adverbial and 8 for the prepositional uses. Many of the most frequent prepositions are in addition ambiguous between a directional and a locative reading, as for example in *The dog runs in the garden*. The directional versus locative interpretation and the adequate semantic modelling of the concepts of PATH and PLACE have been an issue of intensive linguistic research ((Fillmore 1975), (Cresswell 1985), (Talmy 2000), (Jackendoff 1990), and more recently, (Kracht 2002), to mention a few).

In NLP, a reflection of prepositional polysemy is typically encountered in MT, where a single prepositional expression in a source language may correspond to a multitude of expressions in the target language, depending on the object of the prepositional head. The English preposition *on*, e.g., corresponds to the German prepositions *auf, über, an* when combined with an NP expressing place, subject matter or day of the week, respectively. With respect to the problem of determining the correct target language counterpart in such cases, one type of approach which has been developed is symbolic, positing semantic specification in terms of features. Within unification grammar, one of the well known approaches of this type was suggested in (Halvorsen 1995) for Lexical Functional Grammar-based grammar engineering, and

[1] http://heartofgold.dfki.de/

[2] http://www.eurice.de/deepthought

within the machine translation context for prepositional expressions in particular, in (Trujillo 1995), who in turn builds on conceptual distinctions drawn in the linguistic literature (for a survey of this literature, see (Trujillo 1995)).

In the present work we suggest a feature based semantic representation for disambiguation, as part of Minimal Recursion Semantics.

The paper is organized as follows: In section 2 we briefly introduce the Norwegian HPSG grammar 'NorSource', and the Heart of Gold architecture. Section 3 presents the semantics: in section 3.1, we give a short introduction to the MRS formalism, and in section 3.2 we describe our system of sortal specifications on indices as part of the MRS formalism. Section 4 describes the Trailfinder architecture: 4.1 discusses RMRS/XML conversion, and section 4.2 XML transformations. In section 5, we discuss the potential for further developments using the approach instantiated here.

2 The Norwegian HPSG Resource Grammar 'NorSource' and the Heart of Gold

For this project the Norwegian HPSG Resource Grammar 'NorSource'[3] has been used to parse selected sentences from on-line hike descriptions of the Trollheimen region in the middle part of Norway. NorSource was developed as part of the EU-project DeepThought[4] and at present is part of the multilingual grammar engineering initiative Delph-In (http://www.delph-in.net/). It is implemented in the platform Linguistic Knowledge Builder (LKB; Copestake 2002).

In the Trailfinder application, NorSource is used as part of the Heart of Gold component 'PET' ((Callmeier, Schaefer and Siegel 2004)). The Heart of Gold (HoG) is an NLP-architecture which provides, through RMRS ('Robust Minimal Recursion Semantics', cf. (Copestake n.d.)), an interchange format for NLP components of different granularity of processing. In the present application it is used to communicate between parses produced by NorSource and a database for the storage of RMRS representations and a Web Browser. Thus, different from other work on Information Extraction, we do not extract directly from text, but use the markup RMRS produced by our deep processing grammar to encode and store the relevant information. (R)MRS will be described in more detail in the following.

3 Minimal Recursion Semantics

Minimal Recursion Semantics (MRS) representations are 'flat' representations of the elementary predications that compositionally represent the meaning connected to individual constructions, and provide the possibility of underspecifying scope (Copestake et al. 2001, Copestake et al. to appear). The specifications provided by Norsource are, for the present application, in a wholesale fashion carried over to the RMRS markups that we provide for Trailfinder. In the following section we give a short introduction to MRS. Although we use Robust MRS (RMRS) to communicate with Trailfinder, we concentrate in our discussion of prepositional semantics on MRS. (One of the main

[3] More information about NorSource see (http://www.ling.hf.ntnu.no/forskning/norsource).

[4] For more information about the DeepThought project see (http://www.project-deepthought.net)

reasons for the development of RMRS is that it can be used as an output format also for shallower NLP applications such as parts-of-speech parsers and chunkers, and thus allows for a common exchange format between applications of different depth of analysis. For the present application, however, we only work with a deep processing grammar, so that this important aspect of the use of RMRS becomes less relevant.).

