Ontological Knowledgebase for English Syntactic Grammar using Government and Binding Theory

Manpreet Singh Sehgal
Department of Computer Science & Engineering, Apeejay College of Engineering, Sohna, Gurgaon

Twinkle Sehgal
Department of Computer Science, Gurgaon Institute of Technology & Management, Gurgaon

Manjeet Singh
Phd, Department of Computer Engineering, YMCA University of Sc. & Technology, Faridabad

ABSTRACT

Language is important in communication, whether it is in written format for record purpose or it is in spoken format for coordinating our everyday tasks. To analyze and learn a particular language the underlying theories have to be understood. This understanding gets enriched if the knowledge of these theories can be made to store in some knowledgebase. Ontology is one way to store this conceptual database. This paper describes the design and implementation of ontology for syntactic knowledge of English grammar. The knowledgebase in the form of ontology can be further used by other systems involving Language analysis and Language learning. First the Conceptual Design of ontology is provided in the light of both conventional and Government and Binding theory Grammar. Then the ontology is implemented using Protégé platform where the complex relationship between the various grammar entities can be seen as graph using OntoGraf.

Keywords

Ontology, OntoGraf, Government and Binding theory

1. INTRODUCTION

Ontology is defined as “an explicit specification of a conceptualization” [3, 4]. Adapting this framework to the area of Natural Language Analysis is the thrust point of this paper. For definition sake, Natural Language Analysis means taking written language input and mapping it into a representation which can be used by an application [1]. There are six kinds of languages depending upon the order of Subject Object and Verb. For this paper, language of choice is English, which is an SVO (Subject Object Verb) type language. In order to find out more entitles and complex structures or relationships among various syntactic entities we are using Government and Binding Theory [2] as underlying formalism.

2. RELATED WORK

There are various learning applications which are ontology based. [5] Proposed the methodology for developing ontologies for learning systems and presented some ontology based applications in education. An interesting guide to create your first ontology is given in [6]. Henze et al. [7] proposed a framework for personalized e-learning and showed how the semantic web resource description formats can be utilized for automatic generation of hypermedia structures. The research in GB Theory speaks volumes of its profound importance in the field of Natural Language Analysis. [8] Presented concepts of a Government and Binding theory for Hindi Language and showed that the concept can be used for Indian languages as well. C.R. Rene Robin et. al. [9] represented the knowledge representation in ontology for the domain of software risk management. They emphasized on the fact that The ontology model of Software risk management is an effective approach for the intercommunion between people from teaching and learning community, the communication and interoperation among various knowledge oriented applications, and the share and reuse of the software. Many researchers have contributed tremendous literature on Semantic Tools like protégé and Natural Language analysis. Yet the combination package is missing. So this paper introduces this novel idea of merging both the fields to draw something worthwhile.

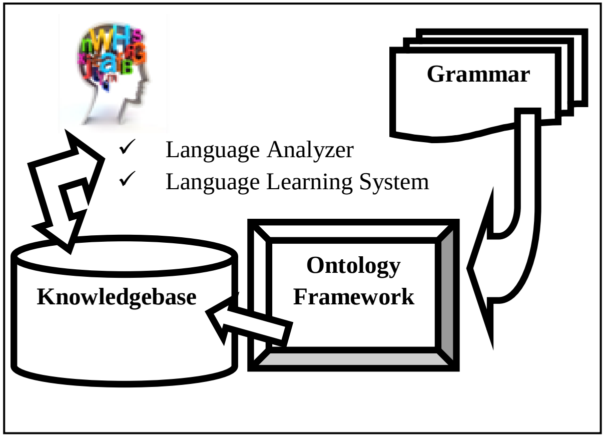


Fig. 1 The Conceptual Framework

3. PROPOSED WORK

Fig 1 depicts the Concept of the proposed work. Grammar expert provides Grammar knowledge to lay out a Grammar Shown in Fig 1. Once Grammar is laid out by the Expert than through the Semantic Web tool (Ontological framework) the grammar is codified and a knowledge base is created. Hence forth this knowledge base can be used as the backbone store for the language learning applications and language analyzers.

4. ARCHITECTURE

The Conceptual design of ontology presented here is set of the definitions configured in the form of knowledgebase that can be easily described in the Ontology framework. For a detailed level analysis Government and Binding Theory is used. The subsequent theories like X bar Theory and Trace Theory are briefly touched for the completion of the concept. The represented concept emphasizes the way to represent the English Language grammar and corresponding Language so as to use it in the Machine Learning, Language analysis and Language Learning. This section illustrates the design and artifacts of proposed work from two angles. 1) Conventional Grammar View 2) Grammar and Binding Theory view.

