DISCOVERING THE DISCOURSE ROLE OF CONVERBS IN TURKISH DISCOURSE

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF INFORMATICS

OF

MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET FARUK ACAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

THE DEPARTMENT OF COGNITIVE SCIENCE

JANUARY 2014

Approval of the Graduate School of Informatics

____

Prof. Dr. Nazife Baykal

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

____

Prof. Dr. Cem Bozşahin

Head of Department

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

____

Prof. Dr. Deniz Zeyrek Bozşahin

Supervisor

Examining Committee Members

Prof. Dr. Cem Bozşahin (METU, COGS) ______________________

Prof. Dr. Deniz Zeyrek Bozşahin (METU, COGS) ______________________

Assist. Prof. Dr. Cengiz Acartürk (METU, COGS) ______________________

Dr. Ruket Çakıcı (METU, CENG) ______________________

Dr. Ceyhan Temürcü (METU, COGS) ______________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this wok.

Name, Last name: Ahmet Faruk Acar

Signature : _________________

ABSTRACT

DISCOVERING THE DISCOURSE ROLE OF CONVERBS IN TURKISH DISCOURSE

Acar, Ahmet Faruk

MS, Department of Cognitive Sciences

Supervisor: Prof. Dr. Deniz Zeyrek Bozşahin

January 2014, 78 Pages

The subordinate verb forms that occur in non-finite adverbial clauses are called converbs (Göksel & Kerslake , 2005). In Turkish, converbs can be discourse connectives as well as acting as the complement of a factive verb or an adverbial. We morphologically analyzed 15 converbs in Turkish Discourse Bank to find out possible morpho-syntactic features in order to distinguish different roles of these converbs. The aim of the study is to find out all possible roles of the converbs and the source of ambiguities as well as to find out beneficial features that may promote automatic methods to disambiguate the discourse role of the converbs, namely Simplex subordinators. For this purpose, we created a converb-corpus out of Turkish Discourse Bank. We conducted an annotation experiment with two annotators and examined the results. Also we trained a decision tree algorithm to see whether the morphological features of the right and left material of the converbs are indicative for the disambiguation task.

According to the annotation results, we observed three kinds of converbs: unambiguous converbs, which always create discourse relations; ambiguous converbs, which are ambiguous between a discourse connective and a non-discourse connective role; and hard cases, which are even more ambiguous, even for the human annotators. In addition to these, we saw that the syntactic features such as the syntactic class of the converb can be essential in automatic disambiguation studies. The distance between the converb and the matrix verb, and the morphological properties of the left and right edge of the converb seem to be good clues according to the machine learning experiment results.

Keywords: Discourse, Discourse Connective, Converbs, Disambiguation, Turkish

ÖZ

TÜRKÇE SÖYLEMDE ULAÇLARIN SÖYLEM ROLÜ

Acar, Ahmet Faruk

Yüksek Lisans, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Prof. Dr. Deniz Zeyrek Bozşahin

Ocak 2014, 78 Sayfa

Ulaçlar, çekimsiz belirteç tümceciklerinde de yer alan yana sıralamalı eylemsilerdir (Göksel & Kerslake , 2005). Türkçe'de ulaçlar, söylem bağlacı olabildiği gibi bir eylemin tamlayanı ya da belirteci gibi de davranabilir. Ulaçların farklı rollerini bulmak ve ayırt edebilmek için, Türkçe Söylem Bankasında yer alan 15 ulacı biçimbilimsel olarak inceledik. Çalışmanın amacı, ulaçların kullanılabildiği tüm farklı rolleri bulmak, anlam belirsizliğinin sebeplerini anlamak ve otomatik işaretleme yöntemlerine faydalı olabilecek özellikleri keşfetmektir. Bu amaçla, bir ulaç derlemi oluşturduk. İki işaretleyici ile işaretleme çalışması başlattık ve sonuçları inceledik. Ayrıca, karar ağacı algoritmasını eğiterek ulaçların sağ ve solundaki biçimbilimsel özelliklerin anlam belirsizliğini çözmede bilgilendirici olup olmadığına baktık.

İşaretleme sonuçlarına göre üç çeşit ulaç gözlemledik: anlam belirsizliği taşımayan ve her zaman söylem ilişkisi oluşturan ulaçlar; anlam belirsizliği taşıyan, söylem bağlacı ve diğer rolleri alabilen ulaçlar; ve işaretleyicilere bile anlamı belirsiz gelen zor ulaçlar. Bunlara ilaveten; ulacın bağlı olduğu sözdizimsel sınıf gibi özelliklerin, ulacın rolünü otomatik belirleme faydalı olabileceğini gördük. Ulaç ile ana eylem arasındaki mesafe ve ulacın sağ ve solundaki kelimelerin biçimbilimsel özelliklerinin de makine öğrenme çalışmaları için iyi ipucu niteliği taşıdığını gözlemledik.

Anahtar Kelimeler: Söylem, Söylem Bağlacı, Ulaç, Anlam Belirsizliği, Türkçe

ACKNOWLEDGMENTS

First of all I would like to thank my supervisor Prof. Dr. Deniz Zeyrek for her guidance and tolerance. I would like to thank Çağrı Çöltekin and Haşim Sak for their technical support with their morphological parsers. I need to thank Işın Demirşahin exclusively for helping me with annotations and answering all my trivial questions with great understanding.

I am grateful to my family and Ayça Moğol for their endless support.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

D-LTAG:	Lexicalized Tree Adjoining Grammar for Discourse
PDTB:	Penn Discourse Treebank
TDB:	Turkish Discourse Bank
MTC:	METU Turkish Corpus
DC:	Discourse Connective
NDC:	Non Discourse Connective
↓:	placed in front of any form that is confined to very informal contexts

CHAPTER 1

Introduction

Turkish Discourse Bank (TDB) recognizes explicit discourse connectives from three grammatical types: simple and paired coordinating conjunctions, simple and complex subordinating conjunctions, and discourse adverbials (Zeyrek & Webber, 2008). In the first release of TDB, coordinating conjunctions, complex subordinators and discourse adverbials are annotated. This study aims to provide a preliminary analysis of the converbs that act as simplex subordinators and build a base for the annotation of the simplex subordinators in order to enrich TDB.

Converbs in Turkish are ambiguous with respect to their discourse vs. non-discourse uses. For example, the factive nominalization with ablative inflection builds the highly frequent suffix group –DIğIndAn, which can be the complement of a verb, the complement of a complex subordinator, or can be a simplex subordinator by itself. Such cases are easily disambiguated by native speakers because of the unambiguous syntactic context. However, some other uses of converbs are ambiguous even for native speaker human annotators because of their capability of creating diverse lexical items such as idioms, collocations, fixed expressions etc. In this thesis, instances of 15 converbs from TDB are annotated by native speakers in terms of their discourse or non-discourse uses.

This thesis aims to investigate following questions:

- What are the possible roles/ambiguity cases for each converb?
- What are the syntactic/semantic features that differentiate discourse role from non-discourse role of each converb?
- What are the common/specific features of converbs in terms of their ambiguity?
- What morphological, syntactic and semantic features are available that will promote automatic annotation of converbs?

Considering the given questions, this thesis will be organized as follows:

Firstly, there will be a section that covers the necessary background knowledge. This section includes three chapters: chapter (2) is an introduction to the discourse studies and discourse structures; where D-LTAG is introduced, and Penn Discourse Treebank is discussed in detail. Especially the connective types and the sense hierarchy of PDTB are essential to this thesis, because realization of discourse connectives and senses in TDB are primarily based on them. In addition, with the latest studies about TDB and the most recent updates are provided, since they are the background for this study. In chapter (4), necessary syntactic and semantic explanations are given. Turkish subordinate clauses and converbs are explained in detail, then terminology from lexical semantics, such as ambiguity, compositionality, and conventionality, is introduced. These terminologies are essential in order to make fine-grained distinctions, especially for different roles of highly ambiguous converbs.

The nest section, Methodology, explains the procedures applied in the thesis. Manual annotation procedure is explained with all preliminary works including the selection of converbs, the guideline and the tag set used to label the converbs. Additionally, the initial plan of the thesis and the reasons to change the thesis' focus is explained at the beginning of the chapter.

The results of the annotations are given in the Result section. Essential questions of the thesis are responded by looking at the annotation results and by interpreting the annotation agreement statistics.

The Conclusion chapter provides the summary of the findings, the possible implications of them and the discussions about what can be done more as future work.

CHAPTER 2

Discourse Studies

A variety of linguistic fields, such as Natural Language Processing, Speech Recognition and Theoretical Linguistics, use large corpora as a source to extract information about language (Marcus, Santorini, & Marcinkiewicz, 1993).

Before delving into specific corpus and discourse subjects, it's better to introduce discourse structures and D-LTAG in particular, as a starting point, since they are the theoretical background of the PDTB.

2.1 D-LTAG

Lexicalized Tree Adjoining Grammar for Discourse (D-LTAG) is an extended version of L-TAG for discourse level (Webber, 2004). According to Lexicalized Tree-Adjoining Grammar (LTAG) each word is associated with a set of tree structures, where the word can appear in one of the *minimal syntactic constructions*. Sample tree structures for the verb *like* are given in Figure 2-1. In LTAG, there are two kinds of tree structures that can be in the tree sets: *initial trees* reflect basic functor-argument dependencies and *auxiliary trees* introduce recursion (Webber, 2004). In Figure 2-1; the trees (a), (b), (c) are examples of initial trees, while (d) and (e) are auxiliary trees. The special symbols used in these trees (↓ and *) relate to the two operations: ↓ indicates a substitution site where an elementary tree can substitute into a derived tree, provided the label at its root matches that of the substitution site; * indicates an adjunction site (or foot node), where an auxiliary tree can adjoin into a root (Webber, 2004).

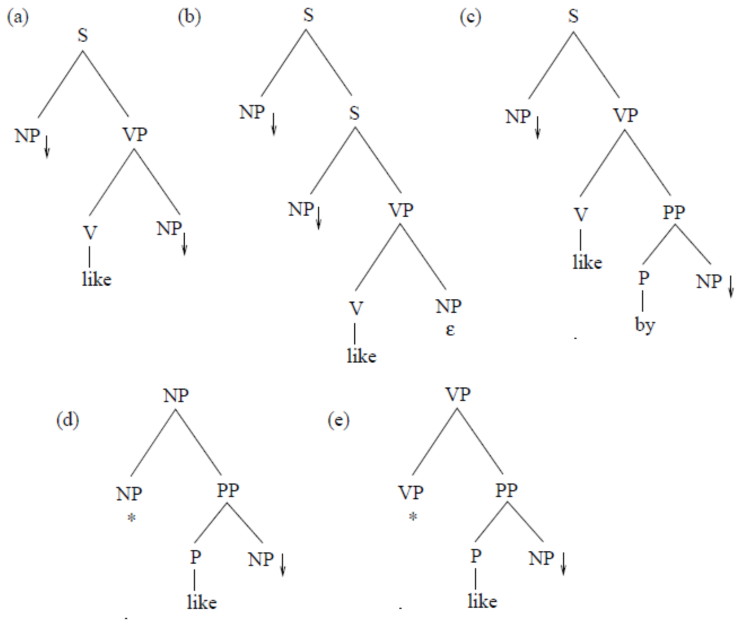


Figure 2-1 Elements of the tree set of like (Webber, 2004, p. 5)

In D-LTAG, low level discourse structure is represented by trees which are anchored by lexico-syntactic items that signify discourse relations such as discourse connectives. Discourse connectives act as predicates, similar to verbs at clausal level (Webber & Joshi, 1998). The arguments of the discourse connectives are text spans that can be interpreted as abstract objects like propositions, facts, descriptions, situations, or eventualities (Asher, 1993). The hierarchy of abstract objects is given in Figure 2-2. To sum up, abstract objects are building blocks of discourse and D-LTAG is the way of building it.