3.1 Introduction to MRS

$$\begin{bmatrix}
\text{LTOP:H1} \\
\text{INDEX E2} \left[\text{E}\left[\text{TENSE:PRES}\right]\right] \\
\text{RELS:}\left\langle \begin{array}{l}
\begin{bmatrix} \text{DEF\_ Q\_ REL} \\ \text{LBL H5} \\ \text{ARG0 X4} \\ \text{RSTR H6} \\ \text{BODY H7} \end{bmatrix},
\begin{bmatrix} \text{\_ BOY\_ N\_ REL} \\ \text{LBL H3} \\ \text{ARG0 X4} \end{bmatrix},
\begin{bmatrix} \text{\_ WALK\_ V\_ REL} \\ \text{LBL H8} \\ \text{ARG0 E2} \\ \text{ARG1 X4} \end{bmatrix},
\begin{bmatrix} \text{\_ ALONG\_ P\_ REL} \\ \text{LBL H8} \\ \text{ARG0 E8} \\ \text{ARG1 E2} \\ \text{ARG2X9} \end{bmatrix}, \\
\begin{bmatrix} \text{DEF\_ Q\_ REL} \\ \text{LBL H11} \\ \text{ARG0 X9} \\ \text{RSTR H12} \\ \text{BODY H13} \end{bmatrix},
\begin{bmatrix} \text{\_ RIVER\_ N\_ REL} \\ \text{LBL H10} \\ \text{ARG0 X9} \end{bmatrix},
\begin{bmatrix} \text{PRPSTN\_ REL} \\ \text{LBL H1} \\ \text{MARG H14} \end{bmatrix}
\end{array}\right\rangle \\
\text{HCONS:}\left\langle \text{H6 QEQ H3, H12 QEQ H10, H14 QEQ H8} \right\rangle
\end{bmatrix}$$

Figure 1: MRS-structure for the sentence *The boy walks along the river*

The above Figure 1 shows a fully specified MRS representation for the sentence

(1) The boy walks along the river

In accordance with a standard MRS set up, for any constituent C (of any rank), the RELS list is a 'bag' of those elementary predications (EPs) that are expressed inside C. The sentence *The boy walks along the river* displays seven EPs, representing the meaning of the six expressions that form this sentence plus one [prpstn_ rel] which reflects the 'message type' of the sentence. The subject argument of the verb *walk* is represented by the coindexation of the verb's ARG1 with the ARG0 of the determiner and the noun that constitute the subject; correspondingly for the ARG2 of the preposition. ARG0 variables are typed: x-type variables correspond to the 'bound variable'of nominal expressions, while 'e' is the type of an event-variable. Scope properties are expressed in the HCONS list, 'x QEQ y' meaning essentially that x scopes over y. HCONS thus records the scopal tree of the constituent in question, as outlined in Copestake et al. (to appear).

The PP *along the river* is interpreted as an event modifier, in the figure represented by the circumstance that the ARG1 of the [_ along_ p_ rel] takes as value the variable 'e2' of the verb, while the handles of the verbal and the prepositional predicate are made identical. A further important feature of MRS structures that carries over to the RMRS structure is that the 'name' of every elementary predication consists of slots, where the first slot corresponds to the morphological stem, and the second slot indicates its categorial type. This information can be used for further extraction of relevant information; in our case, e.g., we were mainly interested in EPs with the the categorial label _ p.

As mentioned before, the additional semantic sort specifications of the (R)MRS used for Trailfinder are directly provided by NorSource. To this end NorSource was made to process additional sortal class information alongside standard semantic information. In (Hellan and Beermann 2005), we discuss other techniques such as the use of an OWL hierarchy to integrate word sense disambiguation into RMRS. Here we now continue with the presentation of MRS structures that accommodate the additional semantic information.

3.2 Enriched MRS structures

In hiking route descriptions, certain features are prevalent, such as the frequent use of implicit subjects, imperatives and the concatenation of PPs specifying stretches of hikes; the Norwegian text in Figure 5 further below illustrates some of these features. Of further interest to the deep parsing of tour descriptions are verbs of movement in space which in Norwegian are often instantiated as verb-particle constructions, such as *gå-opp/ned/bakover*.[5] Our focus however is on the exploration of prepositional and adverbial senses for the language independent representation of movement in space.

