Generating Domain-Restricted Resources for Web Interaction in Several Languages: Hindi, English and Spanish

Marta Gatius, Piyush Paliwal
Department of Computer Science
Technical University of Catalonia
Barcelona, Spain
e-mail:gatius@lsi.upc.edu, piyushpaliwal90@gmail.com

***Abstract*—The aim of our research is to develop domain-restricted resources for web interaction supporting different languages: English, Hindi and Spanish. Many practical natural language systems use linguistic resources adapted to a specific domain because the processing is faster and more robust against errors. Besides, those grammars can be adapted to the language used by different types of users. To facilitate the process of generating linguistic resources for each domain and language, we use ontologies representing the entities and relations in a specific domain. The use of domain ontologies also favors the integration of knowledge from several web sites. For developing the grammar rules for each domain and language, we use Grammar Framework, a powerful tool for writing multilingual grammars that supports several alphabets. Our work is focused on the generation of assisting the user when accessing the web in two different scenarios: searching for information on cultural events and searching for a new medical specialist.**

Keywords-multilingual web interfaces; domain ontologies; semantic grammars.

I. Introduction

As web is becoming more central to our daily activities, the need of web assistants adaptable to different types of users (different languages, ages, skills, and cultural sensitivity) increases. In this context, conversational systems guiding the user to access web information are becoming a good opportunity to enhance web usability.

The main challenge in conversational systems consists of understanding correctly the user's needs. To solve this problem, most practical conversational systems are adapted to a particular domain, thus limiting possible misunderstandings and errors. However, the cost of building the domain-restricted linguistic resources needed for processing the user interventions is high and they are difficult to adapt to a new domain. There are different approaches to face this problem, all requiring human intervention. Statistical techniques require big corpus, manually tagged. In knowledge-based approaches, conceptual and linguistic resources have to be developed by skilled professionals. Although several of the works on domain adaptation use in-domain training data to reduce the adaptation cost, most of these works are based on the use of domain conceptual models.

There are complex research communication systems that use general grammars to syntactically analyze user interventions and, from the results obtained perform semantic analysis, using conceptual domain-restricted knowledge. However, most practical conversational systems use semantic grammars adapted to a specific domain that perform syntactic and semantic analysis in parallel. The reason is that rules in semantic grammars correspond directly to the domain entities and relations, resulting in faster processing and more robust against errors. Besides, domain-restricted grammars are especially appropriate for multilingual systems because similar processing can be done for each language supported. Although domain-restricted grammars are easier to write and maintain than general grammars, they have to be build for each approach. The use of semantic-domain models may also be used to facilitate this task.

The use of a semantic model representing domain entities facilitates the obtaining of the domain-restricted resources needed for interpreting users needs. Furthermore, the semantic models in multilingual (and multimodal) systems provide a common semantic interpretation for different languages (and modes of interaction).

Domain knowledge can be represented in several semantic formalisms. Most used formalism to represent domain entities are database models, frames and ontologies. Database models are useful in applications using databases or web services and have been used in relevant works, such as the dialogue systems described by Polifroni et al. [1] and D'Haro et al. [2]. Frames and ontologies provide a more flexible way of representing domain concepts and have also been used in many interaction systems.

In ontologies, relations and preconditions between entities are defined, thus providing a richer conceptual representation. For this reason, they are very appropriate for complex systems, providing consistent generic processes, reusable for several domains, such as those described by Nesselrath and Porta [3] and Dzikovska et al. [4].

The use of ontologies also favours integration of knowledge sources of different types. Thus, ontologies are especially appropriate for communication systems integrating different types of knowledge, such as the dialogue system Smartweb (described by Sonntag et al. [5]), supporting several languages and modes of interaction and the web assistant Active (described by Guzzoni et al. [6]), integrating language and active learning technologies.

Additionally, the organization of knowledge in ontologies in classes/subclasse helps with under/over specification phenomena and other simple inferences that may appear in communication in complex domains, such as the medical domain (as explained by Milward and M. Beveridge [7]).

In this paper, we present our work on the use of ontologies to generate the domain-restricted grammars needed for a web interface system supporting different languages: English, Hindi and Spanish.

