SEMANTIC ARGUMENT CLASSIFICATION AND SEMANTIC CATEGORIZATION OF TURKISH EXISTENTIAL SENTENCES USING SUPPORT VECTOR LEARNING

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Aylin Koca
September, 2004

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Prof. Dr. Varol Akman (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Prof. Dr. Özgür Ulusoy

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Prof. Dr. Enis Çetin

Approved for the Institute of Engineering and Science:

_____

Prof. Dr. Mehmet Baray

Director of the Institute

ABSTRACT

SEMANTIC ARGUMENT CLASSIFICATION AND SEMANTIC CATEGORIZATION OF TURKISH EXISTENTIAL SENTENCES USING SUPPORT VECTOR LEARNING

Aylin Koca
M.S. in Computer Engineering
Supervisor: Prof. Dr. Varol Akman
September, 2004

There are three types of sentences that form all existing natural languages: verbal sentences (e.g. “I read the book.”), copulative sentences (e.g. “The book is on the table.”), and existential sentences (e.g. “There is a book on the table.”). Syntactic and semantic recognition of these sentence types are crucially important in computational linguistics although there has not been any significant work towards this end. This thesis, in an attempt to fill this evident gap, is on identifying and assigning semantic categories of Turkish existential sentences in print. Existential sentences in Turkish are minimally characterized by the two existential particles *var*, meaning *there is/are*, and *yok*, meaning *there is/are no*. In addition to these most basic meanings, other senses of existential particles are possible, which can be categorized into groups such as case existentials and possession existentials. Our system does shallow semantic parsing in defining the predicate-argument relationships in an existential sentence on a word-by-word basis, via utilizing Support Vector Machines, after which it proceeds with the semantic categorization of the whole sentence. For both of these tasks, our system produces promising results, in terms of accuracy and precision/recall, respectively. Part of this research contributes to the annotation of the METU-Sabancı Turkish Treebank with semantic information.

Keywords: shallow semantic parsing, semantic role labeling, thematic roles, support vector machines, Turkish existential sentences, Turkish Treebank.

ÖZET

TÜRKÇE VAROLUŞSAL CÜMLELERİN DESTEK VEKTÖR MAKİNELERİ KULLANILARAK ANLAMBİLİMSEL ARGÜMAN SINIFLANDIRILMASI VE ANLAMBİLİMSEL GRUPLANMASI

Aylin Koca
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Prof. Dr. Varol Akman
Eylül 2004

Bütün doğal diller üç çeşit cümleden oluşur: fiil cümleleri (ör. "Ben kitabı okudum."), isim cümleleri (ör. "Kitap masanın üzerinde."), ve varoluşsal cümleler (ör. "Masanın üzerinde kitap var."). Bu cümle çeşitlerinin sözdizimsel ve anlambilimsel tanımaları bilişimsel dilbilim için çok önemli olduğu halde, bununla ilgili yapılmış belli başlı bir çalışma bulunmamaktadır. Bu tez, varolan sözkonusu açığı kısmen de olsa kapatmak amacıyla, Türkçe varoluşsal cümlelerin anlambilimsel tanınması ve sınıflaması üzerinedir. Türkçe varoluşsal cümleler, asgari olarak *var* ve *yok* işlevsel sözcükleriyle ıralanır. Bu işlevsel sözcüklerin, varlık bildiren en temel anlamlarının dışında, başka anlamları da mevcuttur. Bunları, örneğin, sahiplik veya hâl/durum bildirenler olarak sınıflamak mümkündür. Sistemimiz öncelikle, destek vektör makineleri yardımıyla, varoluşsal cümlelerin yüklem ve diğer öğeleri arasındaki ilişkileri tanımlamak için kelimeleri baz alan sığ anlambilimsel ayrıştırmasını yapmaktadır. Bunu takiben de cümlelerin anlambilimsel gruplamasını gerçekleştirmektedir. İlk işlem için aldığımız doğruluk, ve ikinci işlem için aldığımız duyarlık/geri çağırma sonuçları oldukça umut vericidir. Bu çalışmamızın bir katkısı da ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemi'nin bir kısmının anlambilimsel bilgi ile etiketlendirilmesi olmuştur.

Anahtar sözcükler: sığ anlambilimsel ayrıştırma, anlambilimsel rol etiketlendirilmesi, anlambilimsel roller, destek vektör makineleri, Türkçe varoluşsal cümleler, Türkçe ağaç yapılı derlem.

Acknowledgements

"Gratitude is the heart's memory."

- French Proverb

First and foremost I would like to express my endless gratitude to my research advisor Prof. Varol Akman for always genuinely believing in me right from the beginning. It was his persistent encouragement and valuable guidance that allowed me to stay resolute and together throughout my graduate studies. One cannot but revere his most pleasant personality, and undoubtedly his qualities both as a researcher and an instructor. I hereby would like to thank him also for his kind generosity he never ceased to show his students in all aspects, by which I am deeply impressed.

Only after I started to build interest in the field of computational *linguistics*, did I come to realize that I was so fortunate enough to actually work with a professional in linguistics: Assoc. Prof. Dr. Engin Sezer, then having only recently arrived from Harvard University, showed keen interest to collaborate, albeit his already overloaded schedule. I am indebted to him for that, and for many long sessions of stimulating discussions that followed. The subject matter of this thesis has been decided on in such one discussion, inspired by a recent work of him. Having the opportunity to share his many bright ideas, closely observe his ways of approach to research was invaluable experience and certainly a privilege for me. For his boundless support in everything, including work for this thesis right from the beginning until the very end, I am truly grateful.

I would like to extend my indebtedness to the members of my oral and reading committee: Prof. Varol Akman, Prof. Özgür Ulusoy, and Prof. Enis Çetin. Assoc. Prof. Engin Sezer also reviewed a draft of this thesis and provided valuable feedback for which I am also thankful. For any possible remaining errors or shortcomings, I am solely to be held responsible.

I should like to acknowledge the generous and punctual assistance of Gökhan Tür at an initial phase of this thesis, when I was not quite sure which approach to take in addressing my problem. Asst. Prof. İlyas Çiçekli also granted his kind assistance by commenting on an early draft of part of the work, and giving advice on an experimental design issue.

The cooperation of the developers of METU-Sabancı Turkish Treebank is greatly appreciated. This work would not be possible without the Treebank. Also my thanks go to Cem Bozşahin of METU, for arranging a meeting exclusively on the Treebank to discuss future directions that should be taken in its further development.

The discussions I had with my colleagues at various stages of this work helped me develop the ideas put forward in this thesis. For that, I would specifically like to thank Rabia Nuray. Throughout this study, she was always a soothing factor whenever I felt unsure or puzzled. Thanks also go to Eray Özkural, Ata Türk, Berkant Barla Cambazoğlu, and Ayışığı Sevdik. The assistance of Sinan Uşşaklı in parsing the Treebank documents with C# of MS Visual Studio .NET is also very much appreciated.

Last but not least, my deep gratitude is to my dear grandfather Major General Selâhattin Kavuştu, Master of Science in Aerospace Engineering, who has been my most significant role model. It was him who first lighted the flame within me and I cannot thank him enough for that. This thesis is reverently dedicated to him.

Contents

List of Figures

List of Tables

List of Abbreviations

1SG, 2SG, 3SG	first, second, third person singular
1PL, 2PL, 3PL	first, second, third person plural
P1SG, P2SG, P3SG	first, second, third person singular possessive
P1PL, P2PL, P3PL	first, second, third person plural possessive
ABL	ablative (+*dAn*)
ADVB	adverbial conversion
APAST	auxiliary past suffix (+*ydH*) different from the verbal past
COND	conditional
COPULA	copula (+*dHr*)
DAT	dative (+*yA*)
E	positive existential particle *(var)*
FUTPART	future participle (+*yAcAk*)
GEN	genitive (+*nHn*)
INS	instrumental (i.e. comitative) (+*lA*)
LOC	locative (+*dA*)
NE	negative existential particle *(yok)*
PASTPART	past participle (+*dHGH*)
Q	yes/no question particle (*mH*)

Chapter 1

Introduction

1.1 Background and Motivation

No later than the beginning of the new millennium, natural language understanding reached a state where semantics plays a greater role than it once did. The need for moving away from carefully hand-crafted, domain-dependent systems[1] towards robustness and domain-independence turned out to be an essential concern. Therefore, the recent advances in domain-independent shallow semantic parsing have been receiving significant attention of the natural language processing community. This is the process of producing a markup for sentences in texts via assigning a simple WHO did WHAT to WHOM, WHEN, WHERE, HOW, etc. structure to them. Although the notion of shallow semantic parsing (i.e. case role analysis) has a long history in computational linguistics literature [JUR2000], the automatic, accurate and wide-coverage techniques that can efficiently annotate text with semantic argument structure have not been quite promising until recently. The case has been even less promising for languages and genre

[1] These are simple speech- and text- based natural language understanding systems that answer questions about flight arrival times (e.g. ATIS in [HEM1990]), give directions, report on bank balances, and the like.

of text for which statistical syntactic parsers are not readily available[2] (e.g. Turkish). Various researchers have cast this problem as a tagging task and have applied supervised machine learning techniques to it [GIL2002a; BLA2000; GIL2002b; SUR2003; GIL2003; CHE2003; FLE2003; HAC2003b; THO2003; PRA2003]. Comparisons of some of these systems are presented in [PRA2003].

When one of several *IOB* representations is utilized [SAN1999], it is straightforward to view shallow semantic parsing as a tagging task. According to these representations, each word in a sentence is labeled with a tag: *I* means that the word is inside a semantic role, *O* means that the word is outside a semantic role, and *B* means that the word is the beginning of a semantic role. Tagging, furthermore, can be formulated as a multi-class classification problem, where the number of classes depends on the number of semantic roles (where each role is filled with one or more words).

In textual classification problems, support vector machines have been shown to be well suited for learning since they are capable of handling a large number of features with strong generalization properties. They also outperform the conventional statistical learning algorithms such as Decision Tree and Maximum Entropy models as has been stated in [JOA1998; KUD2000]. Therefore, we can have support vector machines assign semantic roles to the words of a sentence.