4.1 Design (Conventional Grammar View)

From conventional point of view, all the parts of speech and the relationships among these are encoded in the ontology. The knowledge about how different parts of sentence combine to form a meaningful sentence and what are the various punctuation symbols and their usage is perfectly displayed in the Ontological Database. Grammar Space is classified into appropriate classes and properties are assigned to these classes

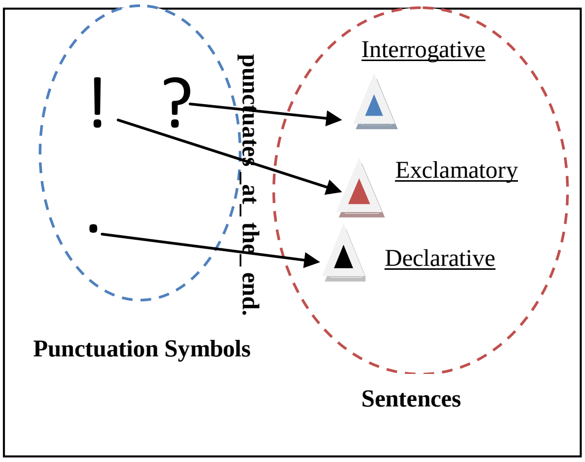


Fig. 2 Concept of Classes and Object Properties

and relationships among the classes are established. Fig 2 describes a subset concept of the work from conventional point of view.

In the above mentioned figure two classes are illustrated, one containing Punctuation symbols and other containing various types of sentences and the relationship among the members of both of the classes is defined, with arrows labeled punctuates_at_the_end. The constraints *"only"* is applied, so that knowledge "Punctuation Mark (!) Punctuates at the end only Exclamation Sentence". On the same lines other classes are linked up with each other as per grammar rules. And more can be added as per the new features of Grammar. As a diagrammatic convention circle represents the super class and triangle represents the sub class and individuals in the super class are just depicted there in a class as for example sign of exclamation (!) is depicted in the circle representing Punctuation Mark. On the similar lines there are many categories that are identified in English grammar and hence class and subclass relation and other equivalent relations can be easily codified in Ontological Framework. The following entities are identified and their relationships are established

4.1.1 Alphabet

There are many kinds of alphabets.

- Armenian
- Cyrillic
- Georgian
- Greek
- Latin

Latin alphabet which is a generalization of Modern English alphabet also generalizes two other alphabets

- Archaic Latin alphabet
- Classical Latin alphabet
- Modern English alphabet

For the properties matter Alphabets possess two properties, one of them is that all these alphabets have varied number of letters but the maximum letters that an alphabet can contain does not exceed 44 letters. Other property is that an alphabet is used by some written language.

4.1.2 Language

As per conventional view, one should be able to write as well as utter the same language, hence, classifying the language on the basis of these abilities; Following are the two classifications of any of the word order language.

- Written Language
- Spoken Language

Written Language uses alphabets and Spoken Language has grapheme as the constituents. Written English language uses only modern English alphabet. Sentence is a part of a language in turn words are parts of sentences.

4.1.3 Sentence

In the field of linguistics, a sentence is an expression in natural language, often defined to indicate a grammatical unit consisting of one or more words that generally bear minimal syntactic relation to the words that precede or follow it. A sentence can include words grouped meaningfully to express a statement, question, exclamation, and request or command [2]. From architecture point of view, sentence is considered as a form which starts with a noun phrase and ends with some verb phrase. Accordingly sentence is categorized to the classes like

- Exclamatory Sentence
- Imperative sentence,
- Interrogative sentence,
- Declarative Sentence,

Exclamatory sentence represents some feeling and punctuates at the end with some exclamation mark. Declarative sentence is some argument or states some fact and this type of sentence is not any kind of command, request, but it states some idea. Declarative sentence punctuates at the end with simple period. Interrogative sentence punctuates at the end with question mark, and these sentences ask some questions. Imperative sentence ask for some action, and this action can be asked with request or without request if it is with request than it is called entreaty else it is referred to as command. A Sentence has an organization based on the type of word order in the grammar, for example SVO (Subject Verb Object) is the word order of English language.