The following are some of the domain specific linguistic features of a text describing hikes: 1. Throughout most of the text, there is a constant ’agent’, which can be conceived either as a ’mover’, a ’tour’, or a road/path - regardless of which perspective is taken, this will be one and the same ’mileage consumer’. An essential aspect of inter-sentential anaphora in these texts is thereby fixed, so that in the summarization of each sentence taken separately, the semantic argument linked to the syntactic subject, that is the ARG1, will have a fixed value. 2. Consecutive sentences, and consecutive directional specifications inside each sentence, generally describe temporally and/or linearly consecutive stretches of path or path-consumption. Also this aspect of intersentential anaphora can be externally superimposed on the representation of each sentence (we return to this issue as we proceed).

An interesting exception are phrases like:

(2) up along the path

where the specifications *up* and *along...*, as long as they are not separated by a comma, typically co-specify the same stretch or move: this situation is represented through the assignment of identical ARG0-values in the EPs representing *up* and *along...*, as opposed to distinct values in the case of consecutive construals. The timeline of a hike is thus reflected in the value of the predicate’s ARG0 and distinct ARG0 values map into consecutive ’stages’ of the hike.

[5] A further element relevant for information extraction from hike descriptions are place names which often constitute a combination of proper names (e.g., ’Storli’) with nouns denoting landscapes, such as ’valley’, ’creek’, ’lake’, etc.; an illustration is given in the translation underneath Figure 5. In the present application we leave place names unanalyzed.

A further analytic concern implemented is the difference between static or locative expressions and directionals. While 'static' modifiers like *in the valley* in phrases like:

(3) they walk around in the valley

(4) the house in the valley

are treated as modifiers predicated of (that is, having as value of their ARG1) the index of the head (that is, for example, *walk* in (3)), directional phrases like:

(5) to the cabin

take as their ARG1 the 'path consumer' expressed, be it as a moving individual or as a road/path. Formally, this is shown by (5) always having an x-variable as its ARG1, whereas ((3) has an e-variable as ARG1. An illustration of the latter case is given in figure 2 below for the preposition *along*. We thus take the approach of (Jackendoff 1990) to directionals and implement the 'mover'as the entity 'measuring out the path'. However, nothing basic to this application hinges on this decision: if we were to treat directionals as event modifiers on a par with stative expressions, both expressions would still be internally distinguished by the value of their SORT attribute. So in short, in the present context, crucial to the summarization of a hiking text is whether a certain location plays the role of starting point, via-point or end point, or of a path or line followed. Important to notice is that in a purely monolingual application, these semantic differences could in principle be accommodated through the representation of the prepositional or adverbial lemmas themselves. However, in a multilingual setting this will not suffice. For example, in an MT application it is the representation of the ambiguity of the English sentence *He walked in the forest* that, for successful generation of a corresponding expression in, e.g., Norwegian or German, needs to lead to two distinct strings, one of which will represent the locative and the other one the directional reading. Likewise for IE, an extraction algorithm based on language independent sortal features is clearly to be preferred over one using language specific features, lending itself more readily to cross linguistic application.

For the application in question, this means that the MRS produced by NorSource will have to supply the arguments of prepositions and adverbs with semantic specifications indicating whether a relation expressed is one of movement to endpoint of path, via viapoint of path, from startpoint of path, or movement along a path: these are, for the time being, specifications under ARG0.SORT.[6] Moreover, for the ARG2 of prepositions (reflecting the governed NP), there will be a SORT specification of whether this is an endpoint, viapoint, etc. This design is illustrated in figure 2 below. The prepositional relation [[_ along_ p_ rel] is annotated with sortal specifications for its ARG0 and its ARG2, indicating that *along* is a preposition of the semantic type 'along-path-motion' and that the ARG2 of this type of preposition denotes a semantic argument of the type 'path-followed'. The full range of prepositions and adverbs in the locative domain are analyzed according to these parameters.