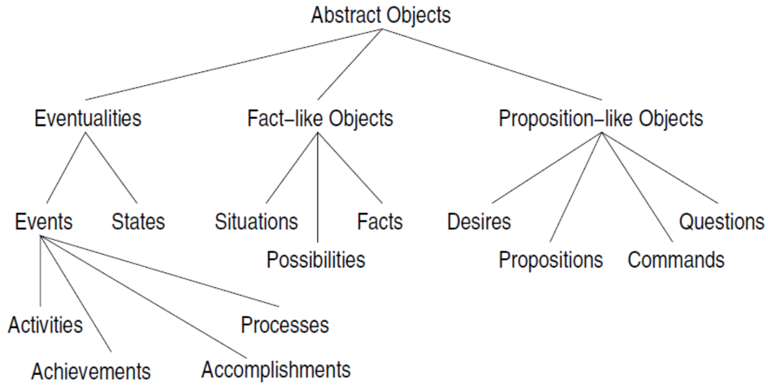


Figure 2-2: Hierarchy of Abstract Objects (Asher, 1993)

2.2 Penn Discourse Treebank

Penn Treebank (PTB) is one of the largest corpora which is annotated for part-of-speech information and contains more than 4.5 million words from Wall Street Journal (Marcus, Santorini, & Marcinkiewicz, 1993). However, developing Natural Language Processing applications require a richer annotation (Kingsbury & Palmer, 2002). For this reason, a part of PTB with over 1 million words from Wall Street Journal (WSJ) was annotated for their discourse relations and arguments; thus The Penn Discourse Treebank (PDTB) was created.

The framework of discourse annotation depends on the theoretical work, whose underlying principles can be found in D-LTAG (Webber, 2004). PDTB supports the extraction of useful features related to syntax, semantics and discourse at the same time since it was built on PTB, which already had sentence level syntactic annotations, and Propbank, which had predicate-argument structure annotation (Prasad, et al., 2008). In addition to the annotation of discourse relations, The Penn Discourse Treebank provides sense annotations for discourse connectives, which can have more than one meaning just like verbs (Miltsakaki, Robaldo, Lee, & Joshi, 2008). PDTB is argued to be theory independent and as a result it can be used by any linguist as a means for their studies. Four fundamental benefits of PDTB are listed as below (Webber, et al., 2005):

1. *It clearly defines discourse structure, which is theory-neutral and useful for researches from different frameworks. Additionally, it can be used as a resource to validate existing theories.*
2. *Since it supports the observation of syntactic and discourse annotation at the same time, researchers can examine the relationship of syntactic structure and discourse structure easily as well as the relationship between clausal and discourse-level semantics.*
3. *It can serve as a basis for more complex NLP tasks such as Machine Translation, Question Answering and Natural Language Generation.*
4. *It can help the development of automatic procedures to identify discourse connectives and their arguments.*

2.2.1 Discourse Connectives in PDTB

The discourse connectives in PDTB are divided into two main categories according to their realization in the corpus. The first type of discourse relations is *explicit relations*. Discourse relations are explicit when they are signaled directly by an appropriate discourse connective. In this case, the arguments of explicit connectives are unconstrained in terms of their location, and can be referred anaphorically. The second type of discourse relations are called *implicit relations,* which exist between two adjacent sentences in the absence of an *explicit connective* (Prasad, et al., 2008).

Regardless of the type of the connective, a discourse connective can take only two arguments which are simply called as **Arg1** and **Arg2**. Arg2 is the argument which is syntactically bound to connective and Arg1 is the other argument. In order to represent the discourse relations consistently and their arguments clearly, the <u>connectives</u> are underlined, *Arg1* is given in italics and **Arg2** is written in bold face in PDTB and TDB publications. We will follow this convention throughout the thesis.

2.2.1.1 Explicit Connectives

Explicit discourse connectives are identified from certain syntactic classes: *subordinating conjunctions, coordinating conjunctions* and *discourse adverbials*:

Subordinating conjunctions connect subordinate clauses to the main clause. They usually express temporal, causal, purpose, concessive and conditional relations. *Although* in (1) is an example of a subordinating conjunction.

(1) "Michelle lives in a hotel room, and <u>although</u> **she drives a canary-colored Porsche**, *she hasn't time to clean or repair it.*" (Prasad, et al., 2008, p. 2)

Coordinating conjunctions connect two independent clauses and include the highly frequent connectives *and*, *but*, *or* etc. *But* in (2) is an example of a coordinating conjunction.

(2) *"The House has voted to raise the ceiling to $3.1 trillion*, but **the Senate isn't expected to act until next week at the earliest."**

Adverbial connectives include adverbs like *however*, *therefore*, and *then,* which modify the sentence and express the discourse relation between two events or states. Prepositional phrases such as *as a result*, *in addition*, and *in fact* are also included in this class since they show similar relations. *As a result* in (3) is an example of a discourse adverbial.

(3) "*...many analysts expected energy prices to rise at the consumer level too*. As a result, **many economists were expecting the consumer price index to increase significantly more than it did**]" (Miltsakaki, Prasad, Joshi, & Webber, 2004, p. 3)

2.2.1.2 Implicit connectives

Implicit connectives are identified between adjacent sentences which are not explicitly anchored with any discourse connective from the syntactic groups above. Therefore, annotation of implicit connectives consists of inserting an appropriate connective between these adjacent sentences that describes the inferred relation best. In (4) the implicit connective *because* is inserted between the two sentences that are inferred to have a causal relation but lack an explicit connective to convey that relation.

(4) *To compare temperatures over the past 10,000 years, researchers analyzed the changes in concentrations of two forms of oxygen.* (Implicit=because) **These measurements can indicate temperature changes,** … (Contingency:Cause:reason) (Prasad, Joshi, & Webber, 2010, p. 3)

However, there are situations where annotators cannot find any appropriate connective to insert between adjacent sentences. In such cases, three distinct labels are used: *EntRel* label is used for an entity-based coherence relation, in which the second sentence seems to continue the description of some entity mentioned in the first; *NoRel* is used if there is no relation between adjacent units; and *AltLex*, stands for *Alternative Lexicalization*, whose instances are annotated if the following conditions are held (Prasad, Joshi, & Webber, 2010):

1. *A discourse relation can be inferred between adjacent sentences.*
2. *There is no explicit connective present to relate them.*
3. *The annotator is not able to insert an implicit connective to express the inferred relation (having used "NONE" instead), because inserting it leads to an awkward redundancy in expressing the relation.*

Further analysis of *AltLex* annotations shows that Discourse Relation Markers (DRMs) are a lexically open-ended class of elements which may or may not belong to well-defined syntactic classes such as conjunctions, prepositional phrases, subordinators etc. For example, Example (5) was annotated as *AltLex* because inserting a connective like *because* result redundancy in discourse relation. The phrase *One reason is* is taken to denote the relation and is marked as *AltLex*.

(5) *Now, GM appears to be stepping up the pace of its factory consolidation to get in shape for the 1990s.* **One reason is mounting competition from new**

Japanese car plants in the U.S. that are pouring out more than one million vehicles a year at costs lower than GM can match. (Contingency:Cause:reason) (Prasad, Joshi, & Webber, 2010, p. 3)

Examples for *EntRel* and *NoRel* are given in Examples (6) and (7) below (Prasad, et al., 2008):

(6) "*Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern*. EntRel **Mr. Milgrim succeeds David Berman, who resigned last month."** (p. 23)

(7) "*Jacobs is an international engineering and construction concern*. NoRel **Total capital investment at the site could be as much as $400 million, according to Intel.**" (p. 25)

PDTB takes all subordinating conjunctions, coordinating conjunctions, certain adverbials and implicit connectives as discourse connectives. Prasad et al., in their recent paper, argue that placing such syntactic and lexical restrictions on Discourse Relation Markers provides a full understanding of discourse relations, *Alternative Lexicalization* in particular, since they can be realized in other ways as well (Prasad, Joshi, & Webber, 2010).

2.2.2 Discourse Arguments and Minimality Principle

Another important issue related to discourse relations is *what counts as arguments* and *how much an argument extends within the discourse*. Because of the fact that the discourse relations hold between abstract objects, an argument should contains at least one predicate along with its arguments, and of course, a sequence of clauses or sentences may also form a legal argument that comprise multiple predicates (Miltsakaki, Prasad, Joshi, & Webber, 2004). Yet there are exceptions: nominal phrases which express an event or a state; and discourse deictics that resolve to an abstract object can also be interpreted as abstract objects. In Example (8), for instance, *that* denotes the interpretation of the sentence immediately preceding it.

(8) Airline stocks typically sell at a discount of about one-third to the stock market's price-earnings ratio – which is currently about 13 times earnings. [*That's*] because [**airline earnings, like those of auto makers, have been subject to the cyclical ups-and-downs of the economy**]. (Miltsakaki, Prasad, Joshi, & Webber, 2004, p. 4)

In order to determine the location and the extent of the arguments, the *minimality principle* was introduced. The principle requires arguments to contain minimal and sufficient amount of information in order to interpret the discourse relation properly (Prasad, et al., 2008). **Table 2-1** shows the distribution of the location and extent of Arg1 among the Explicit connectives (Prasad, et al., 2008):

	SingleFull	SinglePartial	MultFull	MultPartial	Total
SS	0	11224	0	12	11236
IPS	3192	1880	370	107	5549
NAPS	993	551	71	51	1666
FS	2	0	1	5	8
Total	4187	13655	442	175	18459

Table 2-1 Distribution of the location (rows) and extent (columns) of Arg1 of Explicit connectives. SS = same sentence as the connective; IPS = immediately previous sentence; NAPS = non-adjacent previous sentence; FS = some sentence following the sentence contain **(Prasad, et al., 2008, p. 3)**

There are 40600 annotated relations in the second version of PDTB and explicit connectives include 100 different relation types, in which modified forms of a connective are counted as one. There are also 102 types of implicit connectives in total. Table 2-2 illustrates the distribution of connectives in the PDTB 2.0 (Prasad, et al., 2008):

PDTB Relations	No. Of Tokens
Explicit	18459
Implicit	16224
AltLex	624
EntRel	5210
NoRel	254
Total	40600

Table 2-2: Distribution of Relations in PDTB-2.0 (Prasad, et al., 2008, p. 3)

2.2.3 Sense Annotation

Sense annotation is included in the second version of PDTB for the explicit connectives, implicit connectives and *AltLex* relations. It is accomplished by adding new features to the discourse connectives on PDTB rather than building a new standoff annotated version of it. The purpose of giving sense tags to connectives is to provide a semantic description of the relation between arguments (Prasad, et al., 2008).