We are interested in semantic roles that allow us to capture, represent, and understand the predicate-argument relations of Turkish existential sentences, at an abstract level. In view of that, we developed a set of domain-independent abstract semantic roles, such as THEME, LOCATION, SOURCE, POSSESSOR. Similar sets of roles have been used in [HAC2003b], as well as in FrameNet [BAK1998] and PropBank [KIN2002] corpora. The words that represent our set of semantic roles within a sentence are each further tagged in accordance to the *IOB* representation. The resulting tagged existential sentences, allow for the development of a system that categorizes each

[2] A fundamental assumption in architectures adopting various supervised machine learning techniques is the presence of a full syntactic parser that provides the bulk of the features used in the training stage.

sentence according to the existential group that it belongs. This categorization is done based on the types of semantic roles that are tagged to the words of each sentence.

The automated semantic categorization of Turkish existential sentences first requires a theoretical, in-depth syntactic and semantic analysis of the Turkish existential sentence construct. Various comprehensive grammars of Turkish [UND1976; LEW1967; KON1956] discuss existential sentences. The most relevant and recent work on the specific matter of Turkish existential sentences has been conducted by Sezer [SEZ2003]. In his work, Sezer exclusively concentrates on the interaction of various semantic and syntactic properties of Turkish existentials.

It is important to realize that semantic representations play a central role in natural language interfaces between humans and computers. In simple information retrieval tasks, they are used to understand the user's input. In more complex tasks such as question answering, the semantic representation is used to understand the question, to expand the query, to find relevant documents that match the question, and to present a summary of multiple documents as the answer.

This study covers issues of various strands of linguistics and computer science such as natural language processing, and machine learning. Its results can play a major role in tasks like information extraction, question answering, and summarization. It can also serve as an intermediate step in machine translation. Furthermore, the work can always be extended to cover phonology and speech processing, if we decide to base this system on speech rather than text.

1.2 Overview of the Thesis

This thesis is on developing consistent semantic[3] argument identification and classification of Turkish existential sentences, and then accurately categorizing these existential sentences according to their semantic groups. The system largely makes use

of the syntactic information encoded within the METU-Sabancı Turkish Treebank[4]. The exploitation of Support Vector Machines (SVMs), in tagging the arguments of a sentence with semantic roles, proves useful. This is due to the fact that SVMs are easy to use and are capable of performing good classification on textual data, hence our promising results. On the task of assigning semantic roles to the arguments of the predicate of an existential sentence, the system achieves 71.93% accuracy (via SVMs), and on the task of categorizing existential sentences the accuracy reaches 83.33%. It is possible to improve these results almost by 5% by incorporating semantic information to the input files for the SVM.

The organization of the thesis is as follows: The elaboration on Turkish existential sentences and their semantic categorization is provided in Chapter 2. This includes both the categories that are covered by the system, and those categories that are overlooked. Chapter 3 describes the corpus used and the abstract thematic role schema developed for the semantic annotation of the corpus. Chapter 4 elucidates the methodology of this research in two main steps: shallow semantic parsing via classification, and the process of existential sentence categorization. In doing this, it also provides brief overviews of shallow semantic parsing task in general, and SVMs as multi-class classifiers. Chapter 5 then realizes the methodology described in Chapter 4, and reports the results of the two sets of experiments conducted: First, the set of experiments where no semantic information is used to predict the semantic role of a word; and second, the set of experiments where semantic class labels of previous words within the same sentence and inside a predefined context are used. Finally, Chapter 6 summarizes the thesis, draws conclusions, and discusses future directions.

---

[3] The use of "semantic" here, and throughout this thesis, designates the semantics that is incorporated into some syntactic structure, hence not pure semantics.

[4] www.ii.metu.edu.tr/~corpus/treebank.html

Chapter 2

On Turkish Existential Sentences

The two existential particles *var* and *yok* in Turkish show much resemblance to verbs in having their own argument structure and assigning specific thematic roles. Sezer [SEZ2003] argues that there are two sets of existential particles in Turkish that should hence be recognized as two different lexical entries. Of these, one set has the meaning *present/absent* or *is/is not part of*, which assigns the semantic role <participant> on their subject and <scene> on their locative NP[5]. The other set simply asserts the existence of some object in some location, assigning the thematic role <entity> on its subject, and <location> on its locative NP. The latter set contributes to what is generally referred to as the existential sentence in many languages [SEZ2003]. Apart from this somewhat subtle semantic distinction among the existentials in Turkish, it is still possible to do classification into semantic categories, based on the syntactic properties of words comprising the existential sentences. Sections 2.1, 2.2, and 2.3[6] describe the semantic categories of existential sentences that our system processes, while the categories mentioned in Section 2.4 neither exist nor are handled in our system. Subsequently, Section 2.5 gives listings of the remaining uses of the two particles, which are also

[5] <scene>, roughly means a place where there are already other objects (e.g. a picture that contains other objects, a list, an event, a file).
[6] The example sentences used in Sections 2.1, 2.2, and 2.3 are taken from the Turkish Treebank.

neglected by our system because their meaning pass beyond that of regular existentials (e.g. compound verbs, idioms, etc.), but should still be recognized as being so.

2.1 Bare Existentials

Bare existentials constitute the simplest category of existentials. There is no significant information other than the overt subject. This category would correspond to the second set of existentials in Sezer's work [SEZ2003], and is generally referred to as the existential sentence in many languages. Note that here however, there is no explicit information regarding location. The speaker inherently assumes that the hearer knows about the context –hence the location that is implicitly being referred to– when s/he utters a bare existential sentence. Such deep analysis of the semantics of sentences is beyond the scope of our research. Some examples of this category are as follows:

- *İçki var mı?*
  drink E Q
  "Is there (a) drink?"
- *Korucu yok-tu.*
  guard NE-APAST
  "The guard was not present."

2.2 Case Existentials

Case existentials comprise those sentences in which there is case information in the noun phrase (i.e. NP), such as locative, ablative, dative, and instrumental. This case information directly contributes to the existential sense in such a way that it makes explicit WHERE, FROM WHOM/WHAT/WHERE, TO WHOM/WHAT/WHERE, or IN RELATION WITH WHOM/WHAT/WHERE the overt/covert subject exists/does not exist, respectively. Surely, this is only an oversimplification of the information acquired from the case in the NP. For instance, the locative suffix is used to express not only

location in space but in time as well[7]. However, this fine-grained semantic distinction does not yield different categories in our scheme.

- *Arka bahçe-de kimse yok-tu.*
  back yard-LOC nobody NE-APAST
  "There was no one in the back yard."
- *Tamamlanmayan-a para yok.*
  the_one_that_has_not_been_completed-DAT money NE
  "There is no money for the one that has not been completed."
- *Ben-im kimse-yle yarış-ım yok.*
  I-GEN nobody-INS rivalry -P1SG NE
  "I am not in rivalry with anyone."

Note that the relation between the case in the NP and the semantic category of the sentence that bears it is not bidirectional. In other words, case existential sentences always bear a case in the NP, whereas not all NPs with case markers designate a case existential sentence.

One complex instance of case existentials is the *compound case* existentials. Sentences of this type bear NPs with various case markers. The invented example below demonstrates such a case, where the ablative and the dative case markers coexist:

- *Bugün İstanbul'dan Ankara'ya otobüs yok.*
  today Istanbul-ABL Ankara-DAT bus NE
  "Today there are no buses from Istanbul to Ankara."

[7] As in the case:
O zamanlar-da bilgisayar yok-tu.
that times-LOC computer NE-APAST
"There were no computers at those times."

2.3 Possession Existentials

Existential possession is used in Turkish due to the lack of a verb meaning *to have*[8]. From sentences that belong to this category, it is possible to obtain information regarding the possessor object/person, the possessed object/person, or both.

- *Çocuklar-ı yok.*
  children-P3SG NE
  "He does not have kids."
- *Duş-unuz da var.*
  shower-P2PL also E
  "You also have a shower."

It may well be the case where in a sentence there is both possessor/possessed information and case information. Then the category to which such a sentence belongs is determined by the emphasized component: This typically is the component of the sentence that appears right before the predicate, but may change according to prosody. However, in order to be consistent with category marking of such sentences, a thematic role hierarchy that provides a guideline is devised. This is further detailed in Section 3.2.2 of this thesis.

2.4 Other Categories of Existential Sentences

The semantic categories of existential sentences presented in sections 2.1, 2.2, and 2.3 are the ones that are handled in our system, although there are yet other semantic categories of existentials. One incentive for ignoring the remaining categories for this work is their marked semantic peculiarities. Also, it should be noted that the semantic categories handled in the system differ from each other via the implicit means of the syntax of the lexicon used, whereas the sole use of syntax is not sufficient to

[8] Other Turkic languages also lack an indigenous verb meaning *to have*.

differentiate among the overlooked categories. The description of each overlooked category is the subject matter of this section.

Definite Subjects

Existentials with initial **definite subjects** constitute a representative example to the first set of existentials in Sezer's work [SEZ2003]: those that assign the semantic role <participant> on their subject and <scene> on their locative NP. This is different than the locative case existentials, in which an <entity> is simply acknowledged to exist in a physical <location>. Note that an initially placed <participant>, inherently assumes a more influential role than <entity> within the context of the sentence[9], although contrary readings also exist due to speech prosody. Some selected examples from [SEZ2003][10] illustrate definite subjects in existentials:

- *Ben bu komite-de var-ım.*
  I this committee-LOC E-1SG
  "I am on this committee."
- *Siz o toplantı-da yok mu-ydu-nuz?*
  you that meeting-LOC NE Q-APAST-2PL
  "Were you not at that meeting?"

Picture Existentials

The **picture existential sentences** demonstrate similar semantic properties to the above category, in that they also feature initial definite subjects. However, rather than implying that a particular object is physically existing in some context (e.g. at some physical place, at a meeting, at dinner, etc.) as a participant in a scene, they indicate that an object

[9] This subjective evaluation is done based on the particular information the sentence aims to convey. When the two sets of existentials in Sezer's work is considered, the first set is likely to put emphasis on the <participant>, whereas the second set on the <location>.
[10] See [SEZ2003] for a more detailed discussion on "Definite Subjects".

is *represented* in a <scene> (i.e. a picture, a list, or a file, which includes other objects as well). The following examples are adopted from [SEZ2003]:

- *Siz bütün resimler-de var mı-sınız?*
  you all pictures-LOC E Q-2PL
  "Are you in all the pictures?"
- *Ayşe bu dosya-da yok.*
  Ayşe this file-LOC NE
  "Ayşe is not in this file."