[6] For a development of this approach, see (Hellan and Beermann 2005)

$$\begin{bmatrix}
\text{LTOP:H1} \\
\text{INDEX E2}\begin{bmatrix}\text{E}\begin{bmatrix}\text{TENSE:PRES}\end{bmatrix}\end{bmatrix} \\
\text{REL:}\left\langle \begin{array}{l}
\begin{bmatrix}\text{DEF\_ Q\_ REL}\\ \text{LBL H5}\\ \text{ARG0 X4}\\ \text{RSTR H6}\\ \text{BODY H7}\end{bmatrix},
\begin{bmatrix}\text{\_ BOY\_ N\_ REL}\\ \text{LBL H3}\\ \text{ARG0 X4}\end{bmatrix},
\begin{bmatrix}\text{\_ WALK\_ V\_ REL}\\ \text{LBL H8}\\ \text{ARG0 E2}\\ \text{ARG1 X4}\end{bmatrix}, \\
\begin{bmatrix}\text{\_ ALONG\_ P\_ REL}\\ \text{LBL H8}\\ \text{ARG0 E8}\begin{bmatrix}\text{SORT:ALONG-PATH-MOTION}\end{bmatrix}\\ \text{ARG1 E1}\\ \text{ARG2 X9}\begin{bmatrix}\text{SORT:}\begin{bmatrix}\text{PATH-FOLLOWED}\end{bmatrix}\end{bmatrix}\end{bmatrix},
\begin{bmatrix}\text{DEF\_ Q\_ REL}\\ \text{LBL H11}\\ \text{ARG0 X9}\\ \text{RSTR H12}\\ \text{BODY H13}\end{bmatrix}, \\
\begin{bmatrix}\text{\_ RIVER\_ N\_ REL}\\ \text{LBL H10}\\ \text{ARG0 X9}\end{bmatrix},
\begin{bmatrix}\text{PRPSTN\_ REL}\\ \text{LBL H1}\\ \text{MARG H14}\end{bmatrix}
\end{array}\right\rangle \\
\text{HCONS:}\langle \text{H6 QEQ H3, H12 QEQ H10, H14 QEQ H8}\rangle
\end{bmatrix}$$

Figure 2: Enriched MRS-structure for the sentence *The boy walks along the river*

4 The Trailfinder Architecture

Trailfinder is a client-server architecture implemented in Java. An administrator regulates the communication with the HoG and cleans and filters the RMRS structures received from the HoG for their final use by the web client. Initially an XML/RPC call is placed to the mocomanserver (HoG). The received RMRS structures are stripped of unnecessary information and stored in the database. In a second step the filtered data is analyzed and placed in a database of tour descriptions accessible to the Search Engine. The Trailfinder architecture is illustrated in figure 3 below. The RMRS received from the HoG is marked up in XML. This makes it relatively easy to filter out the contents of RMRS that are not useful for Trailfinder. The filtering is done with XSLT, and only the nodes marked as EP, ARG0, ARG1 and ARG2 (and their children) are stored in Trailfinder's database. The HoG proxy shown in Figure 3 sends and receives only one sentence at a time. To arrive at a complete tour description, sentences have to be grouped into a single document for storage. Each sentence is considered a <em>stage</em> in a trip. All the senteces are grouped under the name <em>trip</em> in the single XML document. This document is then stored as XML in Trailfinder's database.

