UDK 004.822
Original scientific paper

WEATHER FORECAST DATA SEMANTIC ANALYSIS IN F-LOGIC

[1]**Ana Meštrović**, [1]**Sanda Martinčić-Ipšić**, [2]**Mirko Čubrilo**
[1]University of Rijeka, Faculty of Philosophy, Department of Informatics, Rijeka, Croatia
{amestrovic, smarti}@ffri.hr
[2]University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
mcubrilo@foi.hr

Abstract: *This paper addresses the semantic analysis problem in a spoken dialog system developed for the domain of weather forecasts. The main goal of semantic analysis is to extract the meaning from the spoken utterances and to transform it into a domain database format. In this work a semantic database for the domain of weather forecasts is represented using the F-logic formalism. Semantic knowledge is captured through semantic categories in a semantic dictionary using phrases and output templates. Procedures for semantic analysis of Croatian weather data combine parsing techniques for Croatian language and a slot filling approach. Semantic analysis is conducted in three phases. In the first phase the main semantic category for the input utterance is determined. The lattices are used for hierarchical semantic relation representation and main category derivation. In the second phase semantic units are analyzed and knowledge slots in the database are filled. Since some slot values of input data are missing in the third phase, incomplete data is updated with missing values. All rules for semantic analysis are defined in the F-logic and implemented using the FLORA-2 system. The results of semantic analysis evaluation in terms of frame and slot error rates are presented.*

Keywords: *Semantic Analysis, F-logic, Lattice, Knowledge Representation, Dialog Management System, Spoken Dialog System.*

1. INTRODUCTION

This paper presents the semantic analysis of weather forecast data. Semantic analysis is a process whereby meaning representations are composed and assigned to linguistic inputs [6]. Presented semantic analysis is a part of the spoken dialog system for weather information in Croatia. In such a system a user can ask questions about weather forecasts and weather conditions in different regions of Croatia and for different time periods. The dialog system would provide relevant answers using weather knowledge collected from the available Web sites over the Internet and stored into the database. The dialog system has to recognize and "understand" the spoken queries and it has to generate spoken answers [22]. Semantic data analysis is the first step in the dialog management module of the dialog system. The input to the semantic analysis is the text recognized by the speech recognition system or text collected from Web pages. Spoken dialog systems for weather forecast information retrieval have been developed for English [3, 21] and Slovenian language [4] and for same other languages as well.

There are three main approaches for semantic analysis [6]: sintax-driven semantic analysis, semantic analysis using formal grammars and information extraction. Croatian language is highly flective and free word order language (like German [4] or Slovenian [4]). This is the main reason why grammars are not easy to implement. Grammars can be useful in combination with some other techniques. In this paper the strategy of semantic parsing technique for Croatian language is combined with slot filling approach. The slot filling approach for semantic analysis has been used in [1, 18]. Additionally we use lattices for representation of hierarchical semantic relations [2] and simple grammars for incomplete data manipulation. Proposed semantic analysis for Croatian data is conducted through three main phases. In the first phase a semantic category for the input is determined. In the second phase semantic units are analyzed and slots of knowledge database are filled. Since some slots of input data are missing in the third phase incomplete data is updated with missing values. Semantic data analysis is based on a previously defined dictionary, semantic categories, phrases and output templates.

In this work we use a deductive object-oriented logic programming language, F-Logic for semantic analysis and for semantic representation of weather forecast data. F-logic provides a very natural way of defining a conceptual model of data semantics and for Web data manipulation [15]. All rules for semantic analysis are defined in F-logic and implemented in FLORA-2 system.

The second section of this paper describes the spoken dialog system. Since domain knowledge and semantic analysis are in F-logic language, there is a short introduction into F-logic given in the third section. The fourth section shortly presents the weather data domain. In section five different approaches for semantic analysis are explained. In section six knowledge representation concepts in F-logic are described: semantic categories organized into an object oriented data schema, semantic word dictionary, word phrases and output templates. Section seven presents our three phase approach to semantic analysis in F-logic: determination of the semantic category of the input text, analysis of semantic units and updating of incomplete data. In the eighth section the evaluation procedure is described and experimental results are presented. Finally some possible improvements are discussed and some future work plans are presented.