4.1.4 Subject and Predicate

Every sentence is considered having two parts- 1) the part which names the person or thing we are speaking about, called Subject of the sentence and 2) the part which tells something about the Subject, called Predicate of the sentence. In most of the sentence, the Subject of the sentence usually comes first, but occasionally (in case of passive voice) it is put after the Predicate.

4.1.5 Phrase and clause

The group of words which makes sense but not containing Subject is called a phrase, and a part of a sentence that contains both Subject and Predicate called a clause thereby making complete sense of idea to be communicated. The relationship among phrases clauses and sentences and other classes is also established.

4.1.6 Parts of Speech

Words are divided into different kind of classes, called parts of speech, according to their use. Following parts of speech are considered for this work

1. Noun 2. Adjective 3. Pronoun 4. Verb

5. Adverb 6. Preposition 7. Conjunction 8. Interjection

All parts of speech has certain properties so are implemented in Ontology.

4.2 Design (Government & Binding Theory View)

In this section, the various concepts from GB theory are incorporated in the proposed Ontology. GB theory supports the concept of universal grammar in sense that here all phrases have the same structure that implies that large portion of the grammar of any particular language is common to all languages and is therefore part of the Universal Grammar. Further GB view is that Universal Grammar can be broken down into two components, Levels of Representation and system of Constraints. So as to represent this knowledge in the knowledgebase four classes are defined that corresponding to Lexicon, D-Structure, S-Structure, Phonological Form and Logical Form. In order to represent the moves among them a class called Movement Rule is created that encompasses "Move Alpha" subclass, "LF Move-Alpha" subclass, "stylistic and phonological rules" subclass, that respectively maps from D-Structure to S-Structure and Structure to LF (Logical Form) and from S-Structure to PF (Phonological Form). Phonology class is designed, and "is_the_interface_with" object property is defined, so that information point: "PF is the interface with are directly represented can be represented into a knowledgebase. The same object property is utilized in the other information point "LF is the interface with the

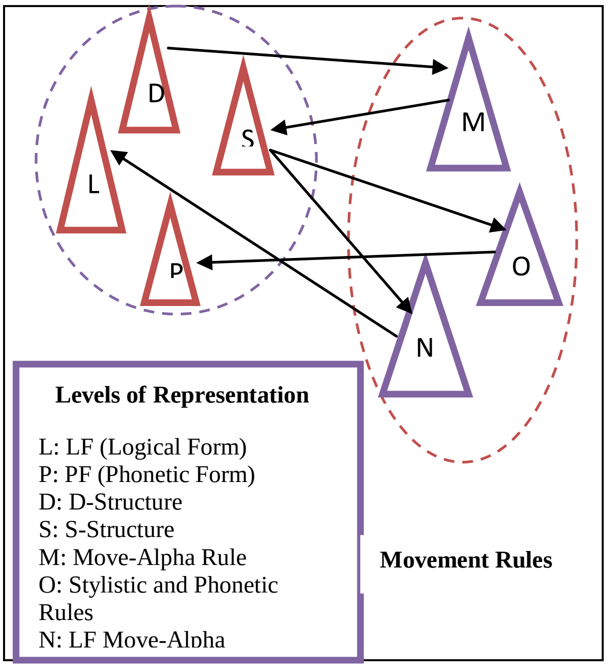


Fig. 3 Levels of Representation and Movement Rules.

semantics, with Semantic as a new class. Fig.3 summarizes the above description.

A word, such as a noun, verb, adjective or preposition is lexical category. In structural terms, they are called heads. Phrases are meaningful groupings of words built up from the lexical category of the same type that they contain. Examples of phrases are: NP, VP, and AP (=AdjectiveP), and PP. But the particular head is choosy about what can combine with it to form a phrase. A 'complement' is a phrase that a lexical category takes or selects as its necessary argument. The type of complements are taken by a particular head is an inherent property of the head. Therefore, Heads, Complements and Specifiers are the three parts of the phrase; Specifiers are the words that precede (in English) the lexical head and further define the scope and range of lexical head.

5. IMPLEMENTATION

In this section, conceptual designed defined in previous section, is implemented using Protégé framework 4.1.alpha from Stanford University and University of Manchester. This framework is used to represent the knowledge in the forms of classes, individuals belonging to the classes and properties among individuals that belong to the classes. As in this work English grammar is considered, the immediate classes to the 'thing' class were chosen to be like alphabet, Word, Sentence, Punctuation Symbol, Representation Level, Movement Rules etc. The detailed level implementation snapshots of the Grammar from both conventional and Government and Binding theory point of view are shown in the various Figures in this section. Figure 4 shows the Alphabet as a Class in the form of a Tree.