Compound Tense Existentials

Compound tense existential sentences are typically characterized by participles in a sentence. The tense of the relevant participle in such sentences semantically contributes to the tense and mood of the whole sentence, hence yielding to a new category of existentials.

- *Giyin-eceğ-im yok.*
  dress_up-FUTPART-P1SG NE
  "I do not feel like dressing up."
- *İki sene-dir on para kazan-dığ-ı yok-tur.*
  two year-ADVB ten buck earn-PASTPART-P3SG NE-COPULA
  "I suppose, he has not earned ten bucks for two years now."

Compound tense existentials inherently give way to ambiguous readings due to the participles they feature. Since participles are verbal adjectives, they can be used to modify nouns. In Turkish, it usually is the case that the modified nouns are absent in the sentence, and are assumed to be implied by the context. For instance, an alternate reading of the first example sentence would be as follows:

- *Giyin-eceğ-im*[11] *yok.*
  the_dress_that_I_will_wear NE
  "The dress that I will wear is gone/not here/missing."

Existentials in the Subordinate Clause

Existential meaning in the subordinate clause is captured either by the two existential particles *var/yok*[12], or the finite verbal stem *ol-* (meaning *to be/become*). For example:

- *Bir derd-in var-sa, [...].*
  a trouble-P2SG E-COND
  "If you have a trouble, […]"
- *Ayakkabılar-ı-nın ol-duğ-u çanta […]*
  shoes-P3SG-GEN be-PASTPART-P3SG bag
  "The bag, in which there were his shoes, […]"

2.5 Uses of *var/yok* beyond the Existential Scope

In the preceding sections, various semantic categories of existential sentences have been exemplified. The uses *var* and *yok* are not however restricted to the construction of existential sentences. This section illustrates these additional uses. The main motivation in presenting these is to complete the big picture about *all* possible uses of *var* and *yok*.

Compound Verbs

Var and *yok* can merge with some common verbs much like ordinary nouns and adjectives to form **compound verb**s. By forming a compound verb, they contribute to the forming of a totally new meaning as shown in the examples below. Therefore, they

[11] The exact sense of the tense and mood indicated by the participle suffix *-eceğ* is also ambiguous, but deemed to be inferred from the context.
[12] E.g. conditional (i.e. *var-sa, yok-sa*), temporal adverbial (i.e. *var-ken*, *yok-ken*)

should no longer be recognized as existential particles that constitute an existential sentence.

- *yok ol-mak*
  NE be-INF (i.e. infinitive)
  "to disappear"
- *var ol-mak*
  E be-INF
  "to come into existence" or "to live"
- *yok et-mek*
  NE do-INF
  "to make disappear" or "to destroy"
- *yok say-mak*
  E count-INF
  "to disregard"

Adnominal Modifiers

Existential particles may also be used as adnominal modifiers such as in the following examples. This is due to the adjectival position that they hold.

- *Var güc-üm-le vur-du-m.* [SEZ2003]
  E strength-P3SG-INS hit-PAST-1SG
  "I hit with all my (existing) might."
- *Yok hâl-im-le ora-ya git-ti-m.* [SEZ2003]
  NE state/energy-P1SG-INS there-DAT go-PAST-1SG.
  "I went there with my depleted energy."
- *Yok paha-sı-na sat-tı-m.*
  NE value-P3SG-DAT sell-PAST-1SG
  "I sold it cheap at half the price."

***Yok* as a Negative Interjection**

Yok is many times used as an interjection meaning *no* in colloquial Turkish:

- *Yok canım!*
  NE dear
  "No way!"
- *Yok, doğru-su iyi adam, kim ne der-se de-sin.*
  NE in_fact good man who what say-COND say-IMP(i.e. imperative)
  "No, in fact he is a good man, no matter who says what."
- *Ver-di-ler, ne âlâ; yok ver-me-di-ler, dön gel.*
  give-PAST-3SG what nice NE give-NEG-PAST-3SG turn come
  "If they give it, fine, if they do not, come back."

Idiomatic Usages

Turkish is a language in which idioms are abundantly used. Accordingly, *var* and *yok* also appear in numerous idioms. It is important for our work to discern all these idioms, not only because they render the language rich, but also because they potentially comprise exceptions when fed into a computer program[13]. Thus, these idiomatic usages should initially be encoded into the system so as not to allow for confusions during processing, due to their irregular linguistic constructs. An incomplete list of such idioms is provided in Appendix A.

[13] This owes to the nature of idioms. An idiom is a speech form or an expression of a given language that is peculiar to itself grammatically, or cannot be understood from the individual meanings of its elements.

Chapter 3

Corpus and Semantic Annotation

In this chapter, we first explicate the content and structure of the corpus that we worked on. Then the semantic annotation schema is presented in detail. This is done by defining the steps in creating such a schema at the outset, and then by systematically elaborating on how each step has been realized. These steps include the delineation of the abstract thematic roles used to define the predicate-argument relations of existential sentences, as well as the thematic role hierarchy among them. Finally, an example sentence from the semantically annotated corpus is presented.

3.1 Corpus Description

All experiments have been performed on the March 2004 release of the Turkish Treebank [OFL2003; SAY2002]. This is a portion of the METU Turkish Corpus[14], which is a 2 million word corpus of post-1990 written Turkish, sampled from approximately 16 main genres: news articles, novels, stories, academic papers, essays, travel writings, discussions, etc. [SAY2002].

[14] http://www.ii.metu.edu.tr/~corpus/corpus.html

The Treebank comprises 7262 sentences in total, accompanied with full morphological and surface dependency annotation on a sentential basis, of which only 292 instantiate the two particles *var* or *yok*. 232 of these sentences have been taken as existential sentences for additional manual semantic annotation (cf. Appendix B). Our manual semantic annotation schema is explained in Section 3.2.

The morphological annotation of the words in the Treebank reveals detailed syntactic information as to their parts of speech, and the sequence of inflectional groups separated by derivational boundaries that construct them. The major parts of speech that are present in our sub-corpus, which has been used for the experimentations described in Section 5, are listed in Table C.1 of Appendix C, with their counts.

The surface dependency annotation of the Treebank, provides yet further syntactic information. This dependency framework, which has been developed with similar motivations as those presented in [HAJ1998; BÉM2001; SKU1997; BRA2001; LEP1998], allows for the representation of the relationships among the lexical items in a sentence. Table C.2 of Appendix C displays the statistics regarding the relations that are present in the sub-corpus that we use.

3.2 Semantic Annotation Schema

In order to develop a semantic annotation schema, one has to first define semantic units at an abstract level that is both generic enough to be domain-independent, and specific enough to capture the whole semantic knowledge that one is interested in. The steps in creating such a schema are described as follows in [HAC2003b]:

- Decide on the type of semantic knowledge required,
- Develop a representation to encode it,
- Prepare annotated data,
- Design a method to acquire that knowledge by a machine.

The type of semantic knowledge that we want to capture from existential sentences is their predicate-argument semantic relations. These relations can best be encoded within certain semantic roles that are inherently assumed by the lexical constituents of a sentence. This set of roles is called as *abstract thematic roles*. According to this representative set, the data (i.e. corpus) is manually annotated. Finally, to automate this whole procedure, a supervised machine learning technique, which is known to be performing well on textual data, is used: Support Vector Learning. While the first three steps are further elaborated in the following sections, discussion on the fourth step will be saved until Chapter 4.

3.2.1 Abstract Thematic Roles

The set of abstract thematic roles depicted in Table 3.1 is developed for use in assigning the relations of the arguments to the predicate in an existential sentence.

Table 3.1: Abstract thematic roles and their definitions

THEME	Overt subject[15] of predicate *var/yok*
LOCATION	Place in which subject is situated
SOURCE	Entity from which subject originates
GOAL	Entity towards which subject heads
RELATION	Entity with which subject shares
POSSESSOR	Referent of subject that possesses
POSSESSED	Entity that is possessed

The main incentive in developing these particular seven thematic roles was to facilitate the consistent identification of the three semantic groups of existentials presented earlier in Chapter 2 (i.e. the bare, case, and possession existential groups), in

[15] Overt subject here should not be marked with possession information. Otherwise, POSSESSOR appears merely to be a special sub-case of THEME.

the later stages. The immediate correlation among these roles and their existential group counterparts is as follows:

- If THEME is the only role present in a sentence, then that sentence belongs to the bare *existentials* group.
- If either one of LOCATION, SOURCE, GOAL, and/or RELATION roles are present in a sentence, then that sentence belongs to the *case existentials* group.
- If there are no case existential roles within a sentence, but either one of POSSESSOR, and/or POSSESSED roles are present, then that sentence belongs to the *possession existentials* group.

The following construct exemplifies how these thematic roles get assigned to the words of an existential sentence[16] so as to define its predicate-argument structure:

- [$_{\text{POSSESSOR}}$ *O-nun*] [$_{\text{LOCATION}}$ *bu ev-de*] [$_{\text{POSSESSED}}$ *yer-i*] [$_{\text{predicate}}$ *yok*] [$_{\text{NULL}}$ *artık*].
  she-GEN this house-LOC place-P3SG NE anymore
  "She has no place in this house anymore."

3.2.2 Thematic Role Hierarchy

As can be inferred from the correlations between the thematic roles and their existential group counterparts presented in Section 3.2.1, there is a precedence relationship among the thematic roles. Accordingly, possession existential arguments have precedence over bare existential arguments, and case existential arguments have precedence over both of the other two. The motivation for this choice follows from the discussions in Chapter 2[17]. Any further interpretation requires deep semantic analysis and an exhaustive thematic exploration of Turkish existentials, thus is beyond the scope of this work. The hierarchy can be represented as follows:

Case Existentials > Possession Existentials > Bare Existentials

[16] This sentence is taken from the Treebank (cf. Appendix B).
[17] Note that we are making some common-sense assumptions in organizing the roles in the form of a hierarchy.