The tag set for the sense annotation is hierarchically organized into classes and each class contains types and subtypes as shown in the Figure 2-3. Such a hierarchical sense organization has benefits. For example, it allows the annotators to select a suitable tag and thus maintain inter-annotator reliability. Also, the annotators can make inferences at any level where they are comfortable, namely it doesn’t force the annotators to make fine selections between distinct senses. Besides, the hierarchical organization of the senses shows that a very small number of relations may exist between arguments. This small set of relations is represented in the *Class* level and the *Types* and *Subtypes* inherit the primary meaning of their parents.

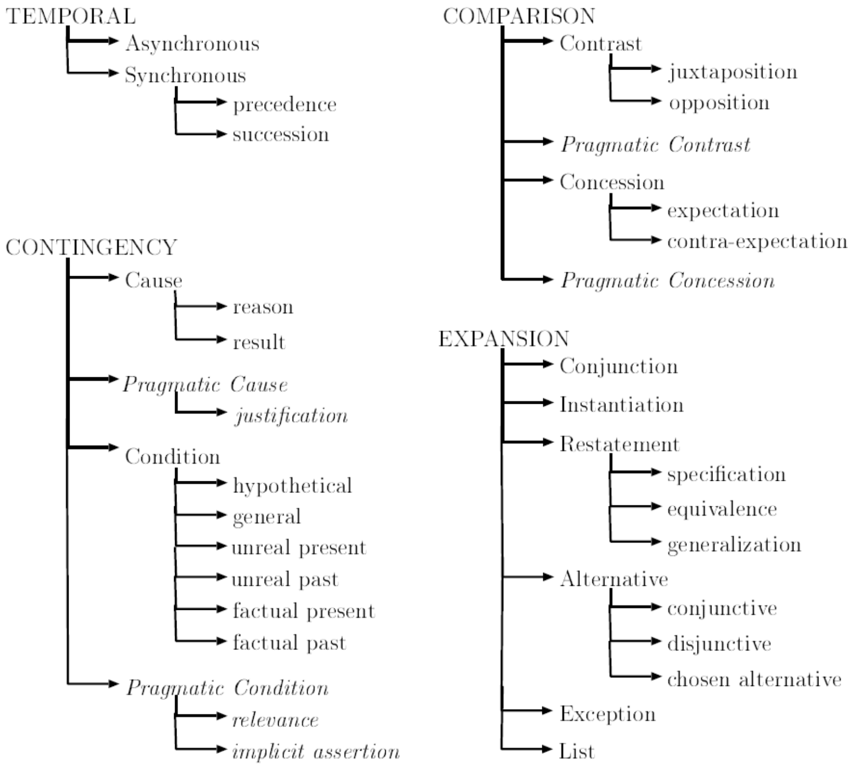


Figure 2-3: Hierarchy of sense tags (Prasad, et al., 2008, p. 5)

It is apparent that the agreement between annotators in the sense annotation will be higher at the Class level and will be lower in the Subtype level. Distribution of inter-annotator agreement (Table **2-3**) and distribution of Class sense tags (Table 2-4) is given below (Prasad, et al., 2008):

LEVEL	% AGREEMENT
CLASS	94%
TYPE	84%
SUBTYPE	80%

Table 2-3 Inter-annotator agreement (Prasad, et al., 2008, p. 5)

"CLASS"	**Explicit (18459)**	**Implicit (16224)**	**AltLex (624)**	**Total**
"TEMPORAL"	3612	950	88	4650
"CONTINGENCY"	3581	4185	276	8042
"COMPARISON"	5516	2832	46	8394
"EXPANSION"	6424	8861	221	15506
Total	19133	16828	634	36592

Table 2-4 Distribution of "CLASS" sense tags (Prasad, et al., 2008, p. 6)

2.3 Turkish Discourse Studies

2.3.1 Metu Turkish Corpus (MTC)

A corpus is a large, usually computerized, database of spoken and/or written texts of a language, which allows for *searching for, retrieving, sorting and calculating linguistic data* (McEnery & Wilson, 1996). A corpus is expected to be representative of its language and can be used for building hypotheses and making generalizations for the language it represents (Tognini-Bonelli, 2001).

METU Turkish Corpus (MTC) is a natural written language source of 2 million words from multiple genres (Say, Zeyrek, Oflazer, & Özge, 2002), and contains texts written between 1991 and 2000. The genres in the corpus include novels, short stories, essays, research monographs, interviews, memoirs and news. All the Turkish example sentences in this thesis are from the MTC, unless stated otherwise.

2.3.2 The Turkish Discourse Bank (TDB)

The Turkish Discourse Bank (TDB) aims to annotate MTC sentences for the discourse connectives and their arguments in order to build a discourse level resource for Turkish (Zeyrek, et al., 2009). TDB follows the principles of PDTB for annotating discourse connectives and their arguments of PDTB with some differences. For instance, TDB aims to annotate only explicit connectives for the time being, and the annotation of implicit connectives remains as future work (Zeyrek & Webber, 2008).

In Turkish, discourse connectives are realized from three syntactic categories which can be further analyzed into five classes (Zeyrek & Webber, 2008):

Simple coordinating conjunctions: Coordinating conjunctions combine two clauses of the same syntactic type. Turkish coordinating conjunctions are single lexical items such as *çünkü* 'because', *ama* 'but,' *ve* 'and', and the particle *dA*.

(9) *Yapılarını kerpiçten yapıyorlar, ama sonra tası kullanmayı öğreniyorlar. Mimarlık açısından çok önemli,* çünkü **bu yapı malzemesini başka bir malzemeyle beraber kullanmayı, ilk defa burada görüyoruz.**

'They constructed their buildings first from mud bricks but then they learnt to use the stone. Architecturally, this is very important because **we see the use of this construction material with another one at this site for the first time.'** (Zeyrek & Webber, 2008, p. 67)

Paired coordinating conjunctions: Paired coordinating conjunctions are composed of two lexical items such as *hem... hem* 'both… and,' and *ne... ne* 'neither… nor' which link two clauses. Example (10) shows the usage of ya .. ya 'either .. or'.

(10) *Birilerinin* ya *işi vardır, aceleyle yürürler*, ya **koşarlar**.
'*Some people are* either *busy and walk hurriedly*, or **they run**.' (Zeyrek & Webber, 2008, p. 67)

Simplex subordinators: Simplex subordinators are mostly converbs, i.e., suffixes forming non-finite adverbial clauses, such as *–(y)ken*, 'while' and *-(y)ArAk* 'by means of'. Since they are the main subject of the thesis, they will be examined in detail in 3.2.

(11) Kafiye Hanım beni kucakladı, **yanağını yanağıma sürt**erek *iyi yolculuklar diledi*.

‘Kafiye hugged me and **by rubbing her cheek against mine**, *she wished me a good trip.*’ (Zeyrek & Webber, 2008, p. 67)

Complex subordinators: Complex subordinators consist of two parts, usually a postposition (*rağmen* ‘despite’, *için* ‘for’, *gibi* ‘as well as’) and an accompanying suffix on the non-finite verb of the subordinate clause.

(12) **Herkes çoktan pazara çıkTIĞI** için *kentin o dar, eğri büğrü arka sokaklarını boşalmış ve sessiz bulurduk.*

‘Since **everyone has gone to the bazaar long time ago**, *we found the narrow and curved back streets of the town empty and quiet.*’

(13) **[Turhan Baytop] Paris Eczacılık Fakültesi Farmakognozi kürsüsünde görgü ve bilgisini arttırMAK** için *çalışmıştır.*

‘**Turhan Baytop** *worked* **at Paris Pharmacology Faculty** so as to **increase his experience and knowledge**,’ (Zeyrek & Webber, 2008, p. 67)

Anaphoric connectives: Anaphoric connectives require only one abstract object syntactically, and they retrieve the other argument anaphorically from the previous discourse. *Ne var ki* ‘however’, *üstelik* ‘what is more’, *ayrıca* ‘apart from this’*, ilk olarak* ‘firstly’, etc. are some examples of Turkish anaphoric connectives.

(14) *Ali hiç spor yapmaz.* Sonuç olarak çok istediği halde **kilo veremiyor**.

‘*Ali never exercises*. Consequently, **he can’t lose weight** although he wants to very much.’

(15) Zeynep önceleri Bodrum’da oturdu. *Krediyle deniz kenarında bir ev aldı.* Evi dayadı, döşedi, bahçeye yasemin ekti. Ne var ki **banka kredisini ödeyemediğinden evi satmak zorunda kaldı.**

‘Zeynep first lived in Mersin. *She bought a house by the sea on credit.* She furnished it fully and planted jasmine in the garden. However, **she had to sell the house because she couldn’t pay back the credit.’** (Zeyrek & Webber, 2008, p. 68)

A list of the discourse connectives from TDB is given in Appendix B.

.

CHAPTER 3

Subordinate Clauses, Converbs and Lexical Semantics

This chapter aims to explain the essential linguistic terms used in the thesis. The chapter consists of three subsections: firstly, there will be an overview of the subordinate clauses in Turkish, since they are the primary context for the subordinate conjunctions; secondly, the converbs will be explained and differentiated from other clausal types; and finally, some terminology from lexical semantics, which is used to discriminate the different roles of the converbs, will be introduced.

3.1 Subordinate Clauses

3.1.1 Subordinate Clauses in Turkish

Like many languages, Turkish has simple sentences that hold only a main clause (16), and complex sentences that have a main clause and one or more subordinate clauses (17) (Göksel & Kerslake , 2005):

(16) Dün okullar açıldı.
'The schools opened yesterday.'

(17) Dün [yolda giderken] [yıllardır görmediğim] bir arkadaşıma rastladım.
'Yesterday, [as I was walking along the street], I ran into a friend [whom I hadn't seen for years].' (p. 109)

The predicate of a subordinate clause can be finite like the predicate of a main clause (18):

(18) [Maç birazdan başla-yacak] de-n-iyor.
match soon start-FUT say-PASS-ıMPF
'It is said [that the match will be starting soon].' (p. 123)

However, subordinate clauses are formed with non-finite predicates most of the time, meaning that their predicate contains one of the subordinating suffixes (19).