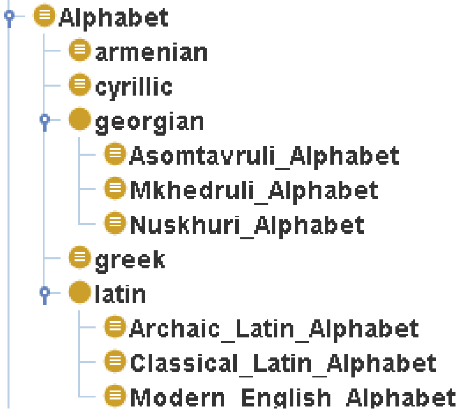
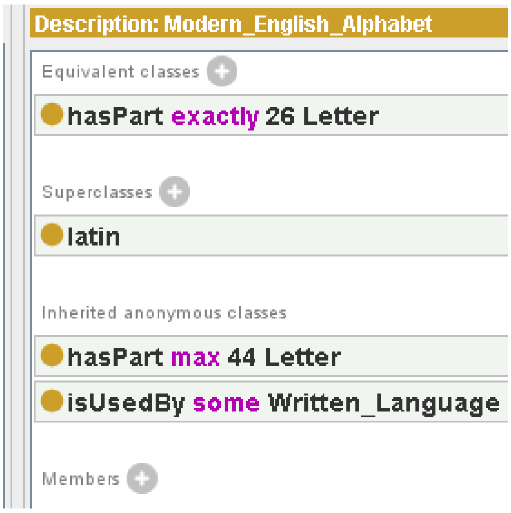

- Alphabet
  - armenian
  - cyrillic
  - georgian
    - Asomtavruli_Alphabet
    - Mkhedruli_Alphabet
    - Nuskhuri_Alphabet
  - greek
  - latin
    - Archaic_Latin_Alphabet
    - Classical_Latin_Alphabet
    - Modern_English_Alphabet

Fig. 4 Tree level implementation of Alphabet

5.1 Properties of Alphabet

The description of every alphabet is stored in the form of equivalent classes, super classes, inherited anonymous classes for example Modern English alphabet has a Super class that is Latin (One of the Latin alphabet is Modern English Alphabet). This alphabet has exactly 26 letters and it inherits all the properties of generic alphabets, like this alphabet is used by some Written Language as shown in Fig. 5.



Description: Modern_English_Alphabet

Equivalent classes

● hasPart exactly 26 Letter

Superclasses

● latin

Inherited anonymous classes

● hasPart max 44 Letter

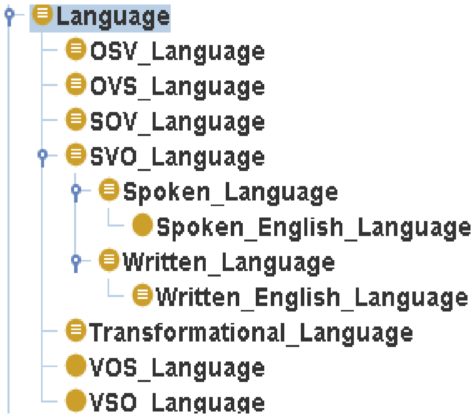
● isUsedBy some Written_Language

Members

Fig. 5 Modern English Alphabet Properties

5.2 Language Representation (GB theory point of view)

As earlier discussed as per the word order there are six languages as shown in Fig.5.3 .One additional language is represented here in order to show some related knowledge.

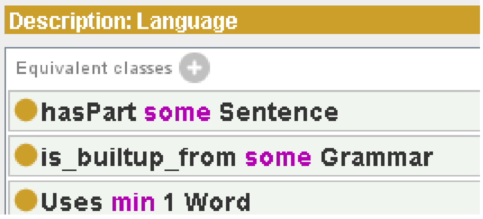


```
Language
  OSV_Language
  OVS_Language
  SOV_Language
  SVO_Language
    Spoken_Language
      Spoken_English_Language
    Written_Language
      Written_English_Language
  Transformational_Language
  VOS_Language
  VSO_Language
```

Fig. 6 Types of Languages

5.3 Language Properties (Conventional View)

Fig. 7 shows that any language has some sentence as a part. And it is built up from some Grammar and at the minimum level it contains at least a word.