Various thematic roles, which designate different existential groups, may often appear in the same sentence. In those cases, the thematic role hierarchy is utilized to determine the category of existential group to which such a sentence belongs. With this reasoning, it can be straightforwardly deduced that the following example is a case existential sentence:

- [$_{\text{POSSESSOR}}$Onun] [$_{\text{LOCATION}}$bu evde] [$_{\text{POSSESSED}}$yeri] [$_{\text{predicate}}$yok] [$_{\text{NULL}}$artık].
  she-GEN this house-LOC place-P3SG NE anymore
  "She has no place in this house anymore."

3.2.3 Example of a *Modified* Treebank Sentence

This section presents an existential sentence as it appears in the customized sub-corpus of the Turkish Treebank. The primary modifications that we integrated into the original version of the Treebank consists of adding thematic role tags (i.e. SEM) to the words of 232 existential sentences, and removing two attributes (i.e. LEM and MORPH) from each word[18]. Figure 3.1 depicts an example structure extracted from the modified sub-corpus of the Turkish Treebank.

```
<S>
<W IX="1" IG='[(1,"o+Pron+PersP+A3sg+Pnon+Gen")]' REL="[4,1,(POSSESSOR)]" SEQ="0" SEM="3"> Onun </W>
<W IX="2" IG='[(1,"bu+Pron+DemonsP+A3sg+Pnon+Nom")]' REL="[3,1,(OBJECT)]" SEQ="0" SEM="7"> bu </W>
<W IX="3" IG='[(1,"ev+Noun+A3sg+Pnon+Loc")]' REL="[5,1,(LOCATIVE.ADJUNCT)]" SEQ="0" SEM="8"> evde </W>
<W IX="4" IG='[(1,"yer+Noun+A3sg+P3sg+Nom")]' REL="[5,1,(OBJECT)]" SEQ="0" SEM="5"> yeri </W>
<W IX="5" IG='[(1,"yok+Adj")]' REL="[7,1,(SENTENCE)]" SEQ="" SEM="15"> yok </W>
<W IX="6" IG='[(1,"artık+Adv")]' REL="[5,1,(MODIFIER)]" SEQ="1" SEM="0"> artık </W>
<W IX="7" IG='[(1,".+Punc")]' REL="[,( )]" SEQ="1" SEM=""> . </W>
</S>
```

Figure 3.1: Sample *modified* Treebank encoding of a Turkish sentence

More detail on the following explanations regarding the definitions of the attributes, except SEQ and SEM, can be found in [OFL2003]:

[18] The LEM attribute, which denotes the lemma of the word as it would appear in a dictionary, and the MORPH attribute, which indicates the morphological structure of the word as a sequence of morphemes, are null valued in the March 2004 release of the Turkish Treebank.

- IX denotes the index of the current word,
- IG is a list of pairs of an integer and an inflectional group,
- REL encodes the relationship of current word, as indicated by its last inflection group, to an inflectional group of some other word,
- SEQ denotes whether the current word appears before or after the predicate,
- SEM numerically encodes the thematic role of the current word, according to the *IOB* representation (e.g. ***O****(utside)* is 0, ***B***_THEME is 1, ***I***_THEME is 2, ***B***_POSSESSOR is 3, ***I***_POSSESSOR is 4, etc[19].).

[19] SEM is valued as null for periods, exclamation marks, question marks, and commas that are *not* used to separate the elements in a series. In those cases, their values are interpreted as 0.

Chapter 4

Methodology

Our approach consists of mainly two tasks: First we do shallow semantic parsing via support vector learning, and then we do sentence categorization to find the semantic group of existentials a sentence belongs to. Section 4.1 discusses our shallow semantic parsing approach. In doing so, the features used in support vector learning, and the learning approach itself has been detailed. A brief overview of support vector machines as our classifier, accompanied with the motivations for choosing it to use in our work has been presented. Finally, the sentence categorization task has been elaborated. The measures of evaluating this system's performance have been defined.

4.1 Shallow Semantic Parsing

Shallow semantic parsing process is regarded as comprising three steps. The first step is the identification of the predicate whose arguments are to be classified. The second step is the identification of words or phrases that represent the semantic arguments of that predicate. Finally, the third step assigns specific argument class labels to those words or phrases.

Variants of shallow semantic parsing have been explored by NLP researchers. In [HAC2003a] two broad classes are described: one is referred to as constituent-by-constituent (C-by-C) and the other as word-by-word (W-by-W) classification. The description goes as follows:

> 'In the C-by-C method, we first linearize the syntactic tree representation of a sentence into a sequence of its syntactic constituents. Then we derive features for each constituent and do classification. […] In the W-by-W method we derive features for each word and decide whether the word is inside a chunk or outside the chunk with a specific role label. As in the former method, this task can also be accomplished in two stages; first segment sentences into chunks and then label them.'

There is yet neither a functional full statistical syntactic parser nor a chunker available for Turkish, which are necessary for architectures employing the C-by-C or W-by-W approaches. Since full parsing is computationally more expensive than chunking, and since it is easier to develop chunkers than full statistical syntactic parsers for new languages, building a phrase chunker seems to be the best possible approach. Nevertheless, the structure of the Turkish Treebank allows us to bypass the development of either one of these two architectural elements. The way each word is tagged in the Turkish Treebank, provides us with sufficient syntactic information that is of comparable use that one would obtain from either a full parser or a chunker for Turkish for the purposes of this research.

Equipped with the information from the Treebank, our system first does semantic classification at the word-level similar to the W-by-W method. In [RAM1995], chunking is proposed as a tagging task and thus a convenient data representation for chunking is presented. Having been inspired by this work, where each word in a sentence is labeled using *IOB* representation, we specifically adopt the *IOB2* representation [SAN1999; RAT1998], according to which each word in a sentence is tagged with either *I*, *O*, or *B*,

respectively meaning that it is inside a chunk[20], outside a chunk, or at the beginning of a chunk. In keeping with this representation, the previous example sentence is tagged as:

[B_POSSESSOR Onun] [B_LOCATION bu] [I_LOCATION evde] [B_POSSESSED yeri] [predicate yok] [NULL artık] [NULL .][21]

4.1.1 Features

There are five features in our baseline system, which encode most of the information given for each word in the Turkish Treebank. These features are the part-of-speech category of the word, the part-of-speech category of the word that this word is linked to within the sentence, the name of this syntactic relation, and whether this word appears before or after the predicate.

The system creates the set of features for each word to be tagged from a fixed-size context that centers the word-in-focus. Therefore, the overall number of features for each word to be tagged comprises not only those that are its own, but also those features that belong to the words that appear before or after it within its context. This notion of context can be illustrated as a forward-sliding window centered at the current word as in Table 4.1. For the first and last words of a sentence, the previous and following words' features are all assigned null values, respectively, since there exist no such words in the context. The idea can be extended to cover the words that follow the first word or precede the last word of the sentence, for the larger-sized context windows.

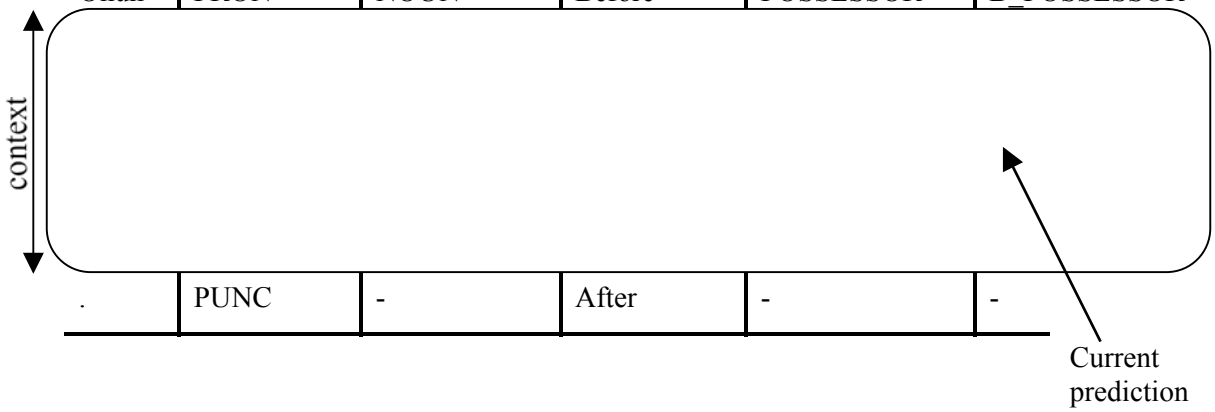
Subsequent to our evaluation of the system with the five features, we added one more feature, and repeated our experiments. This sixth feature is called the semantic class, and its value takes on the IOB tag of the previously classified word(s) that precedes the word-in-focus and appears in the same fixed-size context. For those words that follow the word-in-focus and appear in the same context, the semantic class feature

[20] By c*hunk*, we mean a thematic role chunk, referring to a word group that forms an argument of a predicate. We do not necessarily suggest that *chunk*s be syntactically correlated words within a sentence.
[21] Turkish Treebank treats all punctuation marks as words; hence an argument label is also necessary here.

takes on a null value. Obviously, this feature takes on a null value also for the word-in-focus.

Table 4.1: 5-word context and features used to classify a word

Word	**POS (self)**	**POS (target)**	**Position**	**Relation**	**Semantic Class**
Onun	PRON	NOUN	Before	POSSESSOR	B_POSSESSOR
.	PUNC	-	After	-	-



For purposes of experimentation we tried our system with 3-word, 5-word, and 7-word sized contexts. The evaluation of the results of these various-sized windows is reported in Chapter 5, where comparisons among them are also given.

4.1.2 The Classifier: SVM

Chunking and subsequent labeling of a sentence into its arguments with respect to a given predicate is formulated as a classification-based learning task. As has already been reported in the literature [JOA1998; KUD2000], Support Vector Machines have advantage over conventional statistical learning algorithms, such as Decision Tree and Maximum Entropy models because of their capability of being universal learners and their high generalization performance. SVMs can carry out their learning with all combinations of given features without increasing computational complexity by introducing the Kernel function. Conventional algorithms cannot handle these combinations effectively, thus enforcing the implementer to taking the trade-off between accuracy and computational complexity into account. Furthermore, SVMs' capability to

learn can be independent of the dimensionality of the feature space. Conventional algorithms require careful feature selection to avoid over-fitting [TAI1999].