This paper is organized as follows. Section II introduces our approach and Section III describes its application to two different scenarios: searching for cultural events and searching for a medical specialist. Conclusions and future work are given in the last section.

II. Proposed solution

The work described in this article is related to previous work on a dialogue system guiding the user when accessing web services, described in [8]. For that system, messages were generated using a general method, defined by J.A. Bateman et al. [9], based on a sintactico-semantic taxonomy that relates concepts and attributes in conceptual ontologies to the linguistic structures needed for their expression in several languages (Spanish, English and Catalan). Using this taxonomy grammars and sentences expressing queries and answers about the values of the conceptual attributes can be automatically generated, although they have to be manually supervised. The resulting sentences have been incorporated in the dialogue system with minor changes, however, the resulting grammars have not, because those grammars only recognized correctly a few sentences.

Manually defined grammars can support many more different forms of expressing domain terms, including abbreviations, mistakes, informal expressions, new terms and other forms difficult to represent in a formally. Besides, they can be adapted considering different types of users, different ages and different expertise levels.

Our current work is focused on the study of an appropriate representation of the conceptual and linguistic knowledge involved in communication to facilitate the manually creation of semantic grammars for new domains, new languages and new types of users. As in several of the works mentioned, we represent the general knowledge involved in several domains in reusable representation bases and the specific knowledge for each domain is incorporated manually. In our proposal, concepts appearing in several domains are represented in general ontologies and the linguistic knowledge associated with it in general grammar rules. Then, adapting the system to a new domain requires building the domain ontologies as well as the grammars rules related to this domain knowledge for each language.

Similar approaches have been followed in complex research dialogue systems supporting rich communication in different types of applications. The main difference is that our work was focused on assisting the user when accessing the web. The language supported by our system is limited to that used by the user when asking for information. Thus, the effort of generating the semantic grammar can be limited if first the user needs and expressions in the particular scenario are studied. Additionally, we have used a multilingual grammar environment, Grammatical Framework (GF), [10] specially appropriate for our approach. In GF grammars are represented in two separated modules: conceptual (abstract grammar) and syntactic (concrete grammar). The abstract grammar captures the semantics to be communicated and can be the same for all languages supported in a particular application. The concrete grammar component relates the abstract syntax to the linear strings representations and it is different for each language. This separation of grammars in two components therefore helps the human experts ease the generation of rules in each of the languages.

Following our proposal, the abstract grammars are defined considering the concepts in the domain ontologies. Then, the related rules in each concrete grammar are defined by the language experts.

The integration of the Hindi language with other languages is also an important difference from previous works on communication systems. Our proposed organization of knowledge in separated conceptual and linguistic knowledge and general and domain-specific has also facilitated work with language with a different organization and a different alphabet.

III. Generating domain-restricted grammars

Domain ontologies provide a formal organization of the conceptual knowledge appearing in user intervention when accessing web information in a particular domain. Thus, it can facilitate the integration of domain knowledge that appears in several web sites in different formats and languages. This formal representation of the domain knowledge facilitates the generation of linguistic resources needed for processing the user interventions and generating the system responses. This section describes the use of domain ontologies to generate semantic grammars needed for web interaction in English, Hindi and Spanish. Our proposal is based on a clear separation of general and domain-specific knowledge. General entities common to several domains (such as those related to time and space) are defined in an upper ontology. Grammars rules and vocabulary supporting the expressions related to those general conceptual entities are represented in a general grammar. General grammar also includes rules supporting expressions common to all scenarios (such as those expressing misunderstandings).

For each new scenario, we first study the user needs and the corpus of user's questions, if available. Then, the domain specific entities are described and related to those general entities in the upper ontology. Existing domain ontologies could also be integrated, when needed. For example, in the scenario the user is looking for a specialist an existing ontology describing parts of the body would be incorporated. In the final step, grammars supporting question and descriptions of the concepts in the domain ontologies are developed.