Description: Language

Equivalent classes

● hasPart some Sentence

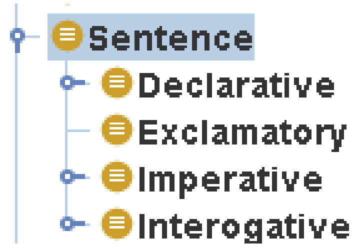
● is_builtup_from some Grammar

● Uses min 1 Word

Fig. 7 Language in terms of Equivalent classes

5.4 Sentence Representation

Fig. 8 depicts the types of sentences.

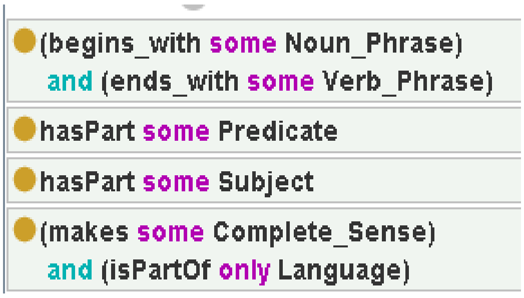


```
Sentence
  Declarative
  Exclamatory
  Imperative
  Interogative
```

Fig. 8 Types of Sentences.

5.5 Sentence Properties

As seen in the snapshot given in Fig. 9, sentence is seen in the form of composing Noun Phrase and Verb phrase.



● (begins_with some Noun_Phrase)
and (ends_with some Verb_Phrase)

● hasPart some Predicate

● hasPart some Subject

● (makes some Complete_Sense)
and (isPartOf only Language)

Fig. 9 Four Important points of Sentence.

The knowledge that sentence has some Predicate and Subject as parts and it makes some complete sense and moreover the sentence is a part of only language is easily represented in the form of Equivalent Classes of a Sentence. Sentence features discussed above are presented in following OntoGraf, given in Fig. 10

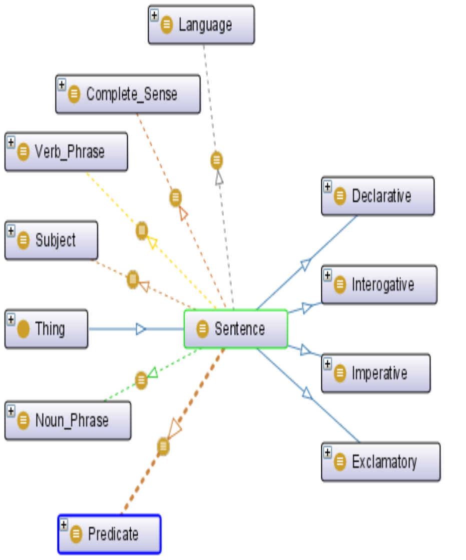


Fig. 10 OntoGraf of a Sentence.

Every entity in the OntoGraf is expandable for example; Declarative Sentence is depicted in Fig 11.

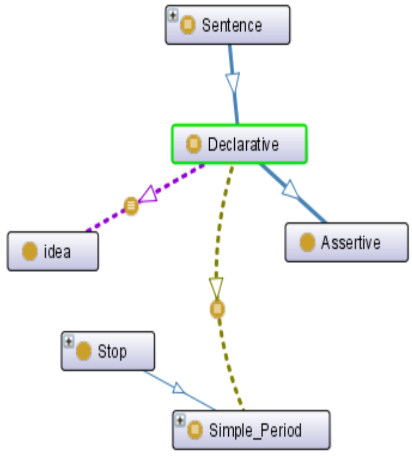
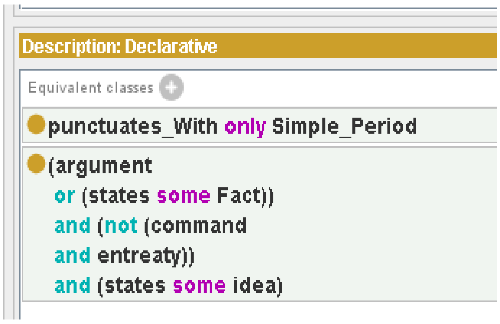


Fig. 11 Declarative Sentence OntoGraf

The OntoGraf here represents that declarative sentence is an assertive sentence and this kind of sentence states some idea. Furthermore this sentence ends with a simple period which is a type of stop. One can click on the "Stop' to find out various types of stops. Here is the description, given in Fig. 12, of what its OntoGraf represents.