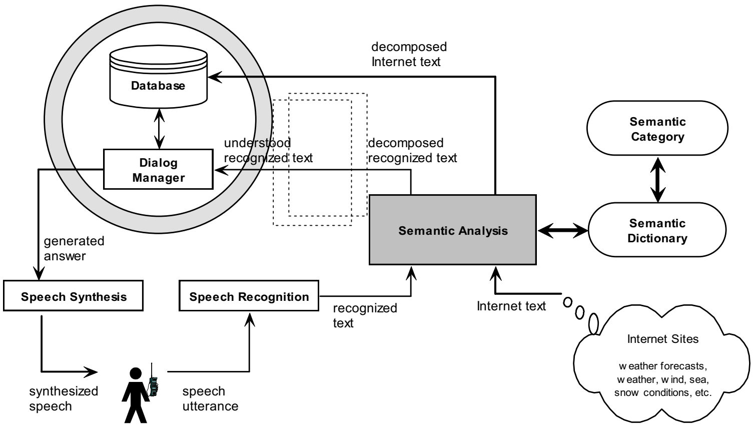

Figure 1. Semantic analysis as a part of a spoken dialog system

2. SPOKEN DIALOG SYSTEM

The development of a spoken dialog system concerns solutions to speech recognition problems, as well as, to speech understanding and human machine interaction problems [16]. The major problems in the development of a continuous speech understanding system arise due to the nature of the spoken language: there are no clear boundaries between words since the phonetic beginning and ending of words are influenced by neighboring words; additionally, variability in speech between different speakers can be noticed, and the speech signal may be affected by noise. To avoid these difficulties, spoken dialog systems are usually limited by different constraints: the vocabulary size is about few thousand words, the communication domain is task oriented, and the sentence structure is usually limited by a simple grammar. The major spoken dialog modules perform word recognition, linguistic analysis, dialog management and speech synthesis. User utterances are first digitalized and transformed into a sequence of speech signal feature vectors. The sequence is passed to the word recognition module, which generates hypotheses of spoken word chains. The recognized words are handed to the linguistic module, which extracts a semantic meaning from the recognized words. The semantic units are passed to the dialog manager. The dialog manager performs a database query if enough parameters are available. According to the dialog history and dialog strategy the user is asked to confirm the parameters or additional parameters are requested. The result of knowledge database query is transformed into sentences. The generated sentences are forwarded to the speech synthesis subsystem and synthesized to the user.

The dialog domain knowledge is captured in the semantic database. In this work the weather information is collected on daily basis from heterogeneous Internet sites and semantically decomposed into slots of the semantic database. The semantic analysis module has to determine the meaning of the words. The semantic analysis module usually consists of a parser and a domain-dependent linguistic knowledge base. In this work semantic data analysis is based on previously defined dictionary, semantic categories, phrases and output templates.

Figure 1 presents the semantic analysis as part of a spoken dialog system. The goal of semantic analysis is two folds. First it has to decompose web weather data into a semantic database and second it has to analyze the text recognized by the speech recognition subsystem. The semantically decomposed recognized text is further processed by the dialog management subsystem. The dialog manager takes the semantic representation of the user utterance and performs the interpretation within the current dialog context and generates system answers. According to the application domain the dialog manager has to answer user inquires or request additional information needed for a successful database query. The dialog manager has to be able to clarify ambiguous dialog situations and generate answers and additional questions to the user in order to keep track of the needed conversational discourse. The dialog manager is in charge of recovering from erroneous situations. The relevant information found by the dialog manager is transformed into the form (sentence, question) that can be physically presented to do user. So the generated response is transferred to the speech synthesis subsystem. Speech synthesis subsystem converts the text of generated answer into a normalized linguistic representation that includes phonetic and prosodic information. The speech processing part of the speech synthesis subsystem is responsible for speech signal generation. And finally the appropriate answer is spoken to the user.

3. F-LOGIC

The frame logic (abbr., F-logic) is a formalism connecting object-oriented approach, frame languages and logical reasoning. The theoretical concept on which the F-logic is based was first presented in 1995. The F-logic syntax and semantics as well as entire theoretical base of F-logic are defined in [8]. Shortcomings in previous attempts of connecting logic programming and object-oriented approach inspired the introduction of F-logic in logic programming. The authors of F-logic extended the classical predicate calculus aiming to define logic that would enable inferring in object-oriented databases. At syntax level, F-logic language is extended with a set of additional symbols, while at semantics level formulas of F-logic assume meaning where it is possible to implement the basic concepts of object-oriented approach. The F-logic retained some quality properties of the classical predicate logic in terms of defining a derivation rule that is analogous to the resolution with a unification procedure used in classical logic.

One of the F-logic language implementations was achieved in the system FLORA-2 [20]. The FLORA-2 system joins three logical formalisms: F-logic, HiLog and transaction logic. FLORA-2 is a flexible system that combines a system of rules, an object-oriented paradigm and is based on higher order logic formalisms. This enables data modelling and representation of knowledge in general [9, 10, 15]. FLORA-2 system is an adequate system for the representation of data semantic components [7, 19].

Basic formulas of the F-logic language show objects and their respective attributes or methods (object[attribute->value]) and appurtenance of an object to a class (object : class) as well as appurtenance of a subclass to a superclass (subclass :: superclass). In our approach, F-logic language is used for semantic dictionary construction, semantic categories and output templates definition. The rules for semantic analysis are also expressed in F-logic.

In this work F-logic language is used for domain knowledge representation through: semantic dictionary, semantic categories and output frames definition. The rules for semantic analysis are expressed in F-logic as well. A simple example of frame instance *wind* with slots: *wind_name*, *wind_intensity*, *place* and *time* is represented as a F-logic formula in Table 1.