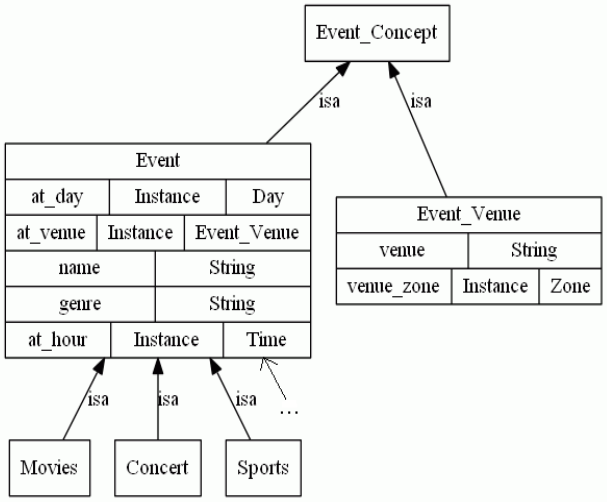
We had studied our proposal in two different scenarios where the user searches for practical information on the web. The first scenario we have studied is that of a user looking for a cultural event in the city. We selected this scenario because we already had previously collected a corpus of user interventions asking for information on cultural events. The second scenario we studied is that of a user searching for a specialist and we did not collect any corpus related to it. Next subsections describe the ontologies and grammars generated for two different scenarios.

A. Building the Domain Ontologies

In the first scenario we studied, the user wants to consult information on the cultural events that take place in the city. We consider the user may ask for information for a specific event (giving its name) or, alternatively, may ask for the events satisfying a specific description. There are many web sites giving information on cultural events. The central concept in those web sites is the same: a cultural event described by a set of attributes. Several of those attributes are the same in most web sites: title, venue, date and time. In some of those web sites additional information could also be given such us participants, price, age.

Figure 1 shows the description of the two entities involve in this scenario: **Event** and **Event-venue**. As can be seen, the concept **Event** is described by the attributes: **name**, **genre**, **at-venue**, **at_day** and **at_hour**. The concept **Event_venue** is described by two attributes: **venue** and **venue_zone**. These two domain concepts are related to the general concepts **Zone** and **Unit_Of_Time**.

In order to support most common user's questions in this domain we have defined the grammar rules supporting the consulting of the attribute values of the two concepts **Event** and **Event_venue**.



Event		
at_day	Instance	Day
at_venue	Instance	Event_Venue
name	String	
genre	String	
at_hour	Instance	Time

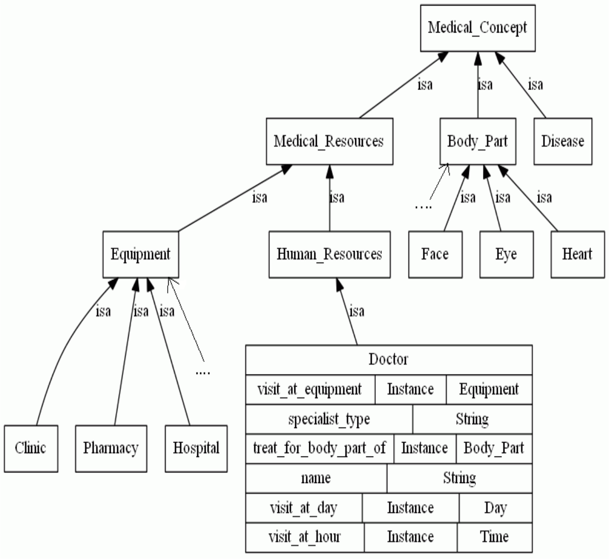
Event_Venue		
venue	String	
venue_zone	Instance	Zone

Figure 1. Conceptual knowledge in the Cultural Events domain.

We have considered a second scenario in a different domain: the health domain. In particular, we have considered the scenario in which the user accesses the web for searching information about a new medical specialist. In this scenario, most common user interventions will consist of giving information about his/her disease and asking for information about specialists.

There are different web sites containing practical information about medical specialists (i.e., the place where they visit and their schedule). In most of these web sites medical specializations and diseases are described using medical terms that could prove difficult for not experts. To solve this problem we have developed the semantic grammar needed for processing the questions asked by general public using non medical terms.