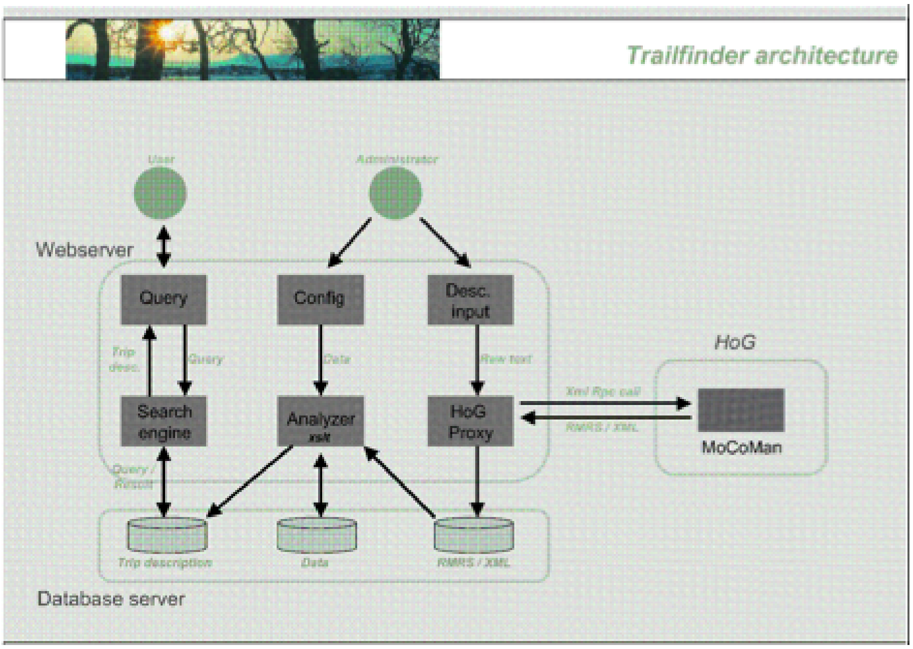


Figure 3: Trailfinder architecture

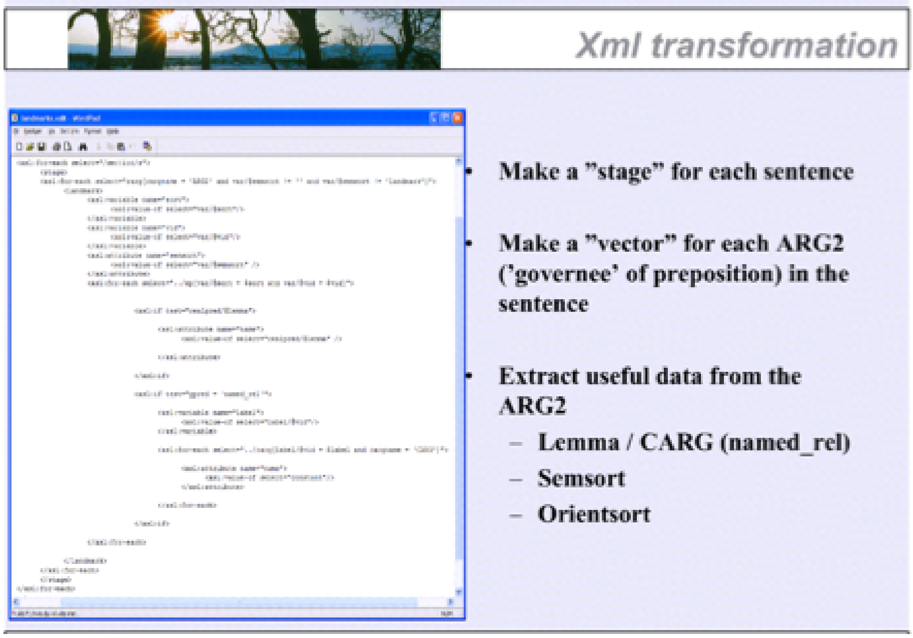
4.1 Filtering of RMRS and grouping of sentences

The information relevant for Trailfinder can be grouped into three classes: its stages, its vectors and its navigational points. As mentioned above, for the present application, we have made use of the fact that consecutive directional specifications divided by comma inside each sentence generally describe temporally and/or linearly consecutive stretches of path or path-consumption, and furthermore that sentences in general correspond to stages of the hike. Among the occurring exceptions to the latter regularity is the first sentence of the hike description given in Figure 5 further below: instead of describing a stage of the tour, this sentence provides an overall characteristics of the tour (as one that goes 'high and free over the mountain tops'). Still, in our extraction algorithm, we represent also this sentence as a stage of the trip, with *over* as a directional preposition with a 'via-point' sortal specification. To impose the time/path line externally, as we consistently do, such that every sentence corresponds to a partial line on the line that the tour as a whole projects, can thus only serve as a first approximation.

Let us now have a closer look at the information that we filter into the final XML document.

4.2 XML transformation

Figure 4 illustrates the final step of RMRS to the tour description XML. On request by the administrator, the analyzer reads all the RMRS documents and transforms them into a markup that only contains the relevant bits of information for the tour description, the final tour/xml. You find this information listed on the right hand side of Figure 4 below. Next to the 'stages', mentioned in the previous section, we are interested in the vectors a person must follow in order to stay on the tour. This information can be found in the value of the ARG2 of the prepositional relation which corresponds to the variable of the argument NP of the preposition. The variable itself will provide us with further information concerning, e.g. the endpoint of a path, while it is the ARG0 of the prepositional relation itself that provides the sortal specification of the vector as such. Navigational points 'en route', finally, are extracted, e.g., from the string value of CARG (constant argument) of named-relations, from where we extract, e.g., proper names relevant for the tour.

