

**SEMANTIC ANALYSIS OF TURKISH THROUGH FORMAL ONTOLOGY
CONSTRUCTED IN COMPUTER**

Yelda Fırat¹, Özlem Uçar², Yılmaz Kılıçaslan²

Çanakkale Onsekiz Mart University, Çanakkale Technical Sciences Vocational High School,
Computer Programming, Çanakkale, Turkey¹

Trakya University, Faculty of Engineering and Architecture, Department of Computer Engineering,
Edirne, Turkey²



Abstract

From a general point of view, models of reality that are defined using a formal language are called formal ontologies. These general purpose models can also be used in special domains of application. Since ontologies are very big by nature, they are prone to errors and omissions if constructed by human labor. Also, manually constructed ontologies are not apt to apply to particular application domains. In this paper, an application for automatically constructing a formal ontology, which is immune to the problems just mentioned, is presented. The application aims at semantic analysis of Turkish within a corpus-based framework of formal concept analysis.

Key words: Formal Ontology, Formal Concept Analysis, Word Sense Disambiguation.

1. INTRODUCTION

With its broadest meaning, language is a set of signs which enables people communicate with each other and express their feelings and thoughts directly or indirectly. It is generally claimed that a language is composed of three components: Morphological / syntactic component which covers the relationship among its arguments, semantic component which includes the relationship of the language with the external world, and the pragmatic component which places the expressions in the speaker and listener context.

The semantic analysis of the language in particular requires the use of a model of reality; in other words, the use of an ontology. The computer-assisted linguistics or natural language processing applications make this ontology formal. As the mentioned ontological models are too big, they carry the risk of having far too many errors and omissions if they are formed as a result of human effort. In addition, the adaptations which can emerge from the application requirements of this ontology formed in this way will be difficult.

The aim of this study is to introduce the application of forming a computer-assisted formal ontology. The application has been developed with the corpus-based annotation method within the framework of formal concept analysis, and semantic analysis of Turkish has been aimed as a particular application field.

In the following parts of the article, background information about the study will be presented. Especially, explanations and examples about the thematic and semantic analysis of the clause arguments will be dealt here. Afterwards, how the thematic role analysis can be modelled with concept lattices will be discussed. Also, an application of forming a corpus-based formal ontology achieved on

this hypothetical ground will be a matter of the following chapter. In the conclusion part, the summary of the article and proposals for the future studies will be presented.

2. THEMATIC ROLES

In many linguistics hypotheses, the term thematic role is used to express the meaningful relationship a noun phrase has with the verb of the sentence. Below is the list of the main thematic roles used in the literature (Jackendoff 1972). The underlined noun phrases in the Turkish and English sentences below carry the thematic role described.

THEME – the one in motion or in somewhere

- (1) a. Kedi odada.
b. Cat is in the room.

AGENT – the one initiates an action of the doer of the action

- (2) a. Can camı kırdı.
b. John broke the glass.

PATIENT – the one who is affected from the action

- (3) a. Can camı kırdı.
b. John broke the glass.

SOURCE – the starting point of the action

- (4) a. Can İstanbul'dan Ankara'ya gitti.
b. John went to Ankara from İstanbul.

GOAL – the thing that the action is directed to

- (4) a. Can İstanbul'dan Ankara'ya gitti.
b. John went to Ankara from İstanbul.

LOCATION – the place where something is

- (5) a. Kedi odada.
b. Cat is in the room.

EXPERIENCER – the one who feels, perceives, or realizes something

- (6) a. Kedi topu gördü.
b. Cat saw the ball.

INSTRUMENT – the thing that is used for performing an action

- (7) a. Can ekmeği bıçakla kesti.
b. John cut the bread with a knife.