Figure 2 shows the concepts involved in the scenario of searching for a specialist. There are three top concepts: **Disease**, **Body_Parts** and **Medical_Resources**.**Medical Resources** is subclassified in three concepts: **Doctor**, **Equipment**, and **Others**. The main concept, **Doctor**, is described by a set of attributes: **name**, **specialist_type** **treat_of_body_part**, **visit_at_equipment**, **visit_at_day**, **visit_at_hour**. The concept **Body_Parts** has been included because one common way to ask for a specialist consists of giving the common name of the body part where a problem has been detected, as in the example *"I have stomachache, and I need a doctor"*. As shown in the Figure 2, the concepts in this domain have also been related to the general concepts **Zone** and **Unit_of_Time**. From this domain ontology, the abstract grammar is generated.



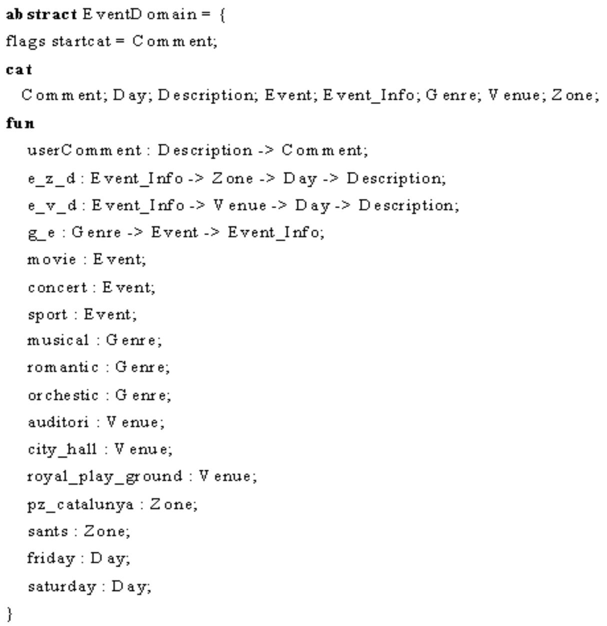
Doctor		
visit_at_equipment	Instance	Equipment
specialist_type	String	
treat_for_body_part_of	Instance	Body_Part
name	String	
visit_at_day	Instance	Day
visit_at_hour	Instance	Time

Figure 2. Conceptual knowledge in the health domain.

We have used Protégé, a knowledge-based framework to develop the ontologies.

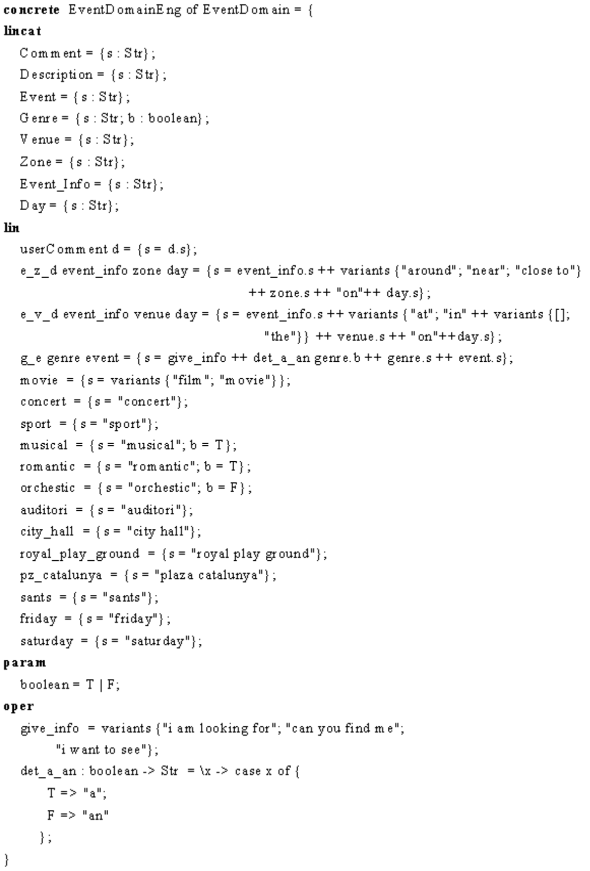
B. Building the Grammars

As we mentioned in the introduction, we have used GF, an open-source environment for writing grammars in different languages. The main reason for using GF is that it supports languages using different alphabets, including Devanagari. However, because not all computers are configured to support Devanagari, we have also transliterated Hindi grammars to the Latin alphabet.