Table 1. The *wind* frame in F-logic

OID:wind.
OID[wind_name->'jugo']. OID[wind_intensity->'strong']. OID[place->'Adriatic']. OID[time->'morning'].

4. TASK DOMAIN DESCRIPTION

The speech corpus is essential part of all spoken technologies systems. The quality and volume of speech data in a corpus directly influences the performance of the system. So recording and transcribing the weather domain related speech data was conducted firstly during the spoken dialog system development. The weather information domain was chosen due to the availability of the speech and text data.

The Croatian weather information speech corpus VEPRAD includes weather forecasts and reports spoken within broadcast news of national radio and TV programs [13]. The collected speech material is divided in several groups: weather forecasts read by professional speakers within national radio news, weather reports spontaneously spoken by professional meteorologists, other meteorological information spoken by different reporters and weather forecasts given by meteorologists within the TV news as well as TV and radio news. Collected speech in the corpora is about weather forecasts, meteorological situation, weather conditions in Croatia, Croatian river water level, sea temperature, snow reports, weather forecasts for sailors at sea, bio-meteorological conditions etc.

This heterogeneous weather data is systematically collected from different Internet sources and semantically analyzed few times a day in order to provide accurate and relevant weather information for the dialog system. Semantically decomposed weather information is stored in the semantic database of the dialog management system. In the spoken dialog system development presented Croatian speech corpus was used for the training of the Croatian speech recognition system [11], for the training of the Croatian speech synthesis system [12] and for the semantic data analysis.

5. SEMANTIC ANALYSIS

Semantic analysis as a part of a spoken dialog system is a process of extracting the meaning of the input written text and transforming into a previously defined knowledge database. There are different approaches defined for semantic analysis [6]. The first approach is syntax-driven semantic analysis. The main idea of syntax-driven semantic analysis is: the meaning of a sentence can be composed from the meaning of its parts. The meaning of a sentence is based on ordering, grouping and relations among the words in the sentence. It is partially based on its syntactic structure. Text as input is first passed through the parser to derive its syntax. The output of the syntactic analysis is passed to semantic analysis. But, syntax-driven semantic analysis has limited scope because semantic representations are assigned exclusively from static knowledge captured in the lexicon and grammar. Two other approaches are more appropriate for practical applications. The first approach with semantic grammars is based on formal grammars that correspond directly to entities and relations from the domain of interest [3]. The second approach is information extraction. Information extraction is used with limited domain and when no detailed understanding of meaning is needed. By information extraction knowledge can be described with simple templates. Templates consist of frames with slots that need to be filled with data from the text. In those situations only relevant information from the input text is used for filling the slots and the rest of the text is ignored. Information extraction with slot filling technique is used in many semantic parsers of spoken dialog systems [5, 18].

The Croatian weather data semantic analysis combines information extraction slot filling technique with grammars. This combined approach is chosen mainly because of the limited weather domain and highly flective nature of the Croatian language. The weather forecasts data can by easily captured in fixed frames with slots. Since Croatian language enables a free word order semantic analysis using grammars is a difficult task.

6. DOMAIN KNOWLEDGE REPRESENTATION IN F-LOGIC

In order to recognize the meaning of the sentence, words from the sentence have to be associated with some kind of semantic representation. Therefore it is necessary to define a semantic background for a chosen domain. Semantic background is the domain knowledge and for the weather forecast domain is captured in: the semantic dictionary, categories, phrases and output templates.

6.1. SEMANTIC CATEGORIES

Each word from the weather domain is associated with an appropriate semantic category. Semantic categories can be naturally represented in F-logic language as classes and subclasses. There are eleven main semantic categories (semantic categories of the first level) defined in the domain of weather forecast: *the weather forecast, the sea weather forecast, the bio weather forecast, the meteorology, the state of the river, the wind, the temperature, the place, the time, the description* and *irrelevant* category. Each of these categories consists of semantic subcategories. Overall there are 36 semantic subcategories at the second level. The hierarchy of semantic categories and subcategories is graphically represented in Figure 2. The hierarchical structure of semantic categories and subcategories is captured in a semantic lattice. The hierarchical model of semantic categories is implemented in FLORA-2 system. An extract of semantic model is given in the Table 2. For example *sea weather forecast* category consists of subcategories: *state, warning, visibility, sea condition*.

Table 2. Semantic categories example

English:	**Croatian:**
forecast::weather_forecast.	prognoza::prognoza_vremena.
state::sea_wforecast.	stanje::prognoza_jadran.
warning:: sea_wforecast.	upozorenje::prognoza_jadran.
weather::weather_forecast.	vrijeme::prognoza_vremena.
sea:: sea_wforecast.	more::prognoza_jadran.
visibility:: sea_wforecast.	vidljivost::prognoza_jadran.

In the approach described in this paper, semantic categories are at the same time used as output templates. That means that categories and their subcategories are forming the database schema of the knowledge base for the weather forecast dialog system. The knowledge database is organized in slots of information of the output template.