Many linguistics describe these thematic roles and their morphological / syntactic features as an important part of the grammar theory (Gruber 1965; Fillmore 1976; Dowty 1991; Jackendoff 1993).¹ Especially, the classification of the arguments date back to ancient times, even to the Panini's Karaka theory (Dahiya 1995; Kasper 2008). Also, in the applications of processing the natural languages, it is of great importance to determine the thematic roles of the sentence arguments in analyzing the sentence meaning (Gildea & Jurafsky 2002).

2.1. Theories developed over the thematic roles

Panini is the first person to establish a relationship between the morphological features of the mother tongue (Sanskrit) and the meaningful categories of thought. The rules available in his grammar are in effort to explain the relationship between meaningful statements and their morphological and syntactic codes. According to him, each verb is a member of a verb class. These verb classes have the common features for the purpose of determining the Karaka roles of the sentence arguments. Panini's Karaka roles are indicated in table 1 (Butt 2005).

Karaka role	Panini's definition	Western thematic role
<i>apadana</i>	the fixed point from which something recedes	ablative
<i>sampradana</i>	the item in view through the <i>karman</i>	goal
<i>Karana</i>	the most effective means	instrument
<i>adhikarana</i>	locus, location	locative
<i>Karman</i>	the thing desired by the <i>kardr</i>	patient
<i>Kardr</i>	the independent one	agent

Table1. Panini's Karaka roles for Sanskrit

Though it isn't being discussed in a very detailed way, Karaka theory is handled as a pioneer theory for the modern theories of thematic roles. In a piece of writing which was published by Frank R. Blake in 1930 and is referred as one of the most important writings about thematic roles, Blake mentions a morphological relationship between the verb of the sentence and the sentence arguments. According to him, the morphological features for the meaningful relationships are named *case form*. In Blake's grammar, morphological features refer to nouns, pronouns, adverbs and verbs, while thematic roles according to morphological features refer to AGENT, PATIENT, INSTRUMENT, THEME, STIMULUS, GOAL, RECEIVER and BENEFACTIVE (Blake 1930). The generalizations obtained from Blake's were made according to morphological features rather than syntactic features of the sentence arguments.

In 1950s, Noam Chomsky described linguistics as a part of the cognitive psychology (Chomsky 1970). The aims and the objectives of the studies carried out until now are in close relationship with his studies. The most striking feature of the Chomsky's theory is that it deals with the language's syntactic component and semantic component separately.

¹ The reader can find a more comprehensive discussion over the literature related to the thematic roles at Kasper (2008).

The association of the syntactic structure formed with morphological categories or word phrase types together with the semantic indications of the sentences bring about the linking problem for us.

- (8) a. Can Meryem'e vurdu.
b. Can hit Meryem

As seen in the sample 8, the verb 'hit / vurmak' and the order of its arguments are the same as the arguments of the verbs 'kill / öldürmek', 'read / okumak', 'send / göndermek'. As can be understood from this, the doers of the action, 'killer / öldüren', 'reader / okuyan', 'sender / gönderen', are always in the position of subject in the structural position. In this context, each linguist should mention the relationship between syntax and meaning to put these generalizations forward. As mentioned above, if there is a doer of the action in a sentence, it is certainly placed in the position of subject. These gain an independent status in the linking theory and become valid. They are named AGENT and of all the arguments of the sentence they become the doer of the action. As a result, the use of the thematic roles becomes inevitable due to the linking problem. Therefore, these roles show the path to the real meanings by indicating the relationships between rule and meaning of the verb and its arguments. In this context, the question What kind of information is needed for location of the thematic roles stated in the special theories? arises. In other words, What kind of information can find the way from the meaning to the rule? The first possibility for the answer to these questions is that the meaning can be equalized to the universal information of the world. The second possibility, on the other hand, is the knowledge that comes from the relationship of the syntax (grammatically related) and the meaning which shares the same structure (the same grammar categories) Chomsky assumes. According to Chomsky's theory, these meaning pieces which are grammatically related makes a part of the rule-meaning interface. That is, what is predicted from the rules are the semantic indications. The first possibility mentioned above has been studied by Jackendoff as a part of the conceptual meaning with the name of conceptual structure since 1983 (Jackendoff 1987). According to him, grammatical meaning is coordinated with the perceived world view. In other words, all knowledge emerging from the human perception is linguistically related. As Chomsky said, besides the innate knowledge, there is knowledge perceived through sense organs.