Description: Declarative

Equivalent classes

punctuates_With only Simple_Period

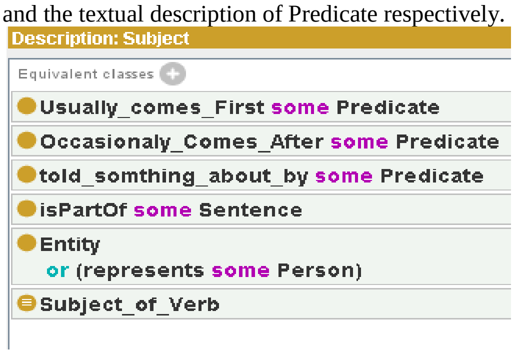
(argument
or (states some Fact))
and (not (command
and entreaty))
and (states some idea)

Fig. 12 Declarative Sentence Properties

On the similar way other types of sentences are represented.

5.6 Subject and Predicate Representation

Fig. 13 and Fig. 14 depict a textual description of a Subject and the textual description of Predicate respectively.



Description: Subject

Equivalent classes

Usually_comes_First some Predicate

Occasionaly_Comes_After some Predicate

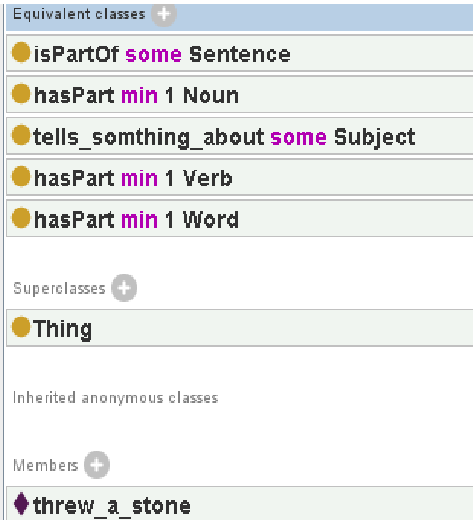
told_somthing_about_by some Predicate

isPartOf some Sentence

Entity
or (represents some Person)

Subject_of_Verb

Fig. 13 Description : Subject



Equivalent classes

isPartOf some Sentence

hasPart min 1 Noun

tells_somthing_about some Subject

hasPart min 1 Verb

hasPart min 1 Word

Superclasses

Thing

Inherited anonymous classes

Members

threw_a_stone

Fig. 14 Description: Predicate

The corresponding OntoGraf of the predicate is shown in Fig 15.

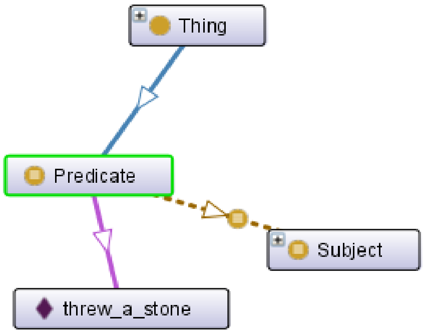


Fig. 15 Predicate OntoGraf

In the above OntoGraf it is represented that Predicate is a Thing and is associated with a Subject (shown in Fig. 13) in a relation that Subject occasionally comes after the Predicate, and predicate tells something about the Subject.

5.7 Phrase Representation (GB theory View)

Phrase is the grouping of some subject and predicate, but still some complete sense is missing. As per Government and Binding theory, phrase is constructed by each kind of head in the sentence. In Fig. 16 each kind of phrase named after each kind of head. For the purpose of X-Bar theory general phrase is named as XP (pronounced as X- Phrase)