There are some semantic categories that refer to relative terms (relative time, relative place, relative forecast). The relative terms are related to the other parts of the sentence. The other part of the sentence carries the information necessary to capture the meaning of the relative term. In order to understand relative term the related parts of the sentence have to be analyzed. The relative terms are resolved in the last phase of semantic analysis by special rules that can handle missing data values.

6.2. SEMANTIC DICTIONARY AND PHRASES

Analysis of the domain vocabulary, the usage of specific phrases and analysis of a common sentence structure was performed in order to determine the semantic structure of data and semantic categories. According to semantic categories of the words in a sentence or from a part of the sentence that is called sentence unit, it is possible to define the dominant semantic category of a whole sentence and to extract the general meaning from the sentence.

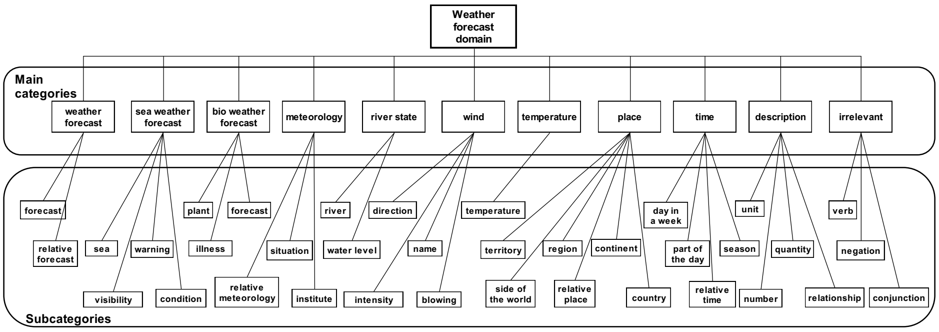

Figure 2. The hierarchy of semantic categories

The word dictionary was prepared from collected texts and contains almost 2300 different words. Each word from the dictionary is assigned to the appropriate semantic category. The number of words in each category is presented in Table 3.

Table 3. Number of words in main categories

Category	Number of words
weather forecast	475
sea weather forecast	17
bio weather forecast	148
meteorology	116
river state	7
wind	82
temperature	38
time	219
place	513
description	117
irrelevant	539
OVERALL	**2271**

The semantic dictionary with all words and their semantic categories is also given in F-logic language. In F-logic each word from the dictionary is represented as an object of an adequate class. In Table 4 the words with their semantic categories as part of the semantic dictionary are presented. In the presented example *morning* and *night* are in *part-of-the-day* category. *Frequently* and *week* are in the *relative time category*.

Table 4. The segment of the word dictionary

English:	Croatian:
"afer":relative_time.	"zatim":relativno_vrijeme.
"country":relative_place.	"zemlje":relativno_mjesto.
"cyclone":meteorology.	"ciklona":meteorologija.
"frequently":relative_time.	"često":relativno_vrijeme.
"morning":part_of_a_day.	"jutro":dio_dana.
"night":part_of_a_day.	"noć":dio_dana.
"week":relative_time.	"tjedan":relativno_vrijeme.
"fog":forecast.	"magla":prognoza.

Croatian language is highly flective and words have different forms, cases and genders. The dictionary comprises all word formats that occur in data. For each word an adequate basic word form (lemma) is chosen. By example a noun in genitive is transformed into basic word form as a nominative noun. Rules for generating the basic word form are represented in F-logic as well.

The weather domain model introduces a set of phrases used in weather forecast reports. That set of phrases includes all phrases that usually appear in spoken or written weather forecasts. A subset of weather forecast phrases can be defined by using simple grammars. Those phrases mainly consist of two words. The first word denotes an attribute/quantity and the second word is a kind of weather (example: '*mostly sunny'*, '*partly cloudy'*). It is not necessary to find all possible phrases, but only the phrases related to the kind of weather. Set of phrases is represented by grammars in F-logic language in a way such that each phrase is treated as an object that belongs to an adequate semantic category implemented as a class in F-logic. The phrases with relative terms, like *relative time* and *relative place* are initially noted as relative terms. The relative terms are analyzed in the last phase of semantic analysis when updating incomplete data with missing values.

6.3. OUTPUT TEMPLATES

Semantic categories that are relevant for generating answer from knowledge base are called output templates. Output template is in F-logic formed as an object belonging to a class with attributes. It represents a frame with slots and adequate values. A slot is represented as an information item for which a value is required [18]. Classes in F-logic are represented as frames. This is the reason why concepts of frames, slots and values are naturally captured in F-logic formalism. Adequate information is extracted from written input and transformed into slot values of the output templates. The task of semantic analysis is to fill all possible slots with input data values. Slot values are used to generate answers in the dialog system. Example 1 presents the weather forecast output template (*time, place, weather forecast, temperature and wind*) in F-logic:

Example 1. Weather forecast output template

```
object_id:dom_class.
object_id [time->Time].
object_id [place->Place].
object_id [w_forecast->> forecast].
object_id [temperature->>Temperaturee].
object_id [wind ->> Wind].
```

Slot values *time* and *place* are common attributes for every output template. Other slot values like *temperature* and *wind* are depending on the type of output template. Some main semantic categories are at the same time output templates as well. Output templates which overlap with semantic categories are: the *weather forecast, the sea weather forecast, the bio weather forecast, the meteorology and the river state*.