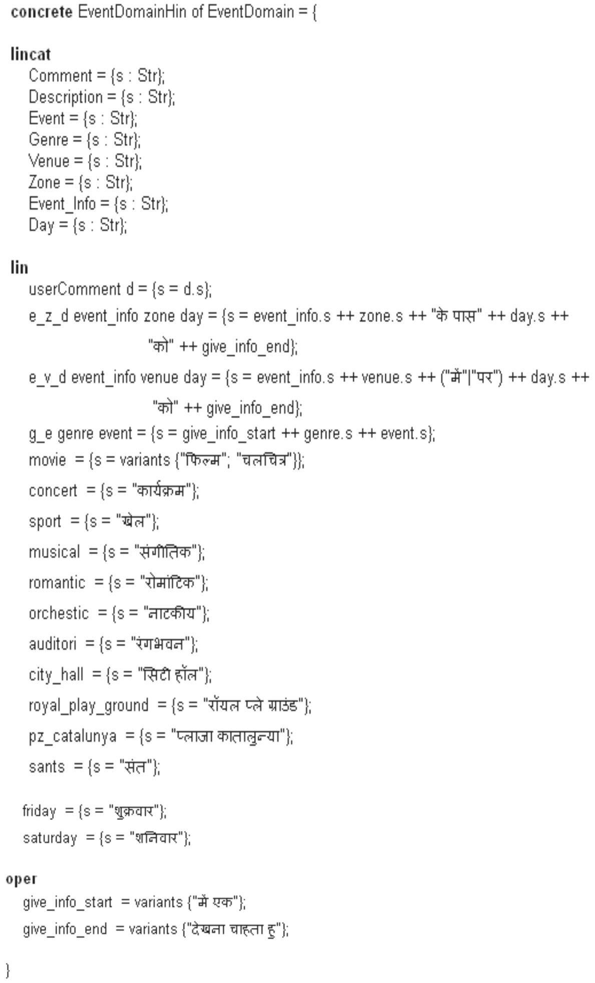
```
abstract EventDomain = {
flags startcat = Comment;
cat
  Comment; Day; Description; Event; Event_Info; Genre; Venue; Zone;
fun
   userComment : Description -> Comment;
   e_z_d : Event_Info -> Zone -> Day -> Description;
   e_v_d : Event_Info -> Venue -> Day -> Description;
   g_e : Genre -> Event -> Event_Info;
   movie : Event;
   concert : Event;
   sport : Event;
   musical : Genre;
   romantic : Genre;
   orchestic : Genre;
   auditori : Venue;
   city_hall : Venue;
   royal_play_ground : Venue;
   pz_catalunya : Zone;
   sants : Zone;
   friday : Day;
   saturday : Day;
}
```

Figure 3. The abstract grammar for the event domain.



```
concrete  EventDomainEng of EventDomain = {
lincat
   Comment = {s : Str};
   Description = {s : Str};
   Event = {s : Str};
   Genre = {s : Str; b : boolean};
   Venue = {s : Str};
   Zone = {s : Str};
   Event_Info = {s : Str};
   Day = {s : Str};
lin
   userComment d = {s = d.s};
   e_z_d event_info zone day = {s = event_info.s ++ variants {"around"; "near"; "close to"}
                                         ++ zone.s ++ "on"++ day.s};
   e_v_d event_info venue day = {s = event_info.s ++ variants {"at"; "in" ++ variants {[];
                                            "the"}} ++ venue.s ++ "on"++day.s};
   g_e genre event = {s = give_info ++ det_a_an genre.b ++ genre.s ++ event.s};
   movie  = {s = variants {"film"; "movie"}};
   concert  = {s = "concert"};
   sport  = {s = "sport"};
   musical  = {s = "musical"; b = T};
   romantic  = {s = "romantic"; b = T};
   orchestic  = {s = "orchestic"; b = F};
   auditori  = {s = "auditori"};
   city_hall  = {s = "city hall"};
   royal_play_ground  = {s = "royal play ground"};
   pz_catalunya  = {s = "plaza catalunya"};
   sants  = {s = "sants"};
   friday  = {s = "friday"};
   saturday  = {s = "saturday"};
param
   boolean = T | F;
oper
   give_info  = variants {"i am looking for"; "can you find me";
          "i want to see"};
   det_a_an : boolean -> Str  = \x -> case x of {
        T => "a";
        F => "an"
      };
}
```