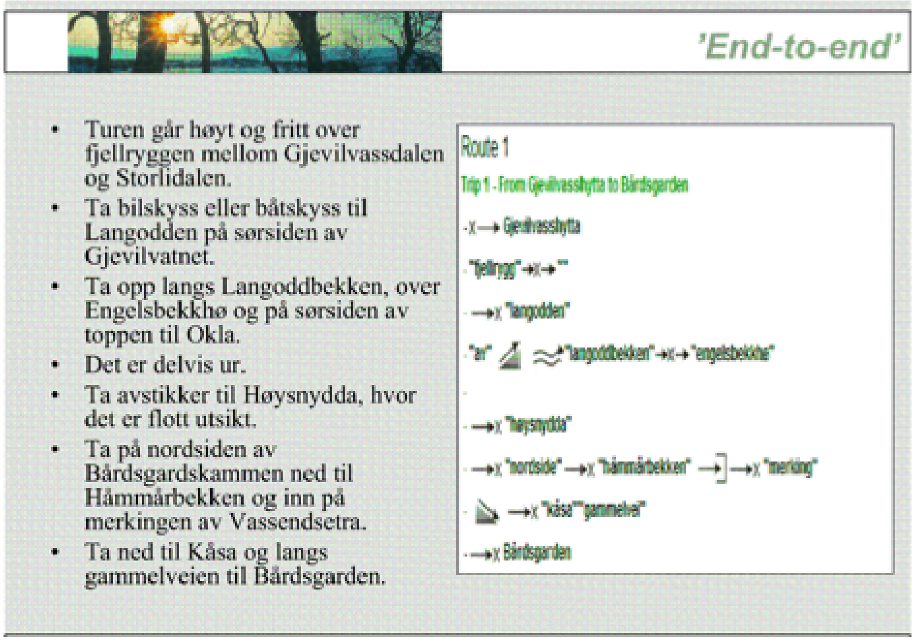


```
[illegible]
    <stage>
    [illegible]
        [illegible]
            [illegible]
                [illegible]
            [illegible]
            [illegible]
                [illegible]
            [illegible]
            [illegible]
                [illegible]
            [illegible]
            [illegible]

                [illegible]

                    [illegible]
                        [illegible]

                    [illegible]

                </xsl:if>

                [illegible]

                    [illegible]
                        [illegible]
                    [illegible]

                    [illegible]

                        [illegible]
                            [illegible]
                        </xsl:attribute>

                    </xsl:for-each>

                </xsl:if>

            </xsl:for-each>

        </landmark>
    </xsl:for-each>
    </stage>
</xsl:for-each>
```

- Make a "stage" for each sentence
- Make a "vector" for each ARG2 ('governee' of preposition) in the sentence
- Extract useful data from the ARG2
  - Lemma / CARG (named_rel)
  - Semsort
  - Orientsort

Figure 4: xml-transformation

It should be mentioned at this point that our primary interest is not so much the pictographic summarization format, illustrated on the right hand side of Figure 5 further below, but rather the language independent semantic encoding of basic spatial and temporal relations. This is the topic of the following final section of this paper. Figure 4 summarizes this section, and Figure 5 illustrates an initial sequence of sentences in its left hand part, and the way they are finally represented in the query interface on the right.



'End-to-end'

- Turen går høyt og fritt over fjellryggen mellom Gjevilvassdalen og Storlidalen.
- Ta bilskyss eller båtskyss til Langodden på sørsiden av Gjevilvatnet.
- Ta opp langs Langoddbekken, over Engelsbekkhø og på sørsiden av toppen til Okla.
- Det er delvis ur.
- Ta avstikker til Høysnydda, hvor det er flott utsikt.
- Ta på nordsiden av Bårdsgardskammen ned til Håmmårbekken og inn på merkingen av Vassendsetra.
- Ta ned til Kåsa og langs gammelveien til Bårdsgarden.