7. SEMANTIC ANALYSIS IN F-LOGIC

This section describes three phases of the semantic analysis procedure. First the semantic category of the input text is determined and input text is decomposed into semantic units. Each unit is analyzed and slots of the output template are filled. Finally incomplete data is updated with missing values and remained empty slots are filled. Figure 3 gives an overview of semantic analysis. The Croatian sentence "*In the morning there will be very sunny in the continent and rain in the other parts of the country*" is analyzed in three phases: the semantic category of the sentence is determined, semantic units of the sentence are analyzed and incomplete data is updated with missing values.

The input to the semantic analysis is Croatian weather forecast text. In the first phase one dominant category of the input text is determined. Sentences of the input text are divided into sentence semantic units. Data from semantic units are extracted in order to fill the slot values of output templates. Each object is generated from one semantic unit. When a semantic unit contains a relative term it can not be directly transformed into slot values. In this case parts of the unit are marked as relative. The problem of relative terms is resolved in the last phase of semantic analysis. The missing slots values are updated with the data value from the preceding or succeeding semantic unit.

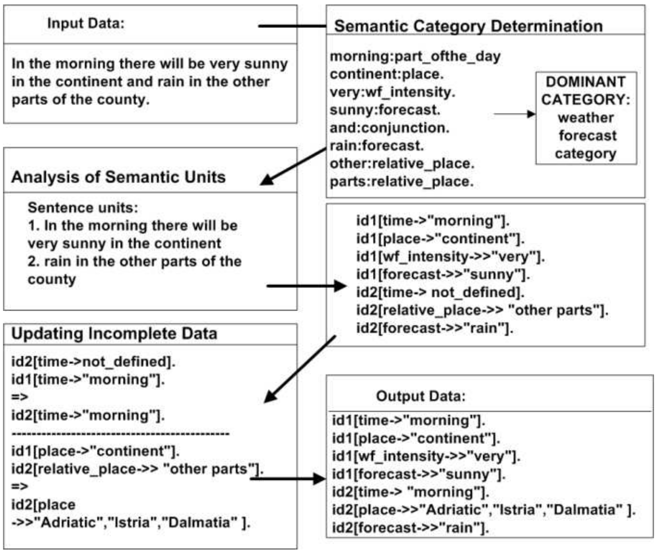

Figure 3. An example of semantic analysis

7.1. SEMANTIC CATEGORY DETERMINATION

Semantic analysis starts with determining a semantic category of written text. Each paragraph can be associated with one semantic category of weather forecast domain. General semantic categories are: weather forecast, sea weather forecast, meteorology, bio weather forecast, and river forecast. These general categories are overlapped with output templates. The first phase of semantic analysis process is to determine a main topic of the paragraph. The main topic is also called dominant semantic category. The dominant category of a text represents the general meaning of sentence.

Some approaches in determining the dominant category are conducted statistically [5]. This work uses lattices as a mathematical formalism for finding the dominant category of input sentence [1]. In lattice theory set L is a set of all classes given by semantic categories. And A is a subset of L. Since classes are organized to form a lattice it is always possible to find an element (class) that is a superclass of all elements in a given subset A. The idea of our approach is to determine the dominant category by generating the set of all categories in text paragraph and to define a superclass of that set as a dominant category. The result of that first phase of semantic analysis is exactly one dominant category. All objects generated from the written text will belong to this dominant category.

All semantic categories are hierarchically organized and represented as classes in F-logic language. We form a lattice structure in a way such that there is exactly one class that is a subclass of all classes and there is exactly one class that is a superclass of all classes given by semantic categories. These two classes represent the minimal and the maximal category. With class inclusion as a partial order relation, this set of classes forms a lattice. Figure 4 shows the graphical representation of semantic categories lattice. The maximal element is a class that represents a dominant weather category. At Figure 4 the dominant category is *general_forecast*. *General_forecast* category is the most general class and every other class is a subclass of *general_forecast*. The minimal element is a class irrelevant, and in further semantic analysis process ignored. The *irrelevant* category contains all words that are not important for analyzing the text.