We employ Support Vector Machines mainly due to their capability to handle a large number of features with strong generalization properties. Nonetheless, SVMs are binary classifiers whereas semantic parsing is a multi-class classification problem. So as to address this issue, several methods have been proposed to extend SVMs for multi-class classification [HSU2002]. All of these methods that are used to extend binary to multi-class classification fall into either one of the following two common approaches: *pairwise* and *one class versus all others*.

In the *pairwise* approach, a separate binary classifier is trained for each of the class pairs, which requires the training of *K*(K-1)/2* binary classifiers. The outputs of all of these classifiers are in the end combined to predict the classes.

In the *one class versus all others* approach, *K* classifiers are trained for a *K*-class problem. Each classifier is trained to discriminate between examples of each class and those belonging to all other classes combined.

Among the two approaches, there is a tradeoff between the number of classifiers to train and the amount of data used in training each classifier. It is a topic of controversy regarding which approach performs better. Some researchers report that *pairwise* approach is better [KRE1999], while others report the opposite [HAC2003a]. Therefore we used both approaches. The results are compared and evaluated in Chapter 5.

4.2 Sentence Categorization

Once the predicate-argument structure of a sentence is captured, subsequent to the classification-based learning task achieved by SVMs, the system categorizes each sentence to the most appropriate one of the existential groups. Table 4.2 depicts the counts and percentages of these existential groups within our sub-corpus.

Table 4.2: Overall numbers and the percentages of each category of existentials

	Count	**%**
Case Existential Sentences	102	43.97
Possession Existential Sentences	68	29.31
Bare Existential Sentences	62	26.72

As have already been described in Chapter 3 of this thesis, there are altogether seven thematic roles that determine to which of the three groups of existentials the sentence belongs. According to the thematic role hierarchy, devised to account for the precedence relationships among the thematic roles, the system automatically and trivially categorizes each sentence as a case, possession, or a bare existential.

We evaluate the performance of this task separately within each existential group based on the following three measures:

- *Precision (P):* Percentage of recognized sentences that are correctly categorized as belonging to a certain existential group.
- *Recall (R):* Number of sentences the system correctly categorized as belonging to a certain existential group divided by the *actual* number of sentences that belong to that existential group.
- *F-score (F):* A combined measure, which is an approximation to the weighted geometric mean of *precision* and *recall*. The F-score is defined as:

$$F_\beta = \frac{(\beta^2 + 1) * P * R}{P + R}$$

where β is a parameter encoding the relative importance of *precision* and *recall*. We take β = 1, meaning that *P* and *R* is weighted equally. These three measures are calculated for each of the three existential groups. For instance, for case existentials, the precision measure is the number of correctly categorized case existentials to the overall number of

the case existentials that the system found. The recall measure, on the other hand, is the number of the correctly categorized case existential sentences to the overall number of case existentials that indeed exist.

Apart from these three measures, we also evaluate the overall accuracy of the categorization process by calculating the percentage of the correctly categorized sentences (that belong to either one of the three existential groups) among all sentences that the system categorized.

Chapter 5

Experiments and Results

For all our semantic tagging experiments, we used the LIBSVM[22] software[23]: Its standard package for the *pairwise* approach (OVO, meaning one versus one), and one of its multi-class classification tools for the *one versus all* (OVA) approach. In our initial experiments, we also tried the DAGSVM[24] method [PLA2000]. However, it produced significantly worse results according to the other two methods. Therefore, this chapter evaluates and compares the results of only the OVO and OVA methods.

In the learning experiments explained in sections 5.1.2 and 5.2.2, the train and test sets have been formed via a 9 vs. 1 split. The statistics regarding the train and test data is depicted in Table 5.1. Note that the percentages of the case, possession, and bare existential sentences are kept constant in both the train and the test files, which also approximately comply with the 9 vs. 1 split. One other issue taken into consideration in forming the test data was to include those instances that have been covered in the train data as much as possible.

[22] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[23] The system uses a radial basis function kernel with cost c = 480, and gamma g = 0.0078125.
[24] DAGSVM stands for Directed Acyclic Graph Support Vector Machines. Its training phase is same as the OVO method, with the additional use of a rooted binary directed acyclic graph in its testing phase.

Table 5.1: Train and test data statistics

	Train Data	Test Data
Number of Words	1538	171
Number of Sentences	208	24
Number of Case Existential Sentences	91	11
Number of Possession Existential Sentences	61	7
Number of Bare Existential Sentences	56	6

The first set of experiments that we conducted has been tested on the input files where one word has five features[25], hence leaving out the semantic class feature. These experiments are explained in Section 5.1. Then, Section 5.2 explains the same experiments conducted on the input files, in which the context information is improved with the introduction of the semantic class feature.

5.1 Without Semantic Information

5.1.1 Cross Validation

In *v*-fold cross-validation, the train set is divided into *v* subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining *v*-1 subsets. Thus, each instance of the whole train set is predicted once, and the cross-validation accuracy is the percentage of data that are correctly classified. Table 5.2 depicts the cross validation accuracies, accomplished via OVO and OVA, with 5, 10, 20, and 30 folds for each three window-sizes. One significant point about this table is that in all cases, OVA method accomplishes higher accuracies. Accordingly, the highest value is the OVA accuracy with the 30-fold cross-validation for the 3-word window.

[25] The window structure is already assumed here, hence the number of features for each word in fact is: 5*(window_size)

Table 5.2: Cross validation accuracy, involving no semantic features

	OVO			OVA		
	3	5	7	3	5	7
5 fold	62.90	61.67	61.56	62.96	61.73	62.43
10 fold	62.32	61.97	62.55	64.77	63.02	64.01
20 fold	61.79	61.85	62.08	65.54	63.66	63.14
30 fold	61.85	62.02	62.20	65.59	-	-

Despite the higher accuracy percentages OVA method accomplishes, its training time is notably longer than that of OVO. Particularly as the number of folds increase, the training time lengthens so much with OVA that it no longer is meaningful to carry out the experiment[26]. This is why we left the 30-fold experiments with OVA for the 5- and 7-word windows unattended.

5.1.2 Classification

The classification accuracy percentages on the test set is reported in Table 5.3. Along with the accuracies of the OVO and OVA methods for the 3-, 5-, and 7-word windows, the mean squared errors (MSE), squared correlation coefficients (SCC), and the accuracy rates are also reported.

Table 5.3: Classification accuracy, involving no semantic features

	OVO			OVA		
	3	5	7	3	5	7
ACC (%)	65.4971	68.4211	71.9298	68.4211	70.7602	68.4211
MSE	6.8538	6.0117	5.3041	7.0702	6.9825	6.1696
SCC	0.7567	0.7826	0.8032	0.7437	0.7478	0.7704
ACC Rate	112/171	117/171	123/171	117/171	121/171	117/171

[26] It should be noted that our data size is nowhere close to being large, when considered among the data files in literature and daily life, and that the training time is bound to increase in line with this size.

The best accuracy is achieved with the OVO method on the 7-word window, such that the error term is the smallest and the SCC value is the largest[27]. An expected effect of the context size is seen in the experiment with the OVO method, where the accuracy increases with the size of the context. A minor point to recognize is the accuracy of the OVA method on the 3-, and 7-word windows: although their accuracy percentages and accuracy rates are the same, the MSE and SCC values reveal that the 7-word context is actually more accurate.

5.1.3 Sentence Categorization

According to the overall sentence categorization accuracy percentages depicted in Table 5.4, the most successful categorization takes place after the OVO tagging for the 5-word window with a value of 83.3333 percent.

Table 5.4: Existential sentence classification accuracy, involving no semantic features

	OVO			OVA		
	3	5	7	3	5	7
ACC (%)	70.8333	83.3333	79.1667	79.1667	70.8333	79.1667

The precision, recall, and F-score values for the categorization process of each of case, possession, and bare existential sentences have been depicted in Table 5.5. The results of previously conducted OVO and OVA experiments influence categorization equably in that they both allow best performance on categorizing case existential sentences. The worst performance, on the contrary, is on categorizing bare existential sentences. This is strongly linked with the fact that case existential sentences occur the most in both the train and test data, whereas bare existentials occur the least. Another interesting issue to note is that the size of the context does not seem to have an effect on the categorization process.

[27] The most it can be is 1 (the system achieves 100% accuracy).

Table 5.5: Precision, recall, F-score values for existential sentence categorization, with no semantic feature

	OVO								
	P (%)			R (%)			F-score		
	3	5	7	3	5	7	3	5	7
CASE	90.91	100.00	91.67	90.91	100.00	100.00	90.91	100.00	95.65
POSS.	50.00	66.67	62.50	57.14	85.71	71.43	53.33	75.00	66.67
BARE	60.00	75.00	75.00	50.00	50.00	50.00	54.55	60.00	60.00

	OVA								
	P (%)			R (%)			F-score		
	3	5	7	3	5	7	3	5	7
CASE	91.67	73.33	100.00	100.00	100.00	100.00	95.65	84.61	100.00
POSS.	62.50	62.50	58.33	71.43	71.43	100.00	66.67	66.67	73.68
BARE	75.00	100.00	100.00	50.00	16.67	16.67	60.00	28.58	28.58

It should be observed that the results presented in Tables 5.4 and 5.5 do not directly correlate with the results presented in Table 5.3. Although some words may be inaccurately tagged via the SVM, the categorization of the sentence containing those words may still be done correctly. So the correlation between these tables might best be described as follows: If SVM correctly tags all the arguments of an existential sentence, then the categorization of that sentence is surely to be done accurately. Otherwise, the sentence categorization process is likely to err, although this may not always be the case.

5.2 With Semantic Information

The addition of the semantic class feature to our input files of 3-, 5-, and 7-word windows, significantly improved the results. Initially, our intention was to incorporate this semantic class feature, for those words that precede the word-in-focus within the fixed-size context, *during* the tagging task. That is, the predicted semantic tag of the previous words in context would be used to predict the semantic tag of the word-in-focus. However, this required the constant interruption of the LIBSVM's prediction

program, after the tagging process of each word within the test data, in order to update the test data for the prediction of the next word. Since it would not be feasible to modify the LIBSVM code to function in this manner, we developed an approach to account for this effect instead.