Figure 4. The English concrete grammar for the event domain



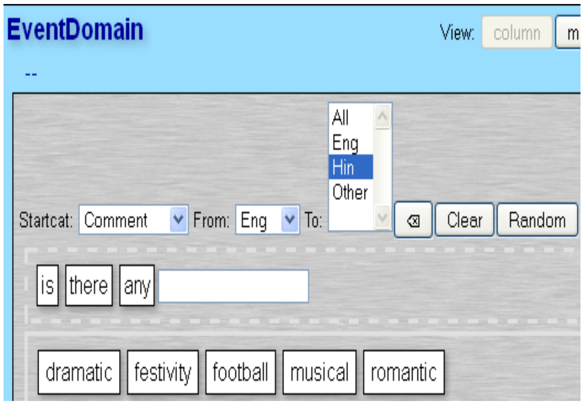
```
concrete EventDomainHin of EventDomain = {

lincat
   Comment = {s : Str};
   Description = {s : Str};
   Event = {s : Str};
   Genre = {s : Str};
   Venue = {s : Str};
   Zone = {s : Str};
   Event_Info = {s : Str};
   Day = {s : Str};

lin
   userComment d = {s = d.s};
   e_z_d event_info zone day = {s = event_info.s ++ zone.s ++ "के पास" ++ day.s ++
                        "को" ++ give_info_end};
   e_v_d event_info venue day = {s = event_info.s ++ venue.s ++ ("में"|"पर") ++ day.s ++
                        "को" ++ give_info_end};
   g_e genre event = {s = give_info_start ++ genre.s ++ event.s};
   movie  = {s = variants {"फिल्म"; "चलचित्र"}};
   concert  = {s = "कार्यक्रम"};
   sport  = {s = "खेल"};
   musical  = {s = "संगीतिक"};
   romantic  = {s = "रोमांटिक"};
   orchestic  = {s = "नाटकीय"};
   auditori  = {s = "रंगभवन"};
   city_hall  = {s = "सिटी हॉल"};
   royal_play_ground  = {s = "रॉयल प्ले ग्राउंड"};
   pz_catalunya  = {s = "प्लाजा कातालुन्या"};
   sants  = {s = "संत"};

  friday  = {s = "शुक्रवार"};
  saturday  = {s = "शनिवार"};

oper
  give_info_start  = variants {"मैं एक"};
  give_info_end  = variants {"देखना चाहता हु"};

}
```

Figure 5. Hindi concrete grammar for event domain

The use of GF also facilitates multilingual applications because all grammars are divided in two components: the abstract syntax component, a tree-like representation that captures the domain knowledge and the concrete syntax that relates the domain knowledge to the corresponding linear strings. We have defined the abstract syntax component (the same for all languages) using the domain representation in the ontology. Additional information obtained from the collected examples of sentences can also be included. Then, the concrete syntax for each language is obtained from the abstract syntax.

The general mechanism to define the abstract grammar from the ontology consists of representing ontology concepts and attributes appearing in the conversation as categories in the grammar (**cat**) and as the right part of one or more rules **(fun)**. There are two types of rules: syntactical rules (if the right side of the rule has more than one category), and lexical rules (if the right side of the rule has only one category). Instances and values of conceptual attributes were represented as lexical rules while syntact rules represent the combination of concepts appearing in user's interventions.