(19) [Maç-ın birazdan başla-yacağ-ı] söyleniyor.
match-GEN soon start-SUB-3SG.POSS
'It is said [that the match will be starting soon].'

3.1.2 Types of Subordinate Clauses

In Turkish, subordinate clauses are created by subordinate suffixes, which are nominalizing suffixes. These suffixes combine with verb stems, and can be inflected with the plural suffix, the possessive markers, or a case suffix to form non-finite verb forms. Subordinate clauses are of three types according to their function in a sentence (Göksel & Kerslake , 2005):

i. **Verbal nouns:** These are the non-finite verbs of **noun clauses,** which function as the subject or object in the sentences (20).

(20) [Sorun **yarat-*acağ*-ı**] belli. (Verbal noun)
problem create-VN-3SG.POSS clear
'It is clear [*that s/he will create* problems].' (pp. 84-85)

ii. **Participles:** These are the non-finite verbs of **relative clauses,** which function as adjectival phrases.

(21) [Sorun **yarat-*an***] kuruluş-lar uyar-ıl-dı. (Participle)
problem create-PART organization-PL admonish-PASS-PF
'The organizations [*that were creating* problems] were admonished.' (p. 85)

iii. **Converbs:** These are the non-finite verbs of **adverbial clauses,** which function as adverbials.

(22) [Sorun **yarat-*maktansa***] sonuç-lar-ı kabullen-di. (Converb)
problem create-CV consequence-PL-ACC accept-PF
'[*Instead of creating problems*] s/he accepted the consequences.' (p. 85)

The majority of the subordinating suffixes in Turkish form only one of the three types of the non-finite verbs. However, certain subordinators, namely *-DIK, -(y)AcAK, -mA* and *–mAK* can form more than one type of subordinate clause. In some cases they do this by combining with other suffixes or postpositions.

3.1.3 Subordinate Suffixes

3.1.3.1 -DIK and -(y)AcAK

-DIK and *-(y)AcAK* form all three types of subordinate clauses when they combine with following possessive suffixes and case suffixes. They can be followed by all of the nominal inflectional suffixes when they function as participles in headless relative clauses including the plural marker, *sat**tık**larımınki* 'the one belonging to those that I sell/sold' (Göksel & Kerslake , 2005).

-DIK suffix typically expresses present or past time and it forms:

(i) **Verbal nouns:** *git**ti**ğini (bil-)* '(know) that s/he has left', *kıskandır**dığ**ınızı (anla-)* '(understand) that you are making/have made [s.o.] envious'.
(ii) **Participles:** *göreme**diğ**im (film)* '(the film) that I was not able to see', *öpüş**tüğ**ü (kız)* '(the girl) whom s/he has kissed/is kissing'
(iii) **Converbs:** *bak**tığ**ımızda* 'when we look/looked', *anla**dığ**ımdan* 'because I understand/(have) understood'. (p. 85)

-DIK has a converbial function with the following suffixes and postpositions:

-DIğIndA:	*yürü**düğ**ümde*'when I walk'
-DIkçA:	*koş**tuk**ça* 'the more [s.o.] runs'
-DIğIndAn (beri/dolayı/ötürü):	*gel**diğ**imizden beri* 'since we arrived'
-DIğI (için/zaman/sırada/anda/halde/kadar ıyla/takdirde/gibi/sürece/ nispette):	*bakma**dığ**ım için* 'because I haven't looked/am/was not looking', *gör**düğ**üm anda* 'the moment I saw [it]'
-DIğInA (göre):	*isteme**diğ**inize göre* 'since you don't/didn't want [it]'
-DIktAn (sonra/başka):	*al**dık**tan sonra* 'after taking [it]', *anla**dık**tan başka* 'in addition to understanding'

Table 3-1 Converbial function of *–DIK* (pp. 85-86)

The subordinator *-(y)AcAK* designates (relative) future time, and forms noun clauses, relative clauses, and adverbial clauses:

- **(i)** **Verbal nouns:** anlayacağımı (san-) '(imagine) that I would understand', iteceğini (düşün-) '(think) that s/he would push'.
- **(ii)** **Participles:** okuyacağım (kitap) '(the book) that I am/was going to read', sevemeyeceğim (bir kişi) '(a person) that I shall/would not be able to like', görüşeceği (doktor) '(the doctor) whom s/he is/was going to see'.
- **(iii)** **Converbs:** öğreneceğine 'instead of learning', isteyeceğimden 'because I am going to want'. (p. 86)

-(y)AcAK has a converbial function when it occurs in one of the following combinations, some of which involve postpositions.

-(y)AcAğI (için/zaman/sırada/anda/halde/gibi):	*kalk**acağ**ın zaman* 'when you are going to get up', *oturma**yacağ**ı için* 'because s/he isn't/wasn't going to stay', *gid**eceğ**i gibi* 'in addition to the fact that s/he is/was going to go'
-(y)AcAğIndAn (dolayı/ötürü):	*satma**yacağ**ından ötürü* 'on account of the fact that s/he is/was not going to sell [it]'
-(y)AcAğInA (göre):	*içme**yeceğ**ime göre* 'since I'm/I was not going to drink [it]'
-(y)AcAk (kadar/derecede):	*sakla**yacak** kadar* 'to the point of hiding [it]'

Table 3-2 Converbial function of *–(y)AcAk* (p. 86)

3.1.3.2 -mA and -mAK

Both *-mA* and *-mAK* create verbal nouns and converbs. These two suffixes differ with respect to which nominal inflectional markers they can combine with (Göksel & Kerslake , 2005). For instance, while *-mA* is often followed by one of the possessive markers, *-mAK* cannot combine with them; or only *-mA takes* the plural suffix.

*git**me**nizi (bekliyor)* 's/he expects you to leave'	*git**mey**i (bekliyor)* 's/he expects to leave'
*şarkı söyle**me**ne (bayılıyor)* 's/he loves [the way] you sing'	*şar kı söyle**mey**e (bayılıyor)* 's/he loves singing'
*koş**ma**mda (ısrar etti)* 's/he insisted that I run/ran'	*koş**mak**ta (ısrar etti)* 's/he insisted on running'
*konuş**ma**mdan (korkuyor)* 's/he is scared that I might talk'	*konuş**mak**tan (korkuyor)* 's/he is scared of talking'

Table 3-3 Combinability of –mA and –mAk with suffixes (p. 87)

-mAK subordinator creates noun clauses and adverbial clauses:

(i) **Verbal nouns:** *al**mak** (iste-)* '(want) to buy', *sev**mey**i (öğren-)* '(learn) to love', *ağla**ma**ya (başla-)* '(start) crying'

(ii) **Converbs:** *iç**meks**izin* 'without drinking'. (p. 87)

-mAK has an adverbial function when it occurs with the following suffixes and postpositions:

-mAk (üzere/için/yerine/suretiyle/şartıyla):	*ver**mek** için* 'in order to give'
-mAklA (birlikte):	*okuyabil**mek**le birlikte* 'although able to read'
-mAksIzIn (formal):	*dön**meksizin*** 'without returning'
-mAktAn (öte/başka/gayrı):	*sat**mak**tan öte* 'apart from selling [it]'
-mAktAnsA:	*bitir**mek**tense* 'rather than finishing [it]'.

Table 3-4 Adverbial functions of *–mAk*

-mA subordinator creates noun clauses and adverbial clauses:

(i) **Verbal nouns:** *anlama**ma**mı (iste-)* '(want) me not to understand'

(ii) **Converbs:** *yürü**mek**ten başka* 'apart from walking'. (p. 88)

-mA has an adverbial function when it occurs with the following suffixes and postpositions:

-mAsI (için/halinde/durumunda/yüzünden):	*öksür**me**si halinde* 'in the event of his/her coughing'
-mAsIndAn (itibaren/önce/sonra/ötürü/başka/dolayı):	*seçil**me**sinden önce* 'before s/he was elected', *isteme**me**mizden ötürü* 'because we don't/didn't want [it]'
-mAsInA (rağmen/karşın):	*anlaş**ma**nıza rağmen* 'in spite of your getting along well together'.

Table 3-5 Adverbial function of *-mA* (p. 88)

3.1.3.3 *-(y)An and -(y)Iş*

-(y)An: This subordinator suffix creates only relative clauses such as *okuy**an** (çocuk)* '(the child) who studies/is studying'. Much less productively, *-(y)An* can be used idiomatically in informal contexts to express the unexpectedly large number of people involved in a particular activity. In these cases, it is reiterated on identical and adjacent verb stems where the second verb has dative case marking, for example, *Konsere gid**en** gid**ene*** 'Masses of people went to the concert', *Şu saçma dergiyi de al**an** al**ana**!* 'Everyone's buying this ridiculous magazine!' (Göksel & Kerslake , 2005).

-(y)Iş: This subordinator suffix can combine with the plural marker, possessive suffixes and case suffixes and creates verbal nouns, for example, *otur**uş**umu (beğen-)* '(like) my way of sitting', *konuş**uş**unuz* 'the way you talk' (Göksel & Kerslake , 2005).

3.1.3.4 Other suffixes that form converbs

Other suffixes that create converbs with some of the suffixes and postpositions given in Table 3-6:

-(y)IncA	*yüz**ünce*** 'when [s.o.] swims/swam', *kalkma**yınca*** 'when [s.o.] doesn't/didn't get up'.
-(y)ArAk	*koş**arak*** 'running', *büyü**yerek*** 'growing up', *çalış**arak*** 'by working'. Also ↓*-(y)ArAktAn:bakaraktan* 'looking'.
↓*-(y)AlI (beri)*	*düşü**neli** (beri)* 'since thinking about [s.t.]', *ge**leli** beri* 'since

	arriving', 'since [s.o.] arrived'. Colloquial form of *-DIğIndAn beri*.
-(y)IncAyA (kadar/değin/dek)/↓*-(y)AnA (kadar)*	*gid**inceye** kadar* 'by the time [s.o.] went'. *-(y)AnA* is a colloquial version: ↓ *otu**rana** kadar* 'by the time [s.o.] sat down'.
-(A/I)r/-(y)AcAk/-mIş/-(y)mIş/-(I)yor gibi	*kalk**acak** gibi* 'as if about to get up', *anla**r** gibi* 'as if understanding', *içki iç**miş** gibi* 'as if having drunk alcohol'.
-(A/I)rcAsInA/-mIşçAsInA	*hisse**dercesine*** 'as if feeling'. With the form *-mIşçAsInA,* there is the possibility of adding person marking: *konuşuyormuşumcasına* 'as if I was talking'.
-(y)Ip	*koş**up** al-* 'run and get', *gir**ip** otur-* 'enter and sit down'. Because of its conjunctive function, this suffix is discussed in 28.2.
-(y)ken	The segment *-(y)-* is the copula: *bakar**ken*** 'when/while ([s.o.] is/was) watching', *çocuk**ken*** 'when/as a child', 'when [s.o.] was a child', *sokaktay**ken*** 'while in the street', *bizim**ken*** 'when [s.t.] is/was ours'. Unlike the other copular markers, it cannot combine with person markers, except optionally with the 3rd person plural suffix *-lAr: gider(ler)**ken*** 'as they go/went'. It is invariable (i.e. its vowel does not undergo vowel harmony).