In F-logic language it is possible to define rules for lattice manipulation. We implement that rules in FLORA-2 system in a module called lattice_manip.flr with predicate dom_cat/2 defined as:

```
dom_cat(List,Max_class):-
          _Num=max{_Z1|_Z1=
count{_X1[_K1]|_X1::_K1@semantika2, member(_X1,List)@prolog(lists),
member(_K1,List)@prolog(lists)}},
          _Max=max{_Z[Max_class]|_Z=
count{_X[Max_class]|_X::Max_class@semantika2,member(_X,List)@prolog(lists),
member(Max_class,List)@prolog(lists)}},
          _Max>=_Num.
```

where *Max_class* is a class that is a minimal superclass of all classes given in a list L. For finding a dominant class we use aggregates max and count defined in FLORA-2. The process of finding dominant semantic category uses predicates from module lattice_manip.flr. We apply predicate dom_cat/2 on a set of classes assigned to a set of words from text paragraph. The result is dominant semantic category which is passed to the next phase of analysis.

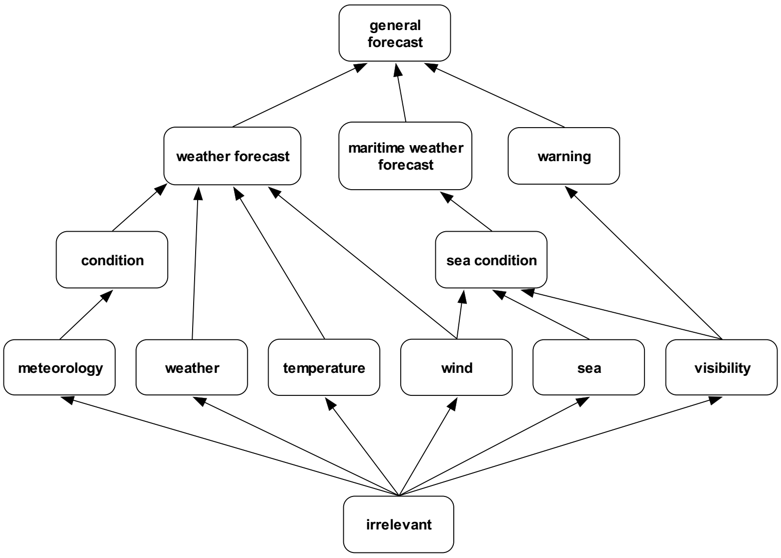

Figure 4. Semantic categories lattice

7.2. ANALYSIS OF SEMANTIC UNITS

Semantic unit is a part of sentence that contains necessary information to fill attribute values of output template slots. Text paragraph is divided into sentences and further, sentences are divided into semantic units. Sentence and semantic units are input data represented in FLORA-2 using two predicates: sentence/2 and sentence_unit/3. Predicate sentence/2 contains *sentence* and *sentence ID.* Predicate sentence_unit/3 defines sentence unit together with *sentence ID* and *unit ID*. *Sentence ID* is generated from date and time of the captured data origin. *Sentence unit ID* is copied from *sentence ID* and concatenated with the unit number within the sentence. Decomposition of one analyzed sentence 3 into semantic units is presented in Example 2.

Example 2. Decomposition of sentence into semantic units

English:	**Croatian:**
sentence('in the morning there will be very sunny in the continent and rain in the other parts of the county.','hrdanas_15072105_1'). ... sentence_unit('in the morning there will be very sunny in the continent ','hrdanas_15072105_1',1). sentence_unit('and rain in the other parts of the county ','hrdanas_15072105_1',2). ...	sentence('ujutro će biti vrlo sunčano u unutrašnjosti, a u ostalim predjelima kiša.','hrdanas_15072105_1'). ... sentence_unit('ujutro će biti vrlo sunčano u unutrašnjosti','hrdanas_15072105_1',1). sentence_unit('a u ostalim predjelima kiša','hrdanas_15072105_1',2). ...

Sentence semantic units are detected by using appropriate delimiters. Delimiters that may be used are commas and conjunctions. The speech recognition subsystem output is text with natural pauses noted in angle brackets (for example <breath>, <sil>) instead of punctuation marks. All events in angle brackets are used as delimiters as well. Physical semantic units are extracted from a sentence on the places of delimiters and special events. If semantic unit carries new information about time or about place, new object in F-logic is generated. If semantic unit contains no new information, then it should be joined to the previous or the next semantic unit. Semantic units that contain new information are referred to as logical semantic units. To generate logical semantic units it is necessary to verify if semantic unit contains new information or is a part of an existing one. The verification on whether a semantic unit contains new information for weather forecast category is implemented in F-logic by a rule:

```
new_information(X,Category) :-
        sentence_unit(X,_,_),
        cat_extract(X,Category),
        (Category::place@semantika; Category::time@semantika).
```

If there is new information about time or about place a new object is introduced. If semantic unit doesn't contain information about time, or place all the gathered data refers to a previously defined object.

At the end each logical semantic unit is mapped into exactly one object with unique ID. Output template slots are filled with appropriate word from semantic unit. Some units can be directly mapped since all needed data slots are extracted from the semantic unit. In many situations it is impossible to define all data slots. In this case the incomplete data is updated in the last phase of the process.