Figure 3 shows a fragment of the abstract grammar obtained from the ontology representing the cultural events entities. Fragments of the concrete grammars for English and Hindi are represented in Figure 4 and Figure 5, respectively. Notice that, even when the two languages are very different, the concrete grammars in both languages have the same organization because the abstract grammar is the same.



EventDomain

Startcat: Comment From: Eng To: Hin

is there any

Figure 6. The system guides the user when building the query.

The GF environment supports another interesting functionality for assisting the user where building the query by presenting the next acceptable options on the screen when writing. When using this functionality, the errors when processing user interventions are minimized, resulting in a friendlier communication. Figure 6 shows how this functionality has been used to guide the user when using the grammars we have developed.

IV. Conclusion and Future Work

In this paper, we describe the generation of domain-restricted grammars for a web interface supporting different languages: English, Hindi and Spanish. To facilitate the process of writing the grammars for different languages, we use domain ontologies. Grammars have been implemented in GF, an open-source environment for multilingual applications. Our work has focused on the language involved when assisting the user when accessing the web in two different scenarios: searching for cultural events and searching for a medical specialist.

The main goal of our work has been to find a general method to facilitate the generation of grammars that are easy to adapt to new languages, new domains and even new users (i.e. young people using informal languages including new words and mistakes).

We have tested the grammars developed by building a toy prototype integrating them, a set of canned system's responses and a small set of databases. Our goal has not been to construct a complete grammar. For this reason, evaluation to study how many sentences can be supported by the grammars has not been done. Instead, we have measured the reusability of grammar components (abstract and concrete rules) across domains and languages, finding high rates of reusability.

Our proposal also includes the semi-automatic generation of system responses by using a syntactic-semantic taxonomy. We have also started working on the adaptation of this taxonomy to the Hindi language, finding it simpler of what we expected. Future work will also include the definition of different types of users for the scenarios considered and the presentation of the results obtained in the most appropriate form for each type of user.

References

[1] J. Polifroni, G. Chung and S. Seneff, "Towars the Automatic Generation of Mixed-Initiave Dialogue Systems from Web Contents," Proc. EUROSPEECH, 2003, pp. 193-195.

[2] L.F. D'Haro, R. Córdoba, J. M. Lucas, R. Barra-Chicote, and R. San-Segundo, "Speeding Up the Design of Dialogue Applications by Using Database Contents and Structure Information," Proc. SIGdial, 2009, pp. 160-169.

[3] R. Nesselrath and D. Porta, "Rapid Development of Multimodal Applications with Semantic Models," Proc. IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, July, 2011, pp. 37-47.

[4] M. Dzikovska, J. F. Allen, and M. Swift, "Finding the balance between generic and domain-specific knowledge: a parser customization strategy," Proc. Workshop on Knowledge and Reasoning in Practical Dialogue Systems, 2003.

[5] D. Sonntag, R. Engel, G. Herzog, A. Pfalzgraf, N. Pfleger, M. Romanelli, and N. Reithinger. "SmartWeb handheld – multimodal interaction with ontological knowledge bases and semantic web services," AI for Human Computing, T.S.Huang et al. (Eds.): LNAI 4451, Springer, 2007, pp. 227-295.

[6] D. Guzzoni, C. Baur, and A. Cheyer, "Active: A unified platform for building intelligent web interaction assistants," Proc. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2006, pp. 417-420.

[7] D. Milward and M. Beveridge, "Ontology-Based Dialogue Systems," Proc. Knowledge and Reasoning in Practical Dialogue Systems, 2003, pp. 9-18.

[8] M. Gatius and M. Gonzalez,"Using Domain Ontologies for Improving the Adaptability and Collaborative Ability of a Web Dialogue System," International Journal of Computer Information Systems and Industrial Management Applications, V5, 2012, pp. 185-194.

[9] J.A. Bateman, B. Magnini, and F. Rinaldi, "The Generalized {Italian, German, English} Upper Model," Proc. ECAI Workshop on Implemented Ontologies, 1994.

[10] A. Ranta, Grammatical Framework: Programming with Multilingual Grammars, CSLI Publications, Stanford, 2011.

ISBN: 978-1-61208-310-0