Table 3-6 Other adverbial suffixes that create converbs (Göksel & Kerslake , 2005, p. 89)

Among the converbial suffixes above, only *–mIşÇAsInA* can combine with person markers (Göksel & Kerslake , 2005). In addition to these, a few converbial subordinators are added to pairs of verbs that follow immediately after each other.

-(y)A…-(y)A	Added to identical or similar verb stems or to semantically contrasting ones: *bak**a** bak**a*** 'staring', *yedir**e** yedir**e*** 'continuously making [s.o.] eat', *bağır**a** çağır**a*** 'at the top of his/her voice', *gid**e** gel**e*** 'going back and forth', *bat**a** çık**a*** 'sinking and rising'.
↓*-DI…-(y)AlI*	Added to identical verb stems. The first stem has person marking: *duy**dum** duy**alı*** 'ever since I heard [it]', *baktır**dın** baktır**alı*** 'ever since you had [it] checked', *alın**dı** alın**alı*** 'ever since it was bought'.
-(A/I)r…-mAz	This pair of suffixes consists of the aorist and negative-aorist position 3 verbal suffixes (8.2.3.3). These produce a converbial form when added to consecutive identical verb stems without any person marking: *ye**r** ye**mez*** 'as soon as [s.o.] eats/ate', *gid**er** git**mez*** 'as soon as [s.o.] leaves/left'.

Table 3-7 Converbial subordinators that are added to pairs of verbs (Göksel & Kerslake , 2005, p. 89)

3.2 Converbs

In Turkish, adverbial clauses can be finite or non-finite. Finite adverbial clauses are formed with *diye, ki, madem(ki), nasıl ki, (sanki)… -mIş/-(y)mIş gibi* and *-DI mI* (Göksel & Kerslake , 2005):

(23) [Çocukları getir-***ir-ler*** **diye**] porselen eşyayı ortadan kaldırmıştı.
bring-AOR-3PL SUB
'[*Thinking* they would bring the children], she had put the china pieces away.' (p. 399)

On the other hand, non-finite forms of adverbials are much more widely used with some other suffixes and postpositions (24) and these are called as **converbs** generally.

(24) Makine [tamir ed-il-**dikten sonra**] yeniden bozul-du.
machine repair AUX-PASS-CV after again break.down-PF
'[*After* being repaired], the machine broke down again.' (p. 405)

In Turkish, converbs followed by a postposition creates discourse relations and they are named *complex subordinators,* where the postposition is considered the discourse connective (Zeyrek & Webber, 2008). On the other hand, converbs without postpositions may encode a semantic relation between abstract objects by taking a small set of suffixes corresponding to English 'while', 'when', 'by means of', 'as if', or temporal 'since', and they are named *simplex subordinators* which have not been annotated yet in TDB. In the rest of this thesis, we will be focusing on the converbs of the latter type since this thesis is specifically interested in the *simplex subordinators* and aims to solve the ambiguity problems regarding the discourse and non-discourse use of converbs.

The frequent order of the arguments of a converb is ARG2-ARG1, where the converb appears as the final element of second argument. Example (25) illustrates the converb, *soracağına* 'instead of asking', with its first argument in italics and second argument in bold and connective underlined.

(25) Vatandaş **bu paranın hesabını bana <u>sora</u>**<u>cağına</u> *bunu seçimi isterken sorsaydı.* (20490000)
The citizens *should have asked for an explanation for this money when they demanded the election* <u>instead of</u> **<u>bringing</u> me to account.**

The suffixes which can form converbs among the given subordinator suffixes in 3.1.3 are: -*DIK, AcAK, -mAK, -IncA, -ArAK, -cAsInA, -Ip, -ken*. These converbial suffixes combine with other inflectional suffixes yielding the following suffix forms which can be converbs that act as *simplex subordinators*: *-(y)AcAğInA, -AcAğIndAn, -AlI, -(y)ArAk, -(A)rcAsIna, -dIğIndAn, -dIğIndAn, -dIkçA, -IncA, -Ip, -ken, -mAksIzın, mAktAnsA, -mIşcAsInA, -sA.*

3.3 Lexical Semantics

Converbs are non-finite verbs, which have no tense, aspect or mood, but they have suffixes that have senses binding discourse units. The *–ken* suffix, for instance, binds a clause with a 'while' meaning to the superordinate clause. *While* is a word in English; on the other hand, *–ken* is not a word in Turkish even it has the similar meaning.

As is well known, compound word forms may have non-compositional meanings. For example, 'to kick the bucket' has nothing to do with 'kick' or 'bucket' but it means 'to die' or *etekleri tutuşmak* doesn't mean 'catching fire on skirts' but 'being alarmed'. In these cases, we cannot extract the meaning of such forms from their components compositionally. Converbs, especially more frequent ones, can create such compound words like *yanıp tutuşmak* 'to yearn for', *son olarak* 'finally' etc. We will call these compounds with non-compositional meanings *lexical items* to express that the combined form has become a single entity in the lexicon. Such cases show that there are challenges to compositionality caused by *conventionality,* which will be explained below, and we need to determine the degree of conventionality of such compound forms to disambiguate their roles in discourse. We give some essential terms from lexical semantics and examples for them below in order to make use of them while analyzing the converbs in the results section.

Compositionality is the property of the meaning of a phrase when it is derived from the meanings of the words in the phrase and the grammatical relations that joins them. **Pure compositionality** means there are no semantic effects of contextual factors that deviate the meaning of the phrase from the sum of the meanings of its parts (26).

(26) syntax: S → NP VP

semantics: S' = F (NP', VP')

(The Degree of) Conventionality is a matter of identifying lexical units, namely **lexical items** between *transparency* and *opaqueness*.

Transparency is the degree of compositionality. If the whole meaning of a lexical expression can be extracted from its parts, then it's transparent.

(27) Yeşil araba -> Green car

Opaqueness is the degree of lexicalization (conventionality). If the meaning of a lexical expression exceeds the sum of the meaning of its parts, it's then opaque.

Lexeme is an abstract minimal unit of morphological analysis in the lexicon of a language that roughly corresponds to a set of forms of a single word (Brinton & Donna, 2010).

(28) Walk—walk, walks, walked, walking
Run—run, runs, ran, running
Sing—sing, sings, sang, sung, singing

Lexical items are meaningful linguistic units which can be a suffix (29) or complex word forms (30) such as fixed expressions, idioms, clichés, etc.

(29) Sen gel[diğinde] biz çıkıyorduk.
[When] you came, we were about the leave.

(30) [İlk adım olarak] da ezan Türkçeleştirilmişti. (10660000)
[As a first step], the call to prayer was translated to Turkish.

Free/productive combination is the tendency of an expression to be compositional.

Collocation is the tendency of words to occur together.

Fixed expressions are highly conventionalized, but still each syntactic component (partly) retains its semantic contribution.

(31) *ENG: "*Fish and chips", "blackboard", "slow motion", "headline"
TUR: *köşe bucak '*lit. corner nook -> all around', *kurufasülye pilav* 'lit. bean rice -> beans served with or over rice" *alet edavat* 'lit. tool insturments -> a set or group of tools', *soyadı* 'lit. lineage name -> surname', *başlık* 'lit. for head -> bonnet'

Idioms tend to take on meanings that go far beyond the sum of the individual meanings of each of their parts. (Gasser) Idioms often involve metaphoric or anectodal meaning extensions.

(32) *ENG*: “kith and kin”, “kick the bucket”, “to be born with a silver spoon in one's mouth”,
TUR: *meteliğe kurşun atmak* ‘lit. to shoot a bullet through a coin -> to be broke’, *anasının gözü* ‘lit. his/her mother’s eye -> a cunning person’

Lexicalized combinations have lost all sense of combination.

(33) *ENG:* “blueprint”, “live wire”, ““hitchhike”, “seahorse”
TUR: *başvurmak* ‘lit. to hit head-> to apply’*, yankesici* ‘lit. side cutter-> cutpurse’

Conventional constructions have very specific grammatical meanings.

(34) *“The more* he likes her, *the more* she dislikes him.” (Jackendoff 1997: 174)
“One more beer *and* I'm living.”
(*Şimdi ‘now’*) V-*past-agr* V-*past-agr.*

An example for the repetition of past inflected verb in (34) would be *Şimdi geldin geldin.* ‘lit. Now you came, you came. -> It’s your last chance to come!’ In this example, the meaning of the expression goes beyond the simple sum of the meaning of its parts.

The spectrum of conventionality in Figure 3-1 explains how semantically transparent or opaque are the types of expressions defined in this section. The most semantically transparent expressions are those that have free composition. As we progress through the spectrum, the phrases get more *fossilized* or more *lexicalized*, until we reach the *lexicalized combinations* which are completely opaque in terms of their semantics.

Free composition > collocations > fixed expressions > idioms > lexicalized combinations
--------------------------------- -> fossilization (lexicalization) -> -------------------------------------

semantically transparent semantically opaque

Figure 3-1: The spectrum of conventionality

The decomposability of a **multi-word expression** (MWE) is the degree to which the semantics of an MWE can be ascribed to those of its parts.

(35) kick the bucket -> die

The syntactic flexibility of an idiom can generally be explained in terms of its **decomposability**, i.e., how much syntactic variation the idiom allows in its use. For example, in (36) the idioms that are marked by an asterisk (*) are rejected as ungrammatical as idioms, because those idioms have the least syntactic flexibility and are not decomposable. The idioms either lose their idiomatic meaning, or turn out to be completely unintelligible. The idiom with the question mark may allow have some degree of flexibility, allowing for some decomposability. While the native speakers may still be able to extract the idiomatic meaning, they might also be uncomfortable with this unconventional use of the idiom, or only accept in under certain stylistic constraints. Yet some other idioms might have complete syntactic flexibility and retain their idiomatic meaning in a variety of syntactic variations. Such idioms could be fully decomposable.

(36)* The considerable advantage that was taken of the situation…
* The bucket was kicked by Kim.
? Strings were pulled to get Sandy the job.
The FBI kept closer tabs on Kim than they kept on Sandy.

CHAPTER 4

Methodology

In order to grasp a full understanding of the ambiguity of the converbs, we created a tag set, a list of annotation guidelines, and a converb-corpus, which consists of all the sentences that contain a converb in TDB. This chapter presents the tag set and the annotation guidelines, which were necessary for annotation procedure. The guidelines explain the distinctions between the discourse connectives and the non-discourse connectives and also they give some specific rules to apply during the annotation procedure. Then, the annotation procedure is explained in detail. Finally, the preliminary studies for creating the converb-corpus are presented. These studies involve searching and selecting converbs within TDB by using a variety tools.