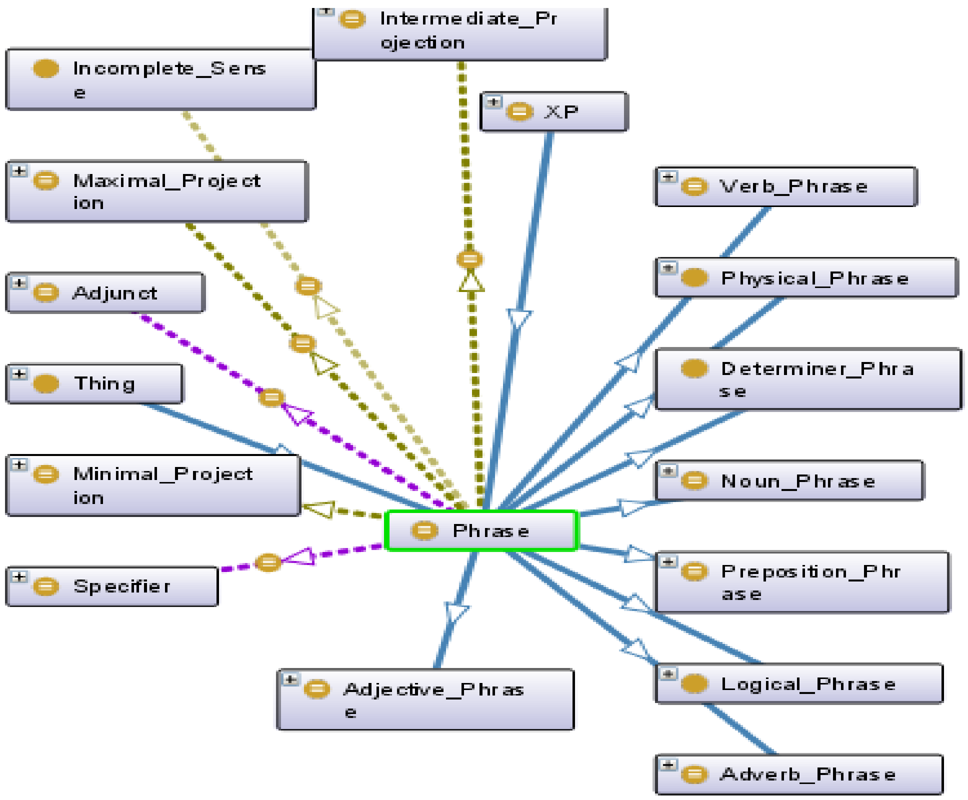
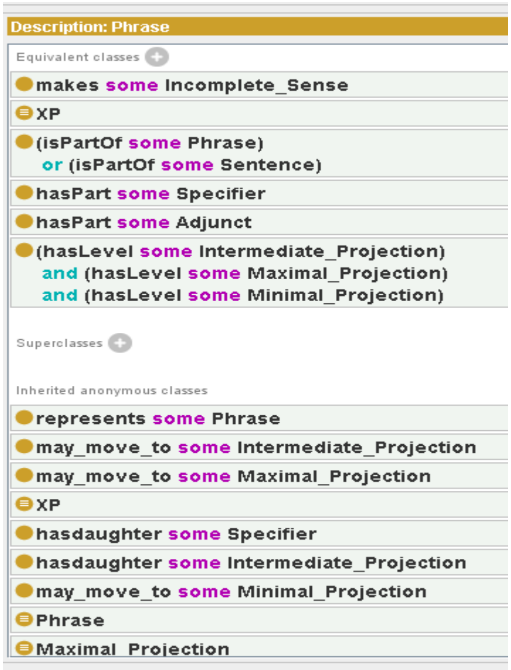


Fig. 16 Phrase: OntoGraf

The other arrows with bubbles represent the Object properties. The movement relation with other classes with the phrase class is represented in the following representation (Textual Representation of a Phrase Fig. 17).



Description: Phrase

Equivalent classes

makes some Incomplete_Sense

XP

(isPartOf some Phrase)
or (isPartOf some Sentence)

hasPart some Specifier

hasPart some Adjunct

(hasLevel some Intermediate_Projection)
and (hasLevel some Maximal_Projection)
and (hasLevel some Minimal_Projection)

Superclasses

Inherited anonymous classes

represents some Phrase

may_move_to some Intermediate_Projection

may_move_to some Maximal_Projection

XP

hasdaughter some Specifier

hasdaughter some Intermediate_Projection

may_move_to some Minimal_Projection

Phrase

Maximal_Projection

Fig. 17 Textual Description of Phrase

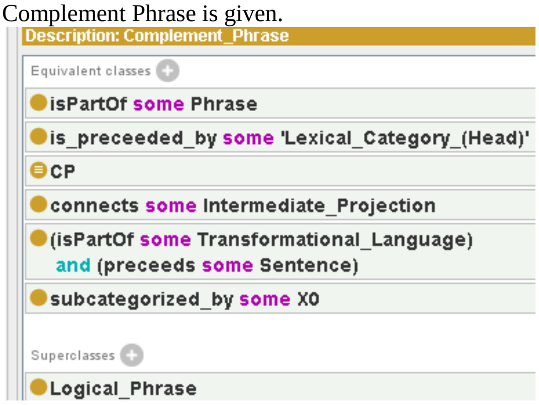
A phrase can further be a part of some phrase or some sentence; it has Specifier and Adjunct as parts. A phrase has three levels, Maximal Projection, Minimal Projection, and Intermediate Projection.