One of the approaches that must be taken into consideration is Dowty's lexical decomposition approach, which he used to determine the thematic roles. Dowty explicitly stated the necessity of these thematic roles in 1980s and put forward the positions of thematic roles clearly. According to him, there is a need only for two types of roles for describing the argument selection effectively. These are Proto-AGENT and Proto-PATIENT (Dowty 1991). Here, the real meanings of the words should be dealt with as cognitive reality by thematic roles. In this context, Croft's roles can be given as the first examples. These roles are based on the cognitive model and they are derived from the causative organization of the events (Croft 1991). Croft and Jackendoff who made a similar role description in 1972 handled the thematic role concept as a part of the semantic indication. In the studies Jackendoff carried out in the following years, the thematic roles were considered as conceptual system.

Since 1970s, the views on the thematic roles have changes owing to the changes in the sentence and its arguments. The argument structures which develop with the syntactic categories which are used to separate the sentence arguments (noun phrase, verb phrase, adjective phrase) also refer to a location which involves the thematic roles in grammar. In short, argument structures are handled as the lexical semantics and syntax interface by most of the linguists. If required to summarize in figure 1;

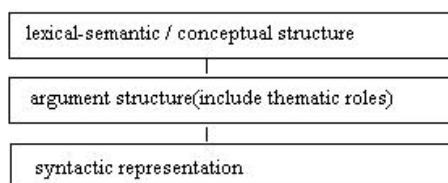


Figure 1. The role of argument structure in theories of grammar

It is seen that the roles have not been the same during the history of the thematic roles, thus these roles are named multi-dimension roles. For instance, Croft described two types of roles. These are direct and indirect roles (Croft 1991). These roles play an highly important role in cognitive importance and causative categorization of the events. Direct roles are normally composed of subject, object, and indirect object. Jackendoff's conceptual structure thought is composed of the thematic tier including the roles about the action and the location and the action tier including the roles about the relationship of AGENT-PATIENT (Jackendoff 1987). The words find their real meanings with the addition of the cognitive tier and the causation tier to these tiers. The thematic roles which are defined in these tiers brings a solution to the formation of the roles at the same position. For example, while the role of AGENT thematic role becomes the role of SOURCE, the role of EXPERIENCER can be analyzed as the role of GOAL.

2.2. *The subjects studied over the thematic role*

With the given sample sentences taken from Turkish and English languages, it has been aimed to indicate the phenomenon of language. These sample sentences should be taken into consideration as an effort for the linking theories according to the argument selection, distribution and thematic roles.

- (9) a. Can ölür.
b. *Can isteyerek ölür.
- (10) a. Can dans ediyor.
b. Can isteyerek dans ediyor.
- (11) a. John dies.
b. *John dies intentionally.
- (12) a. John is dancing.
b. John dances intentionally.

In the sample sentences (9), (11) and (10), (12), it has been seen that there are systematical differences in the intransitive sample sentences. Whereas all the sentences are categorized as intransitive in sample sentences, the sample sentences (10) and (12) are considered agentive due to the word 'isteyerek / intentionally'. It was Fillmore who gave the first systematical approach about intransitivity (Fillmore 1968). Fillmore analyzed the intransitive verbs in the sample sentences (10) and (12) as unergative verbs which have active subjects which can trigger an action. On the contrary, the verbs (unaccusative verbs) in the sample sentences (9) and (11) have an element which is affected by an inactive subject. This type of verbs are called unaccusative verbs the subject of which is in the verb phrase.