4.1 Tag set

Taking the annotation guidelines into consideration, a tag set was created for use during the annotation procedure. The tag set and their explanations are given in Table 4-1.

Category	**Explanation**
DC	*DC* stands for Discourse Connective and is used when a converb is a *simplex subordinator.*
Complement	This tag is used when a converb is the object of a verb or the complement of a postpositional phrase. The second level of the annotation indicates the type of complement i.e., *VP Complement* and *PP Complement.*
Adverb	The *Adverb* tag is used when the only role for the converb is to modify the matrix verb.
Ambiguous	Some instances of the converbs can be interpreted to conform to a more than one role. In those cases, the *Ambiguous* tag is used and the second level of the annotation indicates the type of ambiguity since converb can be ambiguous between *DC-OTHER* roles or *OTHER-OTHER* roles.

Non-converb	There may be some instances where the suffix was tagged as a converb by the disambiguator whereas it in fact fulfills another syntactic role. Such errors frequently occur when the disambiguator mistakes a headless relative clause as a converb. In those cases, the second level of the annotation marks them as *HRC*.
Other	The Other tag is used when the converb does not belong to any of the given categories above. Such cases appear in one of the following conditions listed below, and they are labeled with appropriate tags in the second level annotation. *Lexicalized* tag is used when the converb is the part of a lexicalized expression. *No-arg1* tag is used if the converb is a DC but it misses its first argument and therefore cannot be annotated according to TDB guidelines. *Other DC* tag is used when the converb creates a *complex subordinator* with a postposition, or the whole word is lexicalized as a discourse adverbial.

Table 4-1 Tag set that is used to annotate converbs

4.2 Guidelines for the Converb-Corpus

We prepared the guidelines for the annotation of the converbs by analyzing all converbs in TDB texts. While our guidelines are largely based on the annotation principles of TDB, we created some additional rules specific to the converbs in Turkish.

4.2.1 Syntactic Class

As stated in chapter 2, the discourse connectives come from three syntactic categories, which form five classes In Turkish (Zeyrek & Webber, 2008): *Simple coordinating conjunctions* combine two clauses of the same syntactic type; *Paired coordinating conjunctions* are composed of two lexical items such as *hem… hem* 'both… and,' ne... ne 'neither… nor' which link two clauses; *Simplex subordinators* are also called as converbs, which are suffixes forming non-finite adverbial clauses; *Complex subordinators* ***are*** similar to Simplex subordinators yet they usually contain a postposition (such as *rağmen* 'despite', *için* 'for', *gibi* 'as well as'); and *anaphoric connectives* which require only one abstract object syntactically yet retrieve the other argument anaphorically from the previous discourse. We are only interested in the simplex subordinator type in this thesis.

Therefore, if the converb creates a *complex subordinator* with a postposition, it's annotated as *Other/Other DC* (37).

(37) Mide bulantısından nasıl <u>kurtulacağından</u> [önce], o günün bir iş günü olup olmadığını düşünmüş. (00060111)

[Before] thinking about how <u>to get rid of</u> nausea, s/he had thought if it was a workday or not.

4.2.2 Argument types of the simplex subordinators

In Turkish, the subordinate clauses are usually nominalizations, and when they denote abstract objects they are annotated as arguments of the discourse connective as in (38).

(38) **Dinleyici rolüme büyük bir sadakat <u>göster</u>**<u>erek</u> *başımı salladım.* (00035220)
<u>Displaying</u> great loyalty to my role as the audience, *I nodded*.

Converbs take at least one subordinate clause as argument. They can create a discourse relation between two subordinate clauses or between a subordinate clause and the main clause (39).

(39) **Arabaya binip yola koyuldu**ğumuzda *bir süre susuyoruz.* (00005221)
When we get on the car and hit the road, *we are silent for a while.*

Object relativizers (40) and subject relativizers (41) can also be abstract objects, so they can be the first arguments of *simplex subordinators*.

(40) **Ahmet Metin gibi (ismin baş harfleri bile yazarınınkiyle aynıdır), Rakım Efendi gibi idealize ed**erek *özdeşleştiği* bu roman baş kişileri, aklı başında annelerin tuttuğu İslâm düşünce ve terbiyesine vâkıf eğitmenlerle yetiştirilmiş çocuklardır. (00027113)

These novel antagonists whom *he identifies with* **by idealizing like Rakım Efendi and Ahmet Metin (even his initials are the same as the author),** are kids who were raised by tutors who are well rounded in the Islamic reasoning and etiquette that is favored by sensible mothers.

(41) Menderes döneminde 1958'de, **'başka yere nakledileceği' söylen**erek *kaldırılan* Karaköy Camii'nden 45 yıldır ses çıkmadı. (10310000)

There has been no news for 45 years about the Karaköy Mosque*, which was removed* **saying that it will be transferred to somewhere else** in 1958 during the Menderes era.

4.2.3 Minimality Principle

Minimality Principle is applied in converb-corpus annotation in the same manner as it was applied in PDTB and TDB. For the detailed explanation of the principle, see section 2.2.2.

4.2.4 Shared Objects and Modifiers

In TDB, shared subject and objects are annotated with Shared tag along with the arguments which they belong. Yet, in this study shared subjects aren’t annotated for the sake of simplicity since we are interested in converb itself principally. Therefore, shared arguments or modifiers of *Simplex subordinators* are not annotated. In (42), *imam* is not included to second argument since it’s the shared subject of both arguments.

(42) Oysa imam, **daha kuşluk vakti evine konuk ettiği gencin şu anda ölü olduğuna bir türlü inanamadığı**ndan mıdır nedir, *hayli yavaş hareket edip arada bir durgunlaşıyordu.* (00064211)

Whereas the Imam, probably **because he couldn’t bring himself to believe that the youngster who had hosted just this mid-morning was now dead**, *moved quite slowly, looking dull every now and then.*

Modifiers (43) and focus particles (44) of connectives are not included in arguments or connective.

(43) Annesiz geçen çocukluk yıllarından sonra ona kavuştuğunda [da] şefkat eksikliğini yaşıyor.

After the motherless childhood years [just] when he rejoins with her, he feels the lack of compassion.

(44) Yakın tarihimiz henüz tam olarak aydınlanmadığından [olsa gerek] birbiri ardınca yayımlanan anı kitapları okurlardan beklenenin üzerinde ilgi görüyor. (10220000)

[It must be] because our recent history is not completely enlightened yet that the memoirs published one after another draws more interest from the readers than one would expect.

4.2.5 Unannotated connectives due to the lack of an abstract object

Arguments with copula do not denote abstract objects, so such connectives are not annotated.

(45) Bu soruyu sormasını on yaşındayken öğrenmiştim. (00007121)
I had learned to ask this question when I was 10 years old.

Headless relative clauses don't denote abstract objects.

(46) Bu süre umduğumdan daha da kısa oldu. (10700000)
This took even shorter than I hoped it would.

Converbs aren't considered to be discourse connectives when they indicate manner of a verb.

(47) Hasan koşarak eve girdi.
Lit. Hasan entered the house running
Hasan ran into the house

If a converb misses its first argument or takes it anaphorically, then they are annotated as *Other/No-arg1* tag (48).

(48) Protestolarında, canını dişine takmış bir eski zaman şövalyesinin gözü karalığını göremediğimden belki. (00068131)

Maybe it's because I can't see the recklessness of an antique knight going all out in his protests.

If the converbs are the first item in reduplications, they are annotated as *Other/Lexicalized* even when the reduplication creates an abstract object. In (49), *Zıpladıkça* may create an abstract object by itself, yet in this context reduplication *Zıpladıkça zıplardım* creates abstract object. Same rule applies when reduplication is made by the repetition of same converb (50) or negation of same converb (51).

(49) Zıpladıkça zıplardım ve bir cambaz olmaya karar verirdim ansızın. (00010111)

I would keep jumping and jumping, and then suddenly I would decide to become an acrobat.

(50) Onların dallardan yolup yolup attığı erikleri önlüğündeki torbaya dolduruyordu. (00032161)

She was filling the pocket on her apron with the plums they kept ripping and ripping and throwing of the branches.

(51) Bir gün beni merak ettiğini, fotoğrafımı gönderip gönderemeyeceğimi sordu. (20420000)

One day he said that he was curious about me and asked whether I could send him a photo of him or not.

If the converb has another discourse marker role other than a *simplex subordinator*, it's marked as *Other/Other DC*. In (52), *yoksa* 'or' acts as a coordinating conjunction rather than the conditional inflection of *yok* 'to be absent'.

(52) Ben mi yanlış ya da yetersiz düşünüyorum, yoksa bu işte bir tuhaflık mı var, bilmem. (00054223)

I don't know whether I my thoughts are wrong or insufficient, or there is something fishy about this business.

If the converb is available for different interpretations, they are annotated as *Ambiguous*. In (53), *olarak* can be an auxiliary verb of *para* or a *simplex subordinator*.

(53) Tam diyet 100 deve veya para olarak, bin dinar altın veya onbin dirhem gümüştür. (00023213)

The exact ransom is a thousand gold dinars or ten thousand silver drachmae, paid as either 100 camels or in cash.

4.3 Creating the Converb-Corpus

In order to create the converb-corpus and capture all 15 different converbs within their sentential context, we follow the steps given below:

During **Segmentation,** we used a morphological parser and disambiguator to search for the converbs rather than using regular expressions. This method was expected to result in more precise and accurate search results, was well as providing input for the case study of automatic disambiguation of converbs in the following chapters. Since the disambiguator needs sentence boundaries to disambiguate any given morphologically parsed result, we first split the TDB text into sentences. All TDB texts were segmented into sentences and words by using NLTK's segmentation tools. Consequently, for each of the 197 TDB files, we created a separated text file in which the text is split into sentences. For instance, the raw text file '00001131.txt' of TDB was split into 261 sentences and '00001231.txt' was split into 250 sentences and so on.

The next step was **Parsing the Words**, during which each word of every sentence was morphologically parsed with Boğaziçi Morphological Parser (Sak, Güngör, & Saraçlar, 2008).

The morphological analyses were then disambiguated. For **Disambiguation** we used the morphological disambiguator 'Perceptron'. Table 4-2 shows a sample sentence, its morphologically parsed format, and the disambiguated morphological analysis. In the first row of this table, the sample sentence was extracted from the TDB raw text file '00001131.txt' and when the parser is given the words of this sentence; it produced as output a parsed sentence displayed in second row of the table. Finally, disambiguator took the parsed sentence and disambiguated it by sorting its analysis for each word. For example, the order of the analyses of the word *yere* 'at the floor' is changed and the preferred analysis, shown in bold in table, was selected.