5.8 Logical Phrase Representation

There are two kinds of logical level phrases (in English), one is Complement Phrase (CP) and other is Inflection Phrase (IP).

5.8.1 Complement Phrase

A complement is a phrase that a lexical category takes or selects. In Fig. 18 the textual level snapshot representing Complement Phrase is given.



Description: Complement_Phrase

Equivalent classes

isPartOf some Phrase

is_preceeded_by some 'Lexical_Category_(Head)'

CP

connects some Intermediate_Projection

(isPartOf some Transformational_Language)
and (preceeds some Sentence)

subcategorized_by some X0

Superclasses

Logical_Phrase

Fig. 18 Textual Description of Complement Phrase

OntoGraph representation of the Complement Phrase is shown in Fig. 19

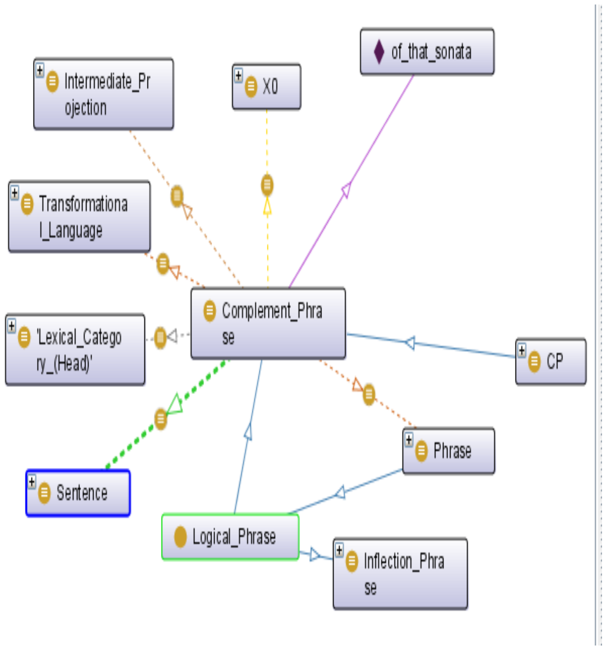


Fig. 19 Complement Phrase: OntoGraf

6. Conclusion and Future Work

In this paper ontological English grammar knowledgebase has been developed. The grammatical knowledge is analyzed using two different views of grammar i.e., conventional grammar view and GB Theory (using its sub theories such as X-Bar theory and Movement theory) view. The various relationships between the grammar constituents have been identified and represented using the Protégé Framework for developing ontologies. Though various grammar constituents and relationship among them have been identified and represented, yet there is a further scope for incorporating more and different type of constituents or entities. For example in future the knowledgebase may be extended by incorporating theta roles knowledge discussed in Theta theory of GB theory.

The present work can be employed for different type of standalone and online services based upon natural language understanding. In particular the proposed work can be used for Natural Language Analysis and Language Learning System.

7. REFERENCES

[1] Marc Moens HCRC Language Technology Group, Edinburgh, “Natural Language Processing”

[2] Cheryl A. Black A step-by-step introduction to the Government and Binding theory of syntax.

[3] GRUBER, T., “A Translation Approach to Portable Ontology Specifications”, *Knowledge Acquisition,* 5(2), 199-220, 1993.

[4] GRUBER, T., “Towards Principles for the Design of Ontologies Used for Knowledge Sharing”, *International Journal of Human and Computer Studies,* 43(5/6), 907-928, 1995.

[5] Dimitris Kanellopoulos, Sotiris Kotsiantis, Panayiotis “ONTOLOGY-BASED LEARNING APPLICATIONS: A DEVELOPMENT METHODOLOGY”. In Proceedings of the 24th IASTED International Multiconference Software Engineering. Feb 14-16, 2006, Innsbruck, Austria

[6] N.F. Noy & D.L. McGuinness, Ontology Development *101: A Guide to Creating Your First Ontology* http://www.ksl.stanford.edu/people/dlm/papers/ontologytutorial-noy-mcguinness.pdf

[7] N. Henze, P. Dolog, & W. Nejdl, Reasoning and Ontologies for Personalized E-Learning in the Semantic Web, *Educational Technology & Society, 7*(4), 2004, 82-97

[8] Lalita Kumari, Radhey Shyam, Swapan Debbarma, Nirmalya Kar, Smita Das, “Government and Binding Theory for Hindi Language” International Journal of Computer Applications (0975-8887) Volume 43-No. 22, April 2012 pp 42-45.

[9] C.R. Rene Robin et. al. / International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5611-5617