Route 1

Trip 1 - From Gjevilvasshytta to Bårdsgarden

-x→ Gjevilvasshytta

"fjellrygg" →x→ ""

→x "langodden"

"ur" ≈ "langoddbekken" →x→ "engelsbekkhø"

→x "høysnydda"

→x "nordside" →x "håmmårbekken" →] →x "merking"

→x "kåsa""gammelvei"

→x Bårdsgarden

Figure 5: End-to-End

Translation:

(In these translations, brackets supply descriptively relevant parts of the proper names that follow.)

The tour goes high and free over the mountain ridge between [the valley] Gjevilvassdalen and [the valley] Storlidalen. Use car- or boat transportation to Langgodden on the south side of [the lake] Gjevilvannet. Go up along [the creek] Langoddbekken, across [the hill] Engelsbekkshø and on the south side of the top of Okla. The terrain is partly rocky. Take a detour to Høysnydda, from where you have a beautiful view. Go along the north side of the [mountain ridge] Bårdsgårdskammen down to [the creek] Hammarbekken and follow the signs to Vassendsetra. Go down to Kåsa and along the old road to Bårdsgården.

5 A Future for the Trailfinder design

We believe that the more principled interest of the Trailfinder application resides in its utilization of interlingua semantic specifications for spatial and temporal expressions, produced by a deep processing grammar, and the usefulness of this information for

IE purposes. An obvious limitation of the present Trailfinder architecture, is its dependency on the grammar's ability to parse running text. The success of any future application will therefore greatly depend on the future embedding of a grammar such as NorSource into an NLP architecture that combines shallow and deep NLP applications to allow a more adequate coverage of diverse textual input. However, independent of these present limitations in parsing coverage, the main aspect of future interest that emerges from Trailfinder is the circumstance that its subject domain is obviously not restricted to routes in mountains, but that it extends to all textual descriptions of spatial navigation. Following a line of research where feature based lexical semantics is combined with semantic formalism suitable for unification based grammars, two considerations are of special importance: The first one concerns the the outline of a more principled system of conceptual distinctions for the spatiotemporal domain. For the present application we have used a flat list of sortal attributes which are hand-tailored for the present application. In (Hellan and Beermann 2005) we, however, outline a more principled approach to a spatial semantics relevant for the description of prepositional and adverbial meaning. The concept of 'line' and 'x-dimensional' are developed and over 100 spatiotemporal senses, embodied by Norwegian prepositions, are described in what we believe is the beginning of a parsimonious system modelling linguistically relevant spatial concepts.

The second concern for a future extension of the work outlined here is the representation of movement in space. With human-machine communication in mind one possible scenario is to use RMRS-mark-ups to, e.g., inform the movement of artificial agents. For any application that, e.g., relates textual given instructions to robots, success will greatly depend on our ability to model spatial anaphors and also the concept of a time line. The present approach has highlighted some of the issues involved concerning the correlated issue of path-stretches; future work needs to show if, e.g., MRS representation should be used to model temporal sequencing.

References

Callmeier, U.and Eisele, A., Schaefer, U. and Siegel, M.(2004), The deepthought core architecture framework, *in* NN (ed.), *Proceedings of the LREC Conference*, NN, pp. –.

Copestake, A.(2002), *Implementing Typed Feature Structure Grammars*, CSLI Publications, Stanford.

Copestake, A.(n.d.), Report on the design of rmrs, Submitted December 2003.

Cresswell, M.(1985), *Adverbial Modification - Interval Semantics and its Rivals*, D. Reidel, Dordrecht, Holland.

Fillmore, C.(1975), *Santa Cruz lectures on deixis, 1971*, Indiana Unmiversity Linguistics Club.

Halvorsen, P.(1995), *Situation Semantics and Semantic Interpretation in Constraint-Based Grammars*, CSLI Publiscations.

Hellan, L. and Beermann, D.(2005), Classification of prepositional senses for deep grammar applications, *Proceedings from SIGSEM Conference on Prepositions, Univ. of Essex, 2005*, pp. 103–108.

Jackendoff, R.(1990), *Semantic Structures*, MIT Press, Cambridge (Mass).

Kracht, M.(2002), On the Semantics of Locatives, *Linguistics and Philosophy 25*.

Talmy, L.(2000), *Towards a Cognitive Semantics*, MIT Press, Cambridge (Mass).

Trujillo, A.(1995), *Lexicalist Machine Translation of Spatial Prepositions, Ph.D. diss*, Cambridge University.