Undoubtly, as can be predicted from the previous studies, as the AGENT role is seen as a central factor, it is the most commonly discussed thematic role. Even Gruber, who is the first person to find the thematic role AGENT in 1965, could not make only one definition for this role. The followings are definitions made by Gruber for the thematic role AGENT (Gruber 1965):

- The one who allows the realization of an action intentionally
- The causer who allows the change of a situation intentionally

Jackendoff, a follower of Gruber, defined the role of AGENT as the one who causes an effect (Jackendoff 1972). In the following years, Jackendoff added the features of will-power, mobility, and initiator for the role of AGENT.

According to van Valin, AGENT was defined for the intended actions. van Valin mentions three factors administering applicability: the semantic features of the noun phrase depending on the action, the structural content of the noun phrase, and the grammatical structure formed by the verb and the noun phrase (van Valin 1996). According to him, the AGENT should have the following features: speaker, discourse participant, human, rational, volitional, animate, and concrete entity (van Valin 1996).

If looked back on the intransitive samples, it can be seen that the applicability is a factor which determines whether the verb is an unergative verb or an unaccusative verb.

The most significant function of the thematic role theory is to show how the role appointments of two sentences of different syntax can be performed more accurately. If the sentences in Turkish and English below are considered in terms of double subject constructions;

- (13) a. Can Meryem-e papağan-ı gönderdi.
Can Meryem-DATIVE parrot-ACCUSATIVE send-PAST
b. Can sent Meryem the parrot.
- (14) a. Can papağan-ı Meryem-e gönderdi.
Can parrot-ACCUSATIVE Meryem-DATIVE send-PAST
b. Can sent the parrot to Meryem.

It is seen that these samples express the same interactions but they are different in perspectives. That is, somebody sends the parrot to somebody else. This difference results from the order of the arguments. Similarly, there are differences in the appointment of the thematic roles, for there are differences in the perspectives of these sentences. These differences result from the use of prepositions ('from' and 'to'). No such difference is witnessed in Turkish language. That is, In the sentence (13), 'Can' is SOURCE and AGENT, 'the parrot' has the role of THEME, and 'Meryem' is in the role of PATIENT and GOAL. However, in the sentence (14b), 'Mary' has only the role of GOAL because of the preposition 'to'. Therefore, a great number of semantic differences occur in these sentence versions. In addition, it is clear that the arguments of the verb have more than one thematic role and word sense disambiguations take place.

If the arguments of the verbs are required to be annotated with a single role, the modelling of such situations can be indicated through Galois Lattices. Hence, the words can be free of ambiguity thanks to the structural lattice in which the thematic roles are available. In this context, the thematic roles are grouped under two tiers. These are:

- Core or action tier: This is the tier in which there are the roles which indicate the distinctive features among the words and should be signed for each noun phrase. (It indicates the relationship between AGENT and PATIENT).
- Non-core or thematic tier: It is the tier which is only ontologically necessary but in which there are roles which are not special like central roles. (It involves the roles about action and location).

Depending on this, because of the insufficiency to find the word meanings of the world model, the conceptual, in other words, real meanings of the arguments of a verb and structural ontologies where the thematic roles are shown are reached, and the structural ontologies are expressed with the concept lattices in which formal concept analysis is applied.

3. LATTICE-BASED ONTOLOGY OF THE WORDS

Semantic roles make avoiding the known difficulties of the abstract thematic roles which express a universal or a smaller role cluster possible. As mentioned above, the semantic roles of the sentence arguments can be easily reached especially with the thematic roles of AGENT and PATIENT. The ambiguities occurring in the arguments of the verb can be easily solved through the indication of the thematic roles in the concept lattices.

3.1. *The modelling of the thematic roles with concept lattices*

Using concept lattices (Galois Lattices) obtained through the formal concept analysis, a mathematical role hierarchy can be formed (Kılıçaslan & Güner 2011). Formal Concept Analysis was first developed by Ganter and Wille (1996). The data given in the formal context is composed of two clusters and the expression which shows the relationship of these two clusters (Kılıçaslan & Güner 2011). While one of the clusters involves the features, the other cluster involves the objects having these features. The concept lattices obtained in this way both allow the modelling of thematic roles in a hierarchical organization and allow a noun phrase to carry more than one thematic role. If it is required to explain in a sample Turkish sentence:

(15) Öğretmen öğrenciye bir kitap verdi.