The system is first supplied with the correct tags as the semantic class features of those words that precede the word-in-focus and appear in the context[28] in both the train and test data. Then it tests the accuracy of classification on this test data, which corresponds to the first row of Table 5.6. This row of results is the best one in the table, since the system is supplied with the correct tags for the semantic class feature initially. To diminish the effect of the all-correct hand-coded semantic tags introduced to the system initially, the process is iterated two more times, in each of which the output file of the previous test phase is incorporated into the test data of the next phase. At the end, the classification results of the system are not fully based on the correct semantic tags that were introduced initially. Instead they are based on the previous prediction values. Hence the system that we had in mind to begin with is impartially simulated. Table 5.6 depicts the classification accuracies on each of the iterations mentioned.

Table 5.6: Classification accuracies for each step, where each test data incorporate prediction values from previous predictions

	OVO			OVA		
	3	5	7	3	5	7
1st ACC %	88.3041	85.9649	87.1345	88.3041	87.7193	87.7193
2nd ACC %	81.2865	76.0234	76.6082	80.7018	77.7778	77.1930
3rd ACC %	76.0234	69.0058	70.7602	76.0234	73.0994	73.0994

5.2.1 Cross Validation

The cross-validation conducted here is similar to the process done for Section 5.1.1. Results are shown in Table 5.7. Except the fact that the highest value of this table is for

[28] Recall that if there are no previous words, then this feature –and all features– of those non-existing previous words are taken as null.

the 3-word window, all the remaining characteristics are quite the opposite of Table 5.2. Here, the OVO method performs better both in terms of higher results and in terms of time. Again, some experiments with the OVA method have not been done, since it takes hours to train the data.

Table 5.7: Cross validation accuracy, involving semantic features

	OVO			OVA		
	3	5	7	3	5	7
5 fold	83.15	82.62	80.75	82.09	80.51	80.05
10 fold	84.38	83.44	81.98	82.80	80.98	80.16
20 fold	84.61	83.62	81.51	83.44	81.92	80.69
30 fold	84.38	83.50	81.63	-	-	80.63

On the average, 20-fold cross-validation seems to have achieved the best accuracy for all window sizes. This pattern could not have been observed from Table 5.2.

5.2.2 Classification

Table 5.8 displays the classification accuracies and their corresponding MSE and SCC values for windows of 3-, 5-, and 7-words. Note that each column in this table is the detailed representation of each corresponding element of the last row of Table 5.6.

Table 5.8: Classification accuracy, involving semantic features

	OVO			OVA		
	3	5	7	3	5	7
ACC (%)	76.0234	69.0058	70.7602	76.0234	73.0994	73.0994
MSE	5.8830	5.7778	4.8304	5.2339	5.0175	5.8363
SCC	0.7876	0.7921	0.8190	0.8039	0.8137	0.7775
ACC Rate	130/171	118/121	121/171	130/171	125/171	125/171

The highest accuracy has been achieved with the OVA method on the 3-word window. This is consistent with the overall performance of OVA for this set of experiments, as it returns higher results for all three window sizes, when compared to the OVO results. Although the size of the context does not seem to have a particular correlation with the accuracy of the classification for either one of the two methods, it is observable that 5- and 7-word windows do not significantly outperform one another, whereas 3-word window outperforms both with each method.

5.2.3 Sentence Categorization

According to the overall sentence categorization accuracy percentages depicted in Table 5.9, the most successful categorization takes place after the OVA tagging for the 5-word window with a value of 87.5 %. This is better than the highest result achieved in the case where semantic features were not included (see Table 5.4).

Table 5.9: Existential sentence classification accuracy, involving semantic features

	OVO			OVA		
	3	5	7	3	5	7
ACC (%)	79.1667	75.0000	79.1667	75.0000	87.5000	75.0000

The precision, recall, and F-score values for the categorization process of each of case, possession, and bare existential sentences have been depicted in Table 5.10. The OVO method results allow the best performance on categorizing bare existential sentences, whereas the OVA method results allow best performance on categorizing case existentials.

In comparison to Table 5.5, Table 5.10 has less deviation in F-score values: There is neither as high values, nor as low values as there are in Table 5.5.

Table 5.10: Precision, recall, F-score values for existential sentence categorization, with semantic features

	OVO								
	P (%)			R (%)			F-score		
	3	5	7	3	5	7	3	5	7
CASE	100.00	100.00	81.82	81.82	81.82	81.82	90.00	90.00	81.82
POSS.	71.43	55.56	62.50	71.43	71.43	71.43	71.43	62.50	66.67
BARE	62.50	66.67	100.00	83.33	66.67	83.33	71.43	66.67	90.91

	OVA								
	P (%)			R (%)			F-score		
	3	5	7	3	5	7	3	5	7
CASE	73.33	84.62	100.00	100.00	100.00	72.73	84.61	91.67	84.21
POSS.	75.00	85.71	60.00	42.86	85.71	85.71	54.55	85.71	70.59
BARE	80.00	100.00	66.67	66.67	66.67	66.67	72.73	80.00	66.67

Chapter 6

Conclusions

In this thesis, we described a novel way of utilizing the Turkish Treebank for domain-independent shallow semantic parsing of Turkish existential sentences by recognizing their predicate-argument structures. This facilitated the system to further categorize these sentences into the most appropriate semantic group of existentials. The categorization task is automatically done by exploiting a thematic role hierarchy that we devised to account for the precedence relationships among the semantic roles, which are assigned to arguments of an existential sentence.

In order to perform this research, we had to complete several preliminaries. At an initial phase of the work, we had to systematically categorize the semantic types of Turkish existential sentences in print. We did this after consulting renowned grammars of Turkish and finding hundreds of existential sentences that later guided us in the process. We then had to develop a set of domain-independent abstract thematic roles to be assigned to the arguments of existential sentences. These thematic roles were then used to semantically hand-annotate 232 existential sentences from the Turkish Treebank. An inevitable and time-consuming stage was the refining of the corpus used for our work.

Our results prove that the incorporation of semantic information to the input files of the SVM is at least a 5% improvement in obtaining higher accuracies on the task of classifying arguments of an existential sentence. This indicates promise for applications in various natural language tasks in Turkish, and those that particularly work with the Turkish Treebank. The results of the task of existential sentence categorization, on the other hand, did not seem to get affected in any way by the incorporation of semantic information to the system. This owes to the fact that classification is done on a word basis and semantic information that we incorporate is also a word-level feature. However categorization is done on a sentence-level and the addition of a word-level feature does not seem to have a significant effect in the sentence categorization results. In spite of this, improving the argument classification system should always have a positive effect on the sentence categorization process, since categorization functions trivially and must return correct categories for sentences whose words are all correctly classified.

The evaluation of the whole system shows that the annotation of the Turkish Treebank is fair enough and that the incorporation of the thematic role tags will indeed be of use to perform research that deals with this Treebank and hence Turkish language in general.

Although our results are promising, there still are various ways to improve them. A more consistently annotated corpus would without doubt yield better results. Any inconsistency in the annotation of the corpus causes SVMs to wrongly model the train data and hence can have disastrous end-effects concerning the classification phase. So as to avoid this, the annotation should be done both correctly *and* consistently, which requires rather conscientious work. Also, the size of the data affects our results, since we are doing machine learning. Increasing the size of the data set will help improve our results. The more instances covered in the train set imply the better learning of the system (i.e. SVMs), hence better classifying the test data.

Many aspects of our system are still quite preliminary. For instance, we currently handle only three types of existential sentences. The system can be extended to

differentiate among senses of all existential sentences, which requires a deeper semantic analysis to be able to encode each one's predicate-argument structure. Moreover, the incorporation of the semantic information into the system can be made more robust such as by disallowing the *I* tags, if they are not at an appropriate point preceded by a *B* tag. With the development of such controlling additional features to the system, the overall accuracy might be increased.

References

[ATA2003] N. B. Atalay, K. Oflazer, and B. Say. The Annotation Process in the Turkish Treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora – LINC*, 2003.

[BAK1998] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet Project. In *Proceedings of the COLING / ACL* , pp. 86-90, 1998.

[BÉM2001] A. Bémová, J. Hajič, B. H. J. Panenová, A. Böhmova, and E. Hajičová. The Prague Dependency Treebank. In A. Abeillé (ed.), *Building and Exploiting Syntactically Annotated Corpora*. Kluwer Academic Publishers, 2001.

[BLA2000] D. Blaheta and E. Charniak. Assigning Function Tags to Parsed Text. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*, pp. 234-240, 2000.

[BRA2001] T. Brants, W. Skut, and H. Uszkoreit. Syntactic Annotation of a German Newspaper Corpus. In A. Abeillé (ed.), *Building and Exploiting Syntactically Annotated Corpora*, pp. 73-88. Kluwer Academic Publishers, 2001.

[CHE2003] J. Chen and O. Rambow. Use of Deep Linguistics Features for the Recognition and Labeling of Semantic Arguments. In *Proceedings of the*

Conference on Empirical Methods in Natural Language Processing, 2003.

[FLE2003] M. Fleischman and E. Hovy. A Maximum Entropy Approach to Framenet Tagging. In *Proceedings of the Human Language Technology Conference*, pp. 22-24, 2003.

[GIL2002a] D. Gildea and D. Jurafsky. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3): 245-288, 2002.

[GIL2002b] D. Gildea and M. Palmer. The Necessity of Parsing for Predicate Argument Recognition. In *Proceedings of the 40th Annual Conference of the ACL*, pp. 239-246, 2002.

[GIL2003] D. Gildea and J. Hockenmaier. Identifying Semantic Roles Using Combinatorial Categorial Grammar. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 57-64, 2003.

[HAC2003a] K. Hacioglu, S. Pradhan, W. Ward, J. Martin, and D. Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. *TR-CSLR-2003-1*, Center for Spoken Language Research, Colorado, 2003.

[HAC2003b] K. Hacioglu and W. Ward. Target Word Detection and Semantic Role Chunking Using Support Vector Machines. In *Proceedings of the Human Language Technology Conference*, 2003.

[HAJ1998] J. Hajič. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In E. Hajičová (ed.), *Issues of Valency and Meaning: Studies in Honour of Jarmila Panenova.* Karolinum – Charles University Press, pp. 106-132, 1998.