7.3. UPDATING INCOMPLETE DATA

Some output template slots can have missing data because they carry relative terms. In the third phase of semantic analysis the incomplete data of relative terms are updated with missing values. Figure 3 shows an example where *place* and *time* attributes for the weather forecast semantic category are missing. Missing categories are updated with the information gathered from the preceding and succeeding semantic units. There are two possible situations: missing data can be derived from relative terms and missing data is really missing, and therefore can not be derived.

The first situation when missing data can be updated is presented in the line (id2[relative_place->>"other parts"].) of example from Figure 3. The phrase *other places ("drugi dijelovi")* in a text indicates that a rule to resolve relative_place for "other parts" has to be used. This rule determines which places are parts of Croatia that are different from inland. The rule derives results *Adriatic, Istria* and *Dalmatia* to update the missing data. This rule is given in F-logic for solving relative place terms with key phrase *'other parts' ('drugi dijelovi'*) is:

```
_X1[place->>Place]:-
        _X1[relative_place->>RP],
        _X[place->>Place1],
        other_parts(RP),
        pokrajna(Place),
        Place\=RP.
```

The second case, when missing data can not be found is written at line (id2[time-> not_defined].) of example from Figure 3. Information about the *time* should be derived from the first part of the sentence by the F-logic rule:

```
_X1[time->Time]:-
        _X1[time->not_defined],
        _X[time->Time].
```

Other relative terms and phrases that are resolved by similar F-logic rules are: *midday*, *the second part of the day, toward the end of the day, western regions, northern parts, along the coast, in the mountains* etc.

8. EVALUATION AND RESULTS

The semantic analysis process described in previous section was evaluated with Web weather data collected in a three month period. The texts for evaluation consist of one general weather forecast for Croatia and three maritime weather forecasts per day. For the test text we manually prepared a reference semantic database that includes all correctly determined slots and slots values [14]. A test semantic database was generated by the F-logic system from the same evaluation texts. The evaluation is performed on the frames and slot value level.

From the test text 2696 slot values of 398 frame instances were generated. First the number of correctly determined frames corresponding to semantic units is evaluated. The semantic unit evaluation results in terms of number of correct frames and error rates are shown in table 5.

Table 5. The frame evaluation results

	Weather forecast	Maritime forecast	OVERALL
Reference	267	131	398
Generated	265	129	394
Correct	264	129	393
Error rate	**1.49%**	**1.52%**	**1.51%**

The slot values are evaluated by slot error rate [14]. The slot error rate is analogous to the word error rate in speech recognition performance. Slot error rate combines the deleted, inserted and substituted types of error. An algorithm implemented in Flora-2 system is used to align generated against reference slot values. The corresponding slots are then matched and scored as either correct or not. If not correct, the error is marked as a substitution (incorrect slot), deletion (missing slot), or insertion (spurious slot). According to a number of errors measure of performance can be computed using the formula:

$$SER = \frac{N_S + N_I + N_D}{N_C + N_S + N_D} \qquad (1)$$

where N_S is a number of substituted slots, N_I is a number of inserted slots, N_D is a number of deleted slots and finally N_C is number of correct slots. The slot error rate (SER) results are shown in Table 6.

Table 6. The slot value evaluation results

	Weather forecast	Maritime forecast	OVERALL
Reference	1655	1041	2696
Generated	1665	1038	2703
Correct	1503	918	2421
SER	**18.97%**	**23.35%**	**20.66%**

The 20% range of slot error rate was expected due to the flective nature of Croatian and due to existence of referenced terms in the input texts. It is reasonable to assume that additional improvement can be achieved in modification of relative terms manipulation.

9. CONCLUSION

This paper presents semantic analysis of Croatian weather data in object-oriented logic programming language F-logic. Domain knowledge representation and semantic analysis is implemented in F-logic using FLORA-2 system. The domain data semantics is captured in a semantic dictionary, semantic categories, phrases and output templates. The proposed stepwise semantic analysis approach combines parsing of Croatian language with slot filling technique. The semantic analysis is conducted in three phases. Initially the dominant semantic category of the input text is determined by using lattices for representation of hierarchical semantic categories relations. Further the input text is divided into semantic units. Sentence semantic units are detected on the places of punctuation marks and special events recognized in the speech. Then the slots of output templates are filled with values. Finally, since some missing data value can occur in slot filling phase, the incomplete data is updated.

Proposed approach for semantic analysis is used in the spoken dialog system for Croatian weather forecast. The aim of semantic analysis in a spoken dialog system is two folded. First it has to decompose web weather data into the semantic database and second it has to analyze the text recognized by the speech recognition subsystem. Since the semantic analysis is the first step in the spoken dialog management subsystem activities towards speech understanding and answer generating should be considered. Presented semantic analysis process was preliminary evaluated using test data from collected Web pages. Preliminary evaluation results are quite promising so an adequate evaluation method should be used for complete semantic analysis evaluation. Final evaluation of Croatian weather data semantic analysis will be the overall evaluation of the weather spoken dialog system.