In the sentence above, 'öğretmen / the teacher' has the role of AGENT and SOURCE, 'bir kitap / a book' has the role of THEME, and the word 'öğrenciye / to the student' has the role of GOAL and PATIENT. That is, in the given sentence, the words 'öğretmen / the teacher' and 'öğrenciye / to the student' have two thematic roles. As shown in figure 2, the thematic role cluster of the each argument is represented with a knot.

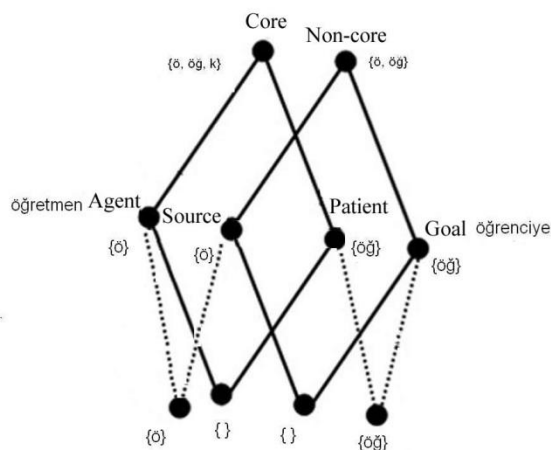


Figure 2. The indication of the thematic roles in lattices

If it is necessary to express the letters forming the object cluster explicitly, *ö* stands for *öğretmen*, *ög* stands for *öğrenci*, and *k* stands for *kitap*. As can be seen from figure 2, the thematic roles which have been expressed in a linear way up to now have been mathematized by being made three dimensional thanks to the lattices.

4. APPLICATION

The aim of this chapter is to introduce the application of forming a computer-assisted formal ontology. The application has been developed with the method of corpus-based annotation within the framework of the formal concept analysis. The corpus used is composed of 1033 sentences. The semantic analysis of Turkish language has been aimed as the private application field.

Annotation is the exemplification of the semantic, morphological, and syntactic features of each argument and part by dividing a sentence into its parts and arguments. This kind of annotation has been achieved through the use of an annotation interface prepared in the computer environment, PHP language, and a database formed in MySQL. This interface is in the location of a page working in the web environment.²

With the help of the interface shown in figure 3, Turkish sentences belonging to our corpus were divided into parts, their thematic roles were annotated, and their syntactic and morphological analyses were performed. The database formed through the data obtained from these annotations involves the definitions of the sentence arguments and provides much more information than traditional dictionaries (Fillmore, Johnson & Petrucci 2003). In other words, this database contains the characteristics that define the morphological, syntactic, and semantic indications of each argument beneath the words' meanings.

² The interface can be accessed via <http://www.yelda1367.com/member/login> address.

155 | Published by Info Invest, Bulgaria, www.sciencebg.net

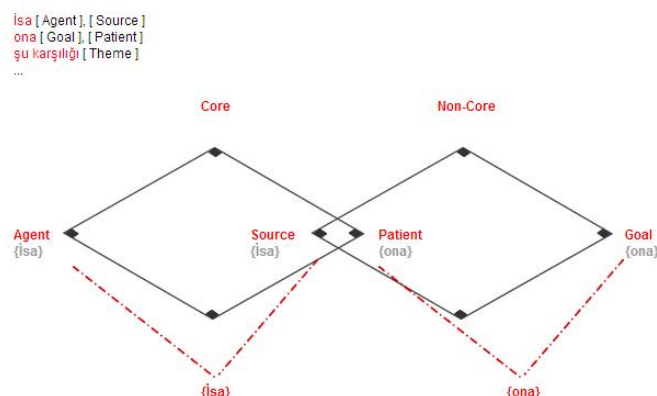


Figure 5. Automatically formed lattice diagram of the annotated sentence

5. CONCLUSIONS

If this study is handled as a pre-study for mapping the ontologies, it gains importance in terms of its contribution to interlingual translation and integration. It gives the researchers studying on Turkish further access to the database. Since Turkish language is agglutinative and has a word order of subject-object-verb, machine learning eliminates difficulties in its translation, and makes the ambiguous words meaningful.