[HEM1990] C. T. Hemphill, J. Godfrey, and G. R. Doddington. The ATIS Spoken Language Systems Pilot Corpus. In *Proceedings DARPA Speech and Natural Language Workshop,* pp. 96-101, 1990.

[HSU2002] C. W. Hsu and C. J. Lin. A Comparison of Methods for Multiclass Support Vector Machines. *IEEE Transactions on Neural Networks*, 13: 415–425, 2002.

[JOA1998] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning (ECML)*, pp. 137-142, 1998.

[JUR2000] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2000.

[KIN2002] P. Kingsbury and M. Palmer. From TreeBank to PropBank. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*. 2002.

[KRE1999] H. G. Kressel. Pairwise Classification and Support Vector Machines. In B. Scholkopf, C. Burges, and A. J. Smola (eds.) *Advances in Kernel Methods – Support Vector Learning*. MIT Press, pp. 255-268, 1999.

[KUD2000] T. Kudoh and Y. Matsumoto. Use of Support Vector Learning for Chunk Identification. In *Proceedings of the 4th Conference on CONLL-2000 and LLL-2000*, pp. 142-144, 2000.

[KON1956] A. N. Kononov. *Grammatika Sovremennogo Tureckogo Literaturnogo Jazyka*. Moskva-Leningrad, 1956.

[LEP1998] Y. Lepage, A. Shin-Ichi, A. Susumu, and I. Hitoshi. An Annotated Corpus in Japanese Using Tesniere's Structural Syntax. In *Proceedings of COLING-ACL'98 Workshop on the Processing of Dependency-Based Grammars*, 1998.

[LEW1967] G. L. Lewis. *Turkish Grammar*. Oxford University Press, 1967.

[OFL2003] K. Oflazer, B. Say, D. Z. Hakkani-Tür, and G. Tür. Building a Turkish Treebank. In A. Abeillé (ed.), *Building and Exploiting Syntactically Annotated Corpora*, pp. 1-18. Kluwer Academic Publishers, 2003.

[PLA2000] J. C. Platt, N. Cristianini, and J. Shawe-Taylor. Large Margin DAGs for Multiclass Classification. *Advances in Neural Information Processing Systems*, 12: 547-553. MIT Press, 2000.

[PRA2003] S. Pradhan, K. Hacioglu, W. Ward, J. H. Martin, and D. Jurafsky. Semantic Role Parsing: Adding Semantic Structure to Unstructured Text. In *Proceedings of the International Conference on Data Mining*, 2003.

[RAM1995] L. A. Ramshaw and M. P. Marcus. Text Chunking Using Transformation Based Learning. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pp. 82-94, 1995.

[RAT1998] A. Ratnaparkhi. Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD Thesis, Computer and Information Science, University of Pennsylvania, 1998.

[SAN1999] E. Sang and J. Veenstra. Representing Text Chunks. In *Proceedings of EACL*, pp. 173-179, 1999.

[SAY2002] B. Say, D. Zeyrek, K. Oflazer, and U. Özge. Development of a Corpus and a Treebank for Present-Day Written Turkish. In *Proceedings of 11th International Conference on Turkish Linguistics*, 2002**.**

[SKU1997] W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. An Annotation Scheme for Free Word Order Languages. In *Proceedings of 5th Conference on Applied Natural Language Processing*, pp. 88-95, 1997.

[SUR2003] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the 41st Annual Meeting of the ACL*, pp. 8-15, 2003.

[SEZ2003] E. Sezer. On Syntactic and Semantic Properties of Turkish Existential Sentences. Unpublished manuscript. Harvard University, 2003.

[TAI1999] H. Taira and M. Haruno. Feature Selection in SVM Text Categorization. In *Proceeedings of AAAI-99*, pp. 480-486, 1999.

[THO2003] C. Thompson, R. Levy, and C. Manning. A Generative Model for Semantic Role Labeling. In *Proceedings of the European Conference on Machine Learning (ECML)*, pp. 397-408, 2003.

[UND1976] R. Underhill. *Turkish Grammar*. MIT Press, 1976.

Appendix A

Idiomatic Uses of Existential Particles

1. var akar, yok bakar
2. var eli titremez
3. var evi, kerem evi; yok evi, verem evi.
4. var mı pulun herkes kulun, yok mu pulun ahrettir yolun
5. var ne bilsin yok halinden
6. var olsun, yerinde olsun
7. var varlatır, yok söyletir
8. var yok
9. vara yoğa
10. vardan yoktan anlamaz
11. varı yoğu
12. varını veren utanmamış
13. varsa yoksa
14. yok canım
15. yok devenin başı/pabucu
16. yok oğlu yok
17. yok satmak
18. yok yere

19. yok yoksul
20. yoktan anlamamak
21. yoktan var etmek

Appendix B

List of All Sentences Used in Experiments

1. Ne kadar çok insan var.
2. Balıklar geldi, ayıklanacak; sebzeler soyulacak, meşrubat taşınacak, siz yoksunuz.
3. Tutunabileceğim bir şey yoktu.
4. Ya reddedilirsem! korkusu var.
5. Günahların bağışlandığı durumlar var.
6. Ama bir ihtimal var, değil mi?
7. Canım, bu kadar tepki gösterecek ne var?
8. Ölmek var dönmek yok.
9. Olayın ilk günü düzenlenmiş bir rapor var örneğin.
10. O kadar çok üvey anne veya babası olan çocuk var.
11. Kot pantolonla gelen hiçbir çocuk yoktu.
12. Bir bardak su var.
13. İçki var.
14. Kendine dönmek isteyen bir insanın gidebileceği neresi vardır?
15. Çok moda bir kahve varmış.
16. Her yaştan insan vardı.

17. Çıplak, doğrudan doğruya tadını duyuran içkiler var.
18. Polisten yardım istememden başka çare yok.
19. Kendimi kandırabileceğim bir şey yok.
20. Ah'lara, vah'lara gerek yok.
21. Vardır, vardır, olmaz olur mu?
22. Var.
23. İlkin hiçbir ışık, hiçbir gölge yoktur.
24. Çiller, Karayalçın, Boyner de yoktu.
25. Onlardan çok var.
26. Kriz nedeni bu yaşamsal sorunlardan biri olsa mesele yok.
27. Ama kriz var.
28. Anlayış birliği var.
29. Böyle bir acı var.
30. Önemli bir şey yok.
31. İki alternatif var.
32. Üniversitelerin tamamının mutabık olduğu ancak sadece sistem üzerinde adeta AKP gibi oturan birkaç kişinin kendi yerlerinden doğan kaygılarıyla rahatsız olduğu durum var.
33. Çünkü ondan uzunu yoktu.
34. Korucu var.
35. Korucu yoktu.
36. Var mı öyle kalleşlik.
37. Dolaştığın o itler var...
38. Benim büyük amcam var...
39. Oldum olası merak ettiğim bir konu vardı.
40. Sol kaşının üstünden başlayan, yanağına kadar inen bu derin iz de yoktu.
41. Deniz saati geçmesine karşın hâlâ yüzen birkaç kişi var.
42. Çünkü onların yararı yok.
43. Hem matematiğin hem de bilim dünyasının kullandığı ortak kavramlar vardır.
44. Bir gün önce her şey var, bir gün sonra hiçbir şey...
45. Senin de kendini asma ihtimalin var.

46. Bir şey yok!
47. Beyefendi, sayın, saygıdeğer demek yoktu.
48. O zaman, klasik şube vardı.
49. Özel televizyonlar da yok.
50. Özel televizyonların çok efendi yorumcu bozumcuları da yok...
51. Yerleştir yerleştirme, devşir devşirme sınavları falan filan yok...
52. Zekâmetre yok!
53. Bilmeyecek ne var?
54. Aklımızdan çıktığı yok.
55. Vaktiyle, genç bir çoban varmış.
56. Şimdi her mevsim var.
57. Patlıcan burunlu var da, neden patlıcan kulaklı yok?
58. O naz ettiyse kurusu var.
59. Bir kere sık sık dışarı çıkarmak yoktu.
60. Abicim kız bir köpek yok mu!
61. Niye bir tane yok?
62. Yağma yok!
63. Aç demeden açmak yok ha!
64. Yürüyelim mi biraz, sakıncası yoksa?
65. Tonik yok.
66. Buz kesiyorum gitgide, aldırdığı yok.
67. Burada herkesin dükkânı var.
68. Benim pantolonda ateş var.
69. Dünyanın merkezinde kendisi vardı.
70. Şuramda kocaman bir delik vardı.
71. Eroin ve tüm uyuşturucu kullanımında bir duyguları öldürme eşiği vardır.
72. Bağımlı kişinin beyninde duygular eşittir acı çekmek formülü vardır.
73. Bizimkinde, bilinmeyen bir Öteki Dünya'da aklımıza hayalimize gelmeyecek cezalara uğrama korkusu var.
74. Tabii bu onlarda da var.
75. Ama arada fark da var.

76. Kaldırımda küçük süt ve yoğurt kutularından bir-iki tane, üzerinden kavun kabukları dökülmüş bir iki çöp poşeti ve izmaritler vardı.
77. Üstelik burada benim bir günahım yok aslında.
78. Ama hiç değilse bir umut var burada...
79. Üstelik Kemal'in dediğinde herkesin her şeyini ortaya dökmek yok.
80. Burada ... yoktum ...
81. Uçak motorlarında buzlanma yoktur.
82. Ben hanımefendiyle beraber davaya başladıktan sonra elimizde belge, doküman yoktu.
83. Kamuoyunda da bu olayın kaza olmadığı, suikast olduğu gibi yaygın bir kanaat vardı, o kadar.
84. İkilinin elinde Askeri Savcılığın verdiği takipsizlik kararı ve KKK Uçuş Emniyet Kurulu'nun düzenlediği müşterek kanaat raporu vardı.
85. Usulde bir hata yoktu.
86. Fakat bu sefer önlerinde araştırmalarını geliştirmelerini önleyen yeni bir engel vardı: zaman...
87. Önlerinde hazırlanan raporların hepsi vardı artık...
88. Heyette dört üye daha vardı: Kurmay Pilot Albay Tünay Çelen, Pilot Binbaşı C.
89. Bizim evde bizimkilerin hep acelesi vardır.
90. Eve geldiklerinde de yapacakları bir şeyler vardır.
91. Gazetelerde kaçak çocuklar var.
92. Bizde sana verecek para yok.
93. Onun bu evde yeri yok artık.
94. Salonda babamın birçok tanıdığı vardı.
95. Ilık et suyunun yanında haşlanmış geyik eti, kızarmış ekmek, bir de ılık çay vardı masada.
96. Bu kamp alanında iki ev var.
97. Bir de en önemlisi, ot toplayıcılığı vardı halk arasında.
98. Bu dağlarda yetişen otların içinde şifa veren ilaçlar vardır.
99. İşte, dünyanın her yanında insanlar vardı; burada, yabancısı olduğum bu kanallar kentinde de.