Sentence (file 00001131.txt) Ben yere bakmazdım.
Parsed Sentence (file 00001131.parse) <S> <S>+BSTag Ben ben[Pron]+[Pers]+[A1sg]+[Pnon]+[Nom] ben[Noun]+[A3sg]+[Pnon]+[Nom] be[Noun]+[A3sg]+Hn[P2sg]+[Nom] yere yer[Verb]+[Pos]+YA[Opt]+[A3sg] **yer[Noun]+[A3sg]+[Pnon]+YA[Dat]** bakmazdım bak[Verb]+mA[Neg]+z[Aor]+YDH[Past]+m[A1sg] . .[Punc] </S> </S>+ESTag
Disambiguated Sentence (file 00001131.disamb) <S> <S>+BSTag Ben ben[Pron]+[Pers]+[A1sg]+[Pnon]+[Nom] ben[Noun]+[A3sg]+[Pnon]+[Nom] be[Noun]+[A3sg]+Hn[P2sg]+[Nom] yere **yer[Noun]+[A3sg]+[Pnon]+YA[Dat]** yer[Verb]+[Pos]+YA[Opt]+[A3sg] bakmazdım bak[Verb]+mA[Neg]+z[Aor]+YDH[Past]+m[A1sg] . .[Punc] </S> </S>+ESTag

Table 4-2 A sentence and its morphological parses

Next, for the **Search for the Converbs** in the data, the morphological analyses were scanned for the words with a converb tag in its disambiguated analysis using the part of speech tags given by the disambiguator.

The final step before the annotation process was the **Selection of the Converbs** to be annotated. We had created a converb-corpus which comprised of 10170 sentences from TDB. In order to capture all possible usages of each converb, 1475 instances were **selected for annotation**. For the converbs which have too many search results, approximately 150 sentences were randomly selected, preserving the genre distribution in TDB (see. Table 4-3).

Converb	**# of Sentences in Converb-Corpus**	**# of Selected Sentences for Annotation**
-AcAğInA	121	121
-AcAğIndAn	31	31
-AlI	22	22
-ArAk	3201	150
-ArcAsInA	48	48
-dIğIndA	501	150
-dIğIndAn	426	152

-DHkçA	126	126
-IncA	472	150
-Ip	2656	150
-ken	1322	150
-mAksIzIn	48	48
-mAktAnsA	7	7
-mIşcAsInA	20	20
-sA	1169	150

Table 4-3 Converbs and their number of instances

4.4 The Annotation Procedure

The selected converbs in the sample sentences were underlined to facilitate the annotation task. Due to the priorities of the thesis, we only annotated the converb itself and its two arguments and did not annotate the shared subjects or modifiers (see 4.2.4). Two annotators looked for the two arguments of the converbs using semantic criteria (see 4.2.5) and the minimality principle (see 4.2.3). For the converbs, the second argument is always syntactically attached to the converb and resides within the sentence. The first argument is expected to follow the converb in most cases, but it can reside in any position within the sentence. The text was preprocessed to underline the converbs, and during the annotation, the annotators followed the convention of the PDTB by annotating the first argument in italics and the second argument in boldface for the instances where the converb was interpreted as a discourse connective.

The annotation process of the converb-corpus involved two steps. First, the selected converb tokens were annotated by two annotators with the given tag set (see 4.1). Second, the disagreements were discussed and resolved during an agreement meeting of the two annotators. In order to test the reliability of the annotations, we measured inter-annotator agreement by means of Kappa statistics, because we used only categorical data and didn't measure the agreement on the argument spans. If the agreement is higher than 0.80 it indicates a good level of agreement. A complete list of how Kappa results (Landis & Koch, 1977) can be interpreted is given in Table 4-4:

Kappa	Interpretation
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Table 4-4 How agreement result is interpreted

CHAPTER 5

Results

This thesis argues that there are three kinds of cases regarding the ambiguity of the converbs: the hard cases in which the abstract object interpretation is so subjective that it is hard to annotate such cases even for the human annotators; the ambiguous converbs with arguments easy to recognize, which can be easily differentiated between their different roles; and the unambiguous converbs which always signify a discourse relation. This chapter proposes methods to clarify the cases for the highly ambiguous converbs, explains *ambiguity resolutions, availability for automatic disambiguation, and the possible morphologic/syntactic/semantic features* for each converb according to the annotation results and inter-annotator agreement statistics. For the senses of the connectives in this section, see Figure 2-3: Hierarchy of sense tags .

5.1 -(y)AcAğInA

The morphology of this suffix is as follows:

(54)–*(y)AcAk* + (-*i*) + [m|n|mIZ|nIZ|lArI(n)] + -*A*.
-NONFACT-ACC-ARG-DAT

The converb -(y)AcAğInA can be a simplex subordinator, which is annotated in this study as a DC; the complement of a verb phrase or a postpositional phrase, which is annotated as NDC; or can take part in a complex subordinator, which is annotated as OTHER DC.

When this suffix is a *simplex subordinator*, it means ‘instead of’ and usually introduces a discourse relation with EXPANSION:Alternative:chosen alternative sense.

(55) Vatandaş **bu paranın hesabını bana** <u>**sora**cağına</u> *bunu seçimi isterken sorsaydı*. (20490000)

The citizens *should have asked for an explanation for this money when they demanded the election* instead of **bringing me to account.**

The suffix *-(y)AcAğInA* may be attached to the complement of certain factive verbs such as *emin olmak* 'to be sure'*, inanmak* 'to believe' etc. (56).

(56) Onu bulacağınıza eminim. (00006231)

I'm sure you'll find it/him/her.

The suffix *-(y)AcAğInA* may also be attached to the complement of a postpositional phrase (57). These postpositions are generally *yönelik, ilişkin, dair*, etc., most of which convey aboutness.

(57) Parti amblemlerinde nelerin kullanılamayacağına ilişkin yasal düzenlemeler var (20250000)

There are legal regulations about what cannot be used in party emblems.

The converb *-(y)AcAğInA* can occur in a complex subordinator with the postposition *göre* (58). This subordinating conjunction has the meaning of *since, because* and thus conveys CONTINGENCY:Cause:reason relation.

(58) Sonra babam, ``Artık İstanbullu olacağına göre vapur düdüklerine alışsan iyi edersin,'' diyerek yeniden güldü. (00008213)

Then dad smiled again and said "Since you will be and İstanbulite soon, you better get used to the steamboat whistles".

In the annotated instances, there are also non-converbial forms of *–(y)AcAğInA*, such as *çalış-acak-lar-a* 'to those who will work', which is a headless relative clause (59). Such erroneous instances are due to deficiency in searching for *–(y)AcAğInA*, and they can be successfully eliminated by looking for (-i) accusative case + (n/m) person agreement markers in the morphology, since these markers only reside in the converbial cases of *–(y)AcAğInA*.

(59) Sermaye Piyasası Kurulu, geçen yıl uygulamaya koyduğu düzenlemeyle sermaye piyasasında çalışacaklara lisans zorunluluğu getirdi. (20320000)

Capital Markets Board brings necessity of license for people who work at capital markets by the regulations last year.

Annotation Results and Disagreements

There were a total of 121 instances of *–(y)AcAğInA*, and 10 of these instances were annotated as DC, 95 of them were annotated as *Complement* and 5 of them were annotated as *Other* by both annotators. The annotators didn't agree on the remaining 11 instances. Only one case was a disagreement between DC and NDC uses and the remaining 10 were disagreements between various NDC uses.

The inter-annotator reliability for the annotators is Kappa= 0.948 (p <0.05) for DC-NDC discrimination, and annotators achieved an almost perfect agreement score.

Most of the disagreements were between *Other* and *Complement* where postpositional phrases like *yapılacağına ilişkin/dair/yönelik* 'about it'll be done', which were annotated as *Other* by *Annotator1* and *Complement* by *Annotator2*. In these cases, converbs are the complement of the postpositional phrases in which *dair/yönelik/ilişkin* are head of the phrase. Thus, *Complement* is the true label for these instances. Additionally, there are 4 Non-Converb instances which are considered as disambiguator errors.

The non-converb cases of *-AcAğınA* can be eliminated by checking the morphology, thus they are not considered as a ground for high ambiguity. Otherwise, *-AcAğınA* can be a *simplex subordinator,* take place in a *Complement* of verb phrases or postpositional phrases, and *complex subordinators*. The complements of postpositional phrases can be simply found by looking for a postposition such as *dair* 'regarding', *yönelik* 'towards', *ilişkin* 'about' after the converb. Similarly, *complex subordinators* can be identified by looking for the postposition *göre* 'since'. On the other hand, distinguishing the *simplex subordinators* from *complements of verb phrases* requires more robust techniques. First of all, *-AcAğInA* creates a subordinate clause which is the object of the superordinate clause when the converb is the *complement of a verb phrase* (60). Conversely, a *simplex subordinator* is not object of any verb phrase and it creates a separate subordinate clause which is not syntactically bound to any superordinate clause (61).

(60) Onlar [silahla bir şeylerin değişeceğine] inanıyordu. (00057221)
They believed [that something would change with guns].
(61) … [**sarkacı durdur**acağına] *var gücüyle aşağı çekmişti* (00068231)
[Instead of stopping pendulum] he/she pulled it down with all his strength.

Additionally, *complements* precede factive verbs such as *inanmak* 'believe', *güvenmek* 'trust', *dikkat çekmek* 'attract attention' since they supply the presupposition created by factive verbs. Eventually, both *simplex subordinators* and verb phrase complements create clauses that can be interpreted as abstract objects. However, only *simplex subordinators* create a discourse relation with another abstract object because of the syntactic availability of its clause. Eventually, *-AcAğInA* can be disambiguated easily by human annotators and it can be disambiguated automatically given syntactic and verb semantic features.

5.2 -(y)AcAğI(n/m)dAn

The morphology of this suffix is as follows:

(62) –*(y)AcAK-I*-[m|n|mIZ|nIZ|lArI(n)]-dAn
-NONFACT-ACC-ARG-ABL

The suffix -(y)AcAğI(n/m)dAn can be a simplex subordinator, annotated as DC; the *complement* of a verb phrase, annotated as NDC; or can take part in a complex subordinator, annotated as *Other*.

When this suffix is a *simplex subordinator*, it means *since,* and therefore creates a discourse relation with CONTINGENCY:Cause:reason sense (63).

(63) Dolayısıyla ''tarihsel ve sosyal değerler ile olaylar**, bu tür yalın bir mantık düzeyinde değerlendirileme**yeceğinden *bu görüşlerin de inceleme alanına girmemekte*'' dir. (10640000)

Therefore, "**since historical and social values and events cannot be evaluated at such a basic logical level**, *they are not in the scope of these views*"

The suffix *-(y)AcAğI(n/m)dAn* can be the complement of a factive verb like *emin olmak* 'to be sure'*, haberdardı olmak* 'to be aware of'*, korkmak* 'to be afraid' etc. (64).

(64) Ama bürokratların personel sayısı konusunda doğru bilgi vereceğinden emin olamayız. (20220000)

But we can't be sure whether the bureaucrats will give accurate information about the number of the personnel.