This study is a annotation for each sentence element. In this technique, the arguments of the Turkish sentences are defined with the developed ontology-based concept lattices. In the future studies, by taking the syntactic and morphological features of the sentence arguments into consideration and reaching the words' semantic roles from their thematic roles, Turkish word groups will be formed from the possible distribution of the words, which is calculated through formulas, with the help of these lattices formed.

In this sense, a bilingual ontology mapping will be provided with the stages of this and future studies. An extensive Turkish dictionary will be prepared in an electronic environment and by finding the correct Turkish equivalents of the English words, machine translation will be achieved.

REFERENCES

- Butt, M 2005, *Theories of Case*, Cambridge University Press, Cambridge.
- Blake, F R 1930, 'A Semantic Analysis of Case', *Curme Volume of Linguistic Studies*, vol. 6, no. 4, Monograph no. 7, pp. 34-49.
- Chomsky, N 1970, *Sprache und Geist*, Suhrkamp, Frankfurt.
- Croft, W 1991, *Syntactic Categories and Grammatical Relations*, University of Chicago Press, Chicago / London.
- Dahiya, Y V 1995, *Panini as a linguist: Ideas and Patterns*, Eastern Book Linkers, Delhi, India.
- Dowty, D 1991, 'Thematic Proto-Roles and Argument Selection', *Language*, vol. 67, no. 3, pp. 547-619.

- Fillmore, C 1968, 'The Case for Case', *Universals in Linguistic Theory*, in E Bach and R Harms (ed.), New York.
- Fillmore, C J 1976, 'Frame semantics and the nature of language', *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, vol. 280, pp. 20-32.
- Fillmore, C J, Johnson, C R & Petruck, M R L 2003, 'Background to Framenet', *International Journal of Lexicography*, vol. 16, no. 3, Special Issue: 1, pp. 235-250, Oxford University Press.
- Ganter, B & Wille, R 1996, *Formale Begriffsanalyse: Mathematische Grundlagen*, Springer, Berlin.
- Gildea, D & Jurafsky, D 2002. 'Automatic Labeling of Semantic Roles', *Computational Linguistics*, vol. 28, no.3, pp. 245-288.
- Gruber, J 1965, *Studies in Lexical Relations*, MIT Dissertation, Indiana University Linguistics Club, Bloomington.
- Jackendoff, R 1972, *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge.
- Jackendoff, R 1987, 'The Status of Thematic Relations in Linguistic Theory', *Linguistic Inquiry*, vol. 18, no. 3, pp. 369-411.
- Jackendoff, R 1993, 'On the Role of Conceptual Structure in Argument Selection: A Reply to Emonds', *Natural Language and Linguistic Theory*, vol. 11, no. 2, pp. 279-312.
- Kasper, S 2008, 'A comparison of 'thematic role' theories', Unpublished master thesis, The Philips University, Marburg.
- Kılıçaslan, Y & Güner, ES 2011, 'Filtering Machine Translation Results with Automatically Constructed Concept Lattices', *Proceedings of 8th International Conference on Concept Lattices and Their Applications (CLA 2011)*, Nancy, France, pp. 59-73.
- van Valin, R D jr 1996, 'The Case for "Effector". Case Roles, Agents and Agency Revisited', in M Shibatani and S. Thompson (ed.), *Grammatical Constructions*, Oxford University Press, Oxford, pp. 289-322.