100. Henüz oraya varmadan önceki ara sokaklardan birinde, göstermek istediğim küçük bir bar vardı.
101. Parkta geçmiştekiler de var...
102. Eski aşıkları, kocaları yok aralarında.
103. Gözünde açık duman rengi gözlükler vardır.
104. Evde başka eşya da vardı.
105. Bunların tümü üzerinde neredeyse hiçbir görüş ayrılığı yok.
106. Bilim tarihinde, çok uzun ve zahmetli bir kolektif sürecin ürünü olan mantığı saymazsak, iki önemli teorik atılım dönemi vardır.
107. Bunun da temelinde insan var.
108. Alışveriş insanın kanında var.
109. Kadınların ve erkeklerin alışveriş şekillerinde ne gibi farklılık var?
110. Bu nedenle yatırımcının tetikte ve likit beklemesinde fayda var.
111. Artık dünyada böyle bir getiri hiçbir yerde yok.
112. Yirmi Ocak tarihli Milliyet'in spor sayfalarında, Anadolu turu köşesinde yer alan bir haberde hata var.
113. Arınç'ın sözleri teypte var.
114. Bu partiler arasındaki DYP'de ise gözle görülür bir hareketlilik var.
115. Şu anda hukuki problem yok.
116. Uçağın Kaptan Pilotu Alaattin Yunak'ın Gölcük'e bağlı Değirmendere beldesindeki baba evinde yas var.
117. Vücudunda yanık ve kırıklar var.
118. Bir kelepçede kaçak vardı.
119. Onlarda da Hıristiyanlık ve demokrasinin bir arada olup uyum içinde yaşaması var.
120. Şimdi sırada gerekli tedbir paketlerinin hazırlanması var.
121. Bisikletleriyle bu sokaktan çok sık geçen, geçerken de bu evin önünde zillerini ya da pilli düdüklerini öttüren delikanlılardan hiçbiri yok ortalıkta.
122. Arka bahçede kimse yoktu.
123. Evde yemek götürecek kimse yoktu.
124. Özellikle gözlerinde tanımı güç bir keder var.
125. Bunda ne anamın suçu var, ne babamın.

126. Doğada hepsinin birer barınağı var.
127. Çalıştığımız fabrikalarda, yararlı olmayan şey yoktur.
128. Sesimde kölesine emreden bir efendinin sesi vardı.
129. Hayatımda hiçbir teselli yoktu.
130. İnsanlara ortada bir cinayet olmadan bu duyguları yaşatabilecek tek şey vardı hayatta.
131. Canım 1956'da da var.
132. Bütün bunlar da yeğenlerde fazlasıyla var doğrusu.
133. Bu oğlana kalsa her zenginliğin altında biraz gözyaşı, hatta kan vardır.
134. Sevişmenin bir önünde, bir de ardında yalan vardır değil mi, biri kadına, ötekisi erkeğe kalan...
135. Poşet yok bu markette.
136. Çevrede in cin yok.
137. Hanın üst katında odalar da mı var?
138. Çay ocağının orada merdiven var.
139. Çevrede kimsecikler yoktu.
140. Mahmut Bey burada yok.
141. Üstünde bir atlet vardı.
142. O günlerde Anayasa düşkünlüğü vardı ülkede.
143. Aralarında rekabet var.
144. Çorum'da nohut bile yoktur.
145. İngilizcede, Eggplant var.
146. Ama abi bu evde iki prens var.
147. Yüzeyinde bazı karaltılar vardı.
148. İngiliz'in ne işi var orada?
149. Yemekte puf böreği var.
150. Ağzının kenarında var mıydı bu derin çizgi?
151. Bana dökmek var.
152. Sizlere anlatacaklarım var.
153. Oysa sizlere ne kadar çok anlatacağım vardı.
154. Dört okurumuza göre var.

155. On milyona şık tişörtler, pantolonlar, 30-40 milyona paltolar var.
156. Tamamlanmayana para yok.
157. Hiç kimsenin rejimin jandarmalığına soyunmasına gerek yok.
158. Yeni bilgilere, olaylara, süreçlere yer yoktur.
159. Kimsenin ona aldırdığı yoktu.
160. Artık o parktan bir çıkış yolu yoktur.
161. Bugüne kadar ifade ettiği hususların, akademik özgürlüklerle ilişkisi yok.
162. Bu yanıtın düşündüklerimle hiç ilgisi yoktu.
163. Benim kimseyle yarışım yok.
164. Ağbimle hiç yok...
165. Aslında sakızla teması yok.
166. Ne işin var kardeşim kız kurularıyla!
167. Kaybedecek neyim var?
168. Aldığım şeyin isminin önemi yoktu.
169. Hayatın boyunca bu korkuyu azaltma imkânın yok.
170. Aynı haltı yiyeceğini bilsen bile bir güvencen var.
171. Herhangi birine sırf anlatmakla hafifletebileceğimiz yüklerimiz yok.
172. Yahu, kızın ne kabahati var!
173. Ölen pilotun kardeşi de vardı.
174. Tek hedefleri vardı: askeri savcılığın elinde bulunan tahkikat dosyasını bulabilmek.
175. Maddiyatla ilgili zaten bir beklentisi yoktu.
176. İyi ki dedem var.
177. Yanlış düşünüyor olabilirim; ama böyle bir saplantım var.
178. Çaresiz bir hâli vardı.
179. Anlatacaklarım var.
180. Bir gerçeklik görünüşü var.
181. Aşıkları yok onun.
182. Pikabı vardı.
183. Daha bunun mezuniyeti, işsizliği, askerliği, Güneydoğu kaygısı var.
184. Nesi var?

185. Son olarak, bin yılın ünlü matematikçilerini ansiklopedik biçimde yansıtan bir çalışmamız da var.
186. İnsanların belli başlı eğilimleri, sınırları ve ihtiyaçları var.
187. Sorumluluk taşıyan insanlarımız var.
188. Partinin bu hale gelmesine neden olanların artık yönlendirme yapmaya hakkı yoktur.
189. Yeni turizm kentleri projemiz var.
190. Baskı politikamız yok
191. Hazine'nin bu hafta beş katrilyon lirası piyasaya olmak üzere toplam altı katrilyon lira tutarında iç borç geri ödemesi var.
192. Dalgası var.
193. Hepsinin var.
194. Kocaman bir kafası, iri iri elleri vardı.
195. Uçları yukarı kıvrık sipsivri bıyıkları vardı babasının.
196. Benim büyük amcam var ya.
197. Hele benim hiç yok.
198. Şu duruşun var.
199. Bunun üç nedeni var.
200. Evin, o eve özgü kokusu, mekânın sesi vardır.
201. İnsanın nesnel gerçekliği yoktur.
202. İçinde altın benekler olan bir eşine bir daha hiç rastlamadığım çok iri yeşil gözleri vardı.
203. Somurtkan dudakları, kocaman gözleri vardı.
204. Kimim kimsem yoktu.
205. Kendine sakladığı bir gizemi vardır.
206. Vizyonum var.
207. Çok sevdiğim bir mahalle arkadaşım var.
208. Erol'un İlhami adında bir ağbisi var.
209. Yalın bir kişiliği vardır.
210. Leblebi beyliğinin üç başkenti var: Çorum, Çankırı, Tavşanlı.
211. Ama beyaz peynirin ihtiyacı var.

212. Yani üçyüzaltmışbeş çeşit peynirleri varmış.
213. Bir keçi peynirleri vardır.
214. Bir de tuzsuz peynirlerimiz var: Lor, dil gibi.
215. Fransa'nın küflü peyniri Roquefort var.
216. Eskiden mevsimi vardı.
217. Tadı tuzu yok.
218. Ama annemin şartları vardı.
219. Bak mesela sadece bir tane annemiz var çünkü bir evde sadece bir tane kraliçe olabilir.
220. Bir kere çok değişik bir rengi var.
221. Şimdiye kadar hiç bir köpekte görmediğim yeşil gözleri, gözünün üstüne düşen alacalı tüyleri var.
222. Beni görür görmez kendini yere atıp göbeğini bir açışı var!
223. Neyin var Ali?
224. Bir şeyim yok.
225. Çocukları yok.
226. DUAL pikabım var.
227. Duşunuz da var.
228. Neniz var.
229. Ne uzun kirpikleri var, o yeşil...
230. Faranjiti var.
231. Pamuğun var.
232. Pasaklıyım var mı diyeceğin!

Appendix C

Partial Statistics of Sub-Corpus

Table C.1: Major parts of speech statistics of sub-corpus

Major Parts of Speech	**Count**	**%**
Adjective	219	12.815
Adverb	83	4.857
Conjunctive	66	3.862
Determiner	96	5.617
Interjection	2	0.117
Noun	640	37.449
Number	25	1.463
Postposition	19	1.112
Pronoun	64	3.745
Punctuation	306	17.905
Question Particle	6	0.351
Verb	183	10.708

Table C.2: Surface dependency statistics of sub-corpus

Syntactic Relation Names[29]	**Count**	**%**
Ablative Adjunct	7	0.484
Apposition	5	0.346
Classifier	65	4.498
Collocation	1	0.069
Coordination	64	4.429
Dative Adjunct	21	1.453
Determiner	91	6.298
Focus Particle	1	0.069
Instrumental Adjunct	5	0.346
Intensifier	29	2.007
Locative Adjunct	89	6.159
Modifier	323	22.353
Object	187	12.941
Possessor	65	4.498
Question Particle	10	0.692
Relativizer	1	0.069
Sentence Modifier	25	1.73
Sentence	243	16.817
Subject	202	13.979
Vocative	11	0.761

[29] For the clarification of these relations, the reader is referred to [OFL2003].