Presented semantic analysis was evaluated using test texts collected from Web pages. The achieved results of 20 % slot error rate are quite promising for further development of Croatian semantic analysis for spoken dialog systems. In the future we will consider using formal grammars in F-logic in order to improve updating of the incomplete data with missing values. Further we will expand the domain of interest with some specific kind of forecasts, such as forecasts for agriculture. That means that we will need to add some new data sources and consider heterogeneous data integration.

REFERENCES

[1] C. Cardie. *Empirical Methods in Information Extraction*, AI Magazine, 18(4) (1997), pp. 65-79

[2] C. Carpineto, G. Romano. *Concept Data Analysis: Theory and Applications*. John Wiley and Sons, 2004.

[3] L. Devillers, H. Bonneau-Maynard. Evaluation of Dialog Strategies for a Tourist Information Retrieval System. *In the proceedings of the ICSLP 1998*, Sidney, Australia, October 1998.

[4] R. Engel, Robust and Efficient Semantic Parsing of Free Word Order Languages in Spoken Dialog Systems. *In the proceedings of the Interspeech 2005*, Portugal. 2005, pp. 3461-3464.

[5] M. Hajdinjak, F. Mihelič. Semantična analiza vremenskih napovedi. *In the proceedings of the 5th International Multi-Conference*, Language Technologies, Ljubljana, 2002.

[6] D. Jurafsky, J. H. Martin. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, Upper Saddle River, New Jersey, 2000.

[7] M. Kifer, R. Lara, A. Polleres, C. Zhao, U. Keller, H. Lausen, D. Fensel. A Logical Framework for Web Service Discovery. *In ISWC 2004 Workshop on Semantic Web Services: Preparing to Meet the World of Business Applications*, 2004.

[8] M. Kifer, G. Lausen, J. Wu. Logical Foundations of Object-Oriented and Frame-Based Languages. *Journal of the ACM*, 42(4), 1995, pp. 741-843.

[9] A. Lovrenčić. Knowledge Base Amalgamation using Higher-Order Logic-Based Language HiLog, *Journal of Inaformation and Organizational Sciences*, Vol. 23, No. 2, Varaždin, Croatia, 1999, pp. 133-147.

[10] A. Lovrenčić, M. Čubrilo. Amalgamation of Heterogeneous Data Sources Using Amalgamated Annotated HiLog. *In Proceedings of the of the 3rd IEEE Conference on Intelligent Engineering Systems*, INES'99, 1999.

[11] S. Martinčić-Ipšić, I. Ipšić. Croatian HMM Based Speech Synthesis. *Journal of Computing and Information Technology- CIT*, Vol. 14(4), 2006, pp. 299-305.

[12] S. Martinčić-Ipšić, I. Ipšić. Croatian Telephone Speech Recognition. Slobodan Ribarić, (ed), Leo Budin, (ed.). *29.th MIPRO 2006*, Opatija, Croatia, 2006. Proceedings. Vol. CTS-CIS. pp. 182-186.

[13] S. Martinčić-Ipšić, M. Matešić, I. Ipšić. Korpus hrvatskog govora. *Govor: časopis za fonetiku*, XXI (2), 2004, pp. 135-150.

[14] J. Makhoul, F. Kubala, R. Schwartz and R. Weischedel**.** Performance Measures For Information Extraction in DARPA Broadcast News Workshop, 1999.

[15] A. Meštrović, M. Čubrilo. Semantic Web Data Integration Using F-Logic. *In the proceedings of 10th International Conference on Intelligent Engineering Systems*, London, 2006.

[16] W. Minker, S. Bennacef. *Speech and Human-Machine Dialog*. Kluwer Academic Publishers. Boston. 2004.

[17] S. Seneff, C. Wang. Statistical Modelling of Phonological Rules Through Linguistic Hierarchies. *Speech Communication*, Vol. 46, 2005, pp. 204-216.

[18] B. Souvignier, A. Kellner, B. Rueber, H. Schramm, and F. Seide.The Thoughtful Elephant. Strategies for Spoken Dialog Systems,. *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 1, 2000, pp. 51-62.

[19] G. Yang, M. Kifer. Inheritance and Rule in Object-Oriented Semantic Web Languages, *In Second International Workshop on Rules and Rule Markup Languages,* 2003.

[20] G. Yang, M. Kifer, C. Zhao. *FLORA-2: User's manual, Version 0.92*, Department of Computer Science, Stony Brook University, http://flora.sourceforge.net/, 2003.

[21] V. Zue et al. JUPITER: A Telephone-Based Conversational Interface for Weather Information. *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No 1, January 2000, pp. 85-96.

[22] J. Žibert, S. Martinčić-Ipšić, M. Hajdinjak, I. Ipšić, F. Mihelič. Development of a Bilingual Spoken Dialog System for Weather Information Retrieval. *In 8th European Conference on Speech Communication and Technology*, September 1-4, 2003, Geneva, Switzerland. EUROSPEECH ′03. Proceedings. ISCA. Vol. 1, pp. 1917-1920.

Received: 15 April 2007
Accepted: 30 October 2007