-(y)AcAğI(n/m)dAn also forms complex subordinators with postpositions like *dolayı* 'since'*, ötürü* 'due to'*, önce* 'before' etc. Postpositions *dolayı* and *ötürü* have a similar sense with its simplex form, yet *önce* means 'before', and it has TEMPORAL:Asynchronous:precedence sense **Hata! Başvuru kaynağı bulunamadı.**.

(65) Mide bulantısından nasıl kurtulacağından [önce], o günün bir iş günü olup olmadığını düşünmüş. (00060111)
[Before] thinking about how to get rid of nausea, s/he had thought if it was a workday or not.

Similar to *-(y)AcAğInA*, *-(y)AcAğI(n/m)dAn* also has non-converb instances such as *gelecek-ler-den* 'from those who will come', which is a headless relative clause (66). Such erroneous instances are due to deficiency in searching for *-(y)AcAğI(n/m)dAn*, and they can be successfully eliminated by looking for (-i) accusative case + (n/m) person agreement markers in the morphology, since these markers only reside in converbial cases of -*(y)AcAğI(n/m)dAn*.

(66)Başıma geleceklerden korkuyorum sonra ve tahta atın aklının ucundan bile geçmiyorum. (00010111)

Then I'm afraid of what would happen, and I don't even cross the mind of the wooden horse.

Annotation Results and Disagreements

There were a total of 31 instances of *–(y)AcAğInA*, and 22 of these instances were annotated as DC, 3 of them were annotated as *Complement*, 2 of them were annotated as non converb and 1 of them was annotates as *Other* by both annotators. The annotators didn't agree on the remaining 3 instances. Only one case was a disagreement between DC and NDC uses and the remaining 2 were disagreements between various NDC uses.

The inter-annotator reliability for the annotators is Kappa= 0,839 and p < 0,05, achieving high reliability. There are 2 disagreements between *Non-Converb* and *Complement*, but in these instances both tags are correct since these instances are non-converbs as well as complements of verbs.

Similar to *–AcAğIndA*, the non-converb cases of *-AcAğındAn* can be eliminated by checking the morphology, and thus they are not considered as a ground for high ambiguity. The suffix *-AcAğındAn* can be a *simplex subordinator,* can occur in *complements* of verb phrases and *complex subordinators*. *Complex subordinators* can be identified by looking for the postpositions like *dolayı* 'because', *ötürü* 'due to', and *önce* 'before'. Therefore, the ambiguity is primarily based on the distinction between the two roles of *–AcAğIndAn*; the *simplex subordinator* and the *complement* of verb phrase. *Simplex subordinator* role of the converb creates a subordinate clause which is not the object of the main clause, whereas the *complement* role creates clauses that are objects in these sentences. Also, the *complements* precedes factive verbs such as *korkmak* 'to be afraid of', *emin olmak* 'to be sure', *çekinmek* 'to shy away from' etc., thus *-AcAğIndAn* can be disambiguated by human annotators with a high agreement (Kappa = 0,839) and it can be disambiguated automatically, given the syntactic and verbal semantic features. Thus, *-AcAğIndAn* is considered to be a less ambiguous converb.

5.3 –(y)AlI

-(y)AlI is a subordinating suffix. It's the colloquial form of *-DIğIndAn beri* 'ever since'.

The suffix *-(y)AlI* is a converbs that can occur with or without a postposition.

When *-(y)AlI* is a *simplex subordinator,* it mean "since" and its meaning is TEMPORAL:Asynchronous:precedence (67).

(67) **Atatürk öle**li *dört yıl kadar olmuştu*; adını ``Ebedi Şef'' koymuşlardı.

It had been four years **since Atatürk died** when they named him "the Eternal Chief".

-(y)AlI composes a complex subordinator with *'beri',* which has the same meaning as the *simplex subordinator* (68).

(68) **Koalisyon kurula**lı beri *buradaki özel timin sorgusunda iki genç hayatını yitirmiş.*

Since the coalition was established, *two youngsters lost their lives during the interrogation by the special task force.*

In addition to the above uses, *-AlI* can be found in reduplications such as *bildim bileli* 'as far as I can remember', gitti gideli 'ever since he/she left'. Since the converb *–AlI* is at the right edge of the reduplication, connecting it with the rest of the sentence with the same sense as the simplex subordinator, these occurrences were also annotated as such.

(69) *Güncel politika olaylarına karşı duyduğum yoğun ilgi* **kendimi bildim bil**eli *yüksekti.*
My intense interest towards current policy events is always high all my life.

Annotation Results and Disagreements

19 of the 22 instance were agreed by both annotators but still inter-annotator agreement is not high because of the small sample size. The inter-annotator reliability for the annotators is Kappa= 0.593 (p <0.05). Besides, 3 differences in annotation are due to different interpretation of the reduplication *bildim bileli* 'since I can remember'. With clear annotation guidelines that include this special use, higher agreements could be achieved.

The converbs *–AlI* creates *simplex* and *complex subordinators*. *Complex subordinators* can be distinguished by the following postposition *beri* 'since'. Therefore, *-AlI* creates converbs which are unambiguous.

5.4 -(y)ArAk

-(y)ArAk is a frequent means of conjoining clauses which are semantically of equal status with respect to tense/aspect/modality and it express manner directly, in terms of an accompanying action or state (Göksel & Kerslake , 2005).

-(y)ArAk can be a *simplex subordinator,* annotated as DC; act as a manner of verbs, annotated as *Manner;* and form discourse adverbials and idiomatic expressions, annotated as *Other*.

When it is a *simplex subordinator -(y)ArAk* has the meaning of 'by doing' and signifies a discourse relation with EXPANSION sense similar to *–Ip* (70).

(70)**Sandalyemin tekerleklerini <u>çevir</u>**<u>erek</u> *koltuğunun önüne gelmiştim*. (00001131)

I had come in front of his/her armchair **<u>ny turning</u> the wheels of my chair.**

Sometimes the verb with *–ArAk* does not denote a separate event, fact or state about the world other than its matrix verb, and it only modifies the matrix verb. In such cases it's annotated as *Adverbial* rather than *simplex subordinator*. For example, *bilerek* 'lit. knowingly -> intentionally' in (71) is not a discrete event; instead, it only modifies the main verb as the sentential adverb.

(71) Mektubu okumayı <u>bilerek</u> geciktirdi. (00054123)

He/she delayed reading the letter <u>intentionally</u>.

When *ol- 'be'* verb creates converb with *–(y)ArAk*, it generally becomes an auxiliary verb of a compound verb and has a meaning of *as* (72). In such cases, *ol-arak* is annotated as *Other*.

(72)Bu yapı birimi, [tapınak] <u>olarak</u> adlandırdığımız bir yer. (00013112)

This construction unit is a location we refer to <u>as</u> [temple].

There are also ambiguous cases. For example, in (73) *ışık topu olarak* can be interpreted both as 'as a light ball' and 'being a big light ball' so it can be either a *simplex subordinator* or an auxiliary verb.

(73)Güneş, denizin üstünde iri bir [ışık topu <u>olarak]</u> alçalıyor suları altın rengine boyayarak. (00005221)
The sun is setting, painting the water in gold [as a big light ball/being a big light ball].

Olarak can also create a discourse adverbial as in (74). Such adverbials are generally formed with lexicalized items like *ilk olarak '*firstly', *son olarak* 'finally' etc.

(74) [İlk adım <u>olarak]</u> da ezan Türkçeleştirilmişti. (10660000)

[As the first step], the call to the prayer was translated to Turkish.

Annotation Results and Disagreements

A total of 150 instances of –*ArAk* were annotated. In 65 cases both annotators agree on DC, on 4 cases on *Manner*, and on 48 cases as *Other*. The annotators could not agree on 33 cases. The inter-annotator reliability is reported with Kappa = 0,674 (p <0.05) for DC-NDC distinction. A large part of the ambiguity is due to Manner and DC role of -*ArAk*. There are 23 instances in which two annotators disagree between DC and Manner roles. The agreement meetings did not result in clear cut guidelines for distinguishing these two uses, since the annotators have trouble both in identifying the difference between them and justifying their own annotations. Therefore, -*ArAk* instances are considered as *Highly Ambiguous*. Moreover, auxiliary verb *olarak* causes disagreements because of the possible different interpretations.

Converbs created by –*ArAk* are ambiguous between *Simplex subordinator*, *Manner* and *Other* categories. *Other* category comprises discourse adverbials and idiomatic expressions, for example, *olarak*, the converbial form of *ol-* 'be' formed with –*ArAk*, creates a lexicalized item which is a marker of certain types of adverbial phrases and such instances can be distinguished from *Simplex subordinators* and *Manners* (Göksel & Kerslake , 2005). However, *Simplex subordinator* and *Manner* categories both modify the main verb, so there is no syntactic clue to differentiate between them.

One possible solution is looking for semantic relations between the converb and its matrix verb. In (75), *Bağırarak* 'shouting' is annotated as *Manner* since it modifies the verb *söyledi* 'told' which is semantically connected to shouting. Yet if *bağırarak* modifies a verb such as *uyandı* 'woke up' then it would be considered as *simplex subordinator* since it denotes a separate event other than waking up (76).

(75)Bağırarak şimdi hatırlayamadığım birşeyler söyledi. (00058211)

He/she shouted out something which I cannot remember now.

(76)Tam o anda **bağır**arak *uyandı*. (00001231)
He/she woke up shouting.

Nevertheless, checking semantic relations between the verbs is not an easy task. Most of the disagreements of –*ArAk* annotations are due to this problem. The annotators cannot agree on the role of the converb *okuyarak* 'by reading' in (77) for instance.

(77)**Gününü kitap oku**yarak *geçiriyordu.*
He/she was spending the day reading a book.

Another clue for the DC-Manner differentiation is that the converbs that have separate objects other than matrix clause are likely to be interpreted as abstract objects. For instance, *yerleşim ve tarım alanları* 'resident and agricultural areas' is the object of the adverbial clause and *doğal yaşamı* 'natural life' is the object of the matrix clause (78). Since the matrix clause and the subordinate clause have their own objects, the subordinate clause is more likely to be interpreted as an abstract object. On the other hand, in (79), only the matrix clause has an object *Mektubu okumayı* 'reading the letter', so the subordinate clause *bilerek* 'intentionally' is considered as a manner rather than an abstract object.

(78)**… [yerleşim ve tarım alanları aç**arak] [*doğal yaşamı tehdit eden]…* (00011112)
The one [*who threatens the natural life*] [by expanding **residential and agricultural areas**]
(79)[Mektubu okumayı] bilerek geciktirdi. (00054123)

He/she delayed [reading letter] intentionally.

As a result, *–ArAk* creates highly ambiguous converbs which cannot be distinguished between their *Simplex Subordination* and *Manner* roles solely based on the syntactic features since they are all manners syntactically. Disambiguation should take place both at semantic level and syntactic level. Therefore, *-ArAk*, along with its adverbial function, must denote a discrete event other than modifying its matrix clause in order to be *Simplex subordinator*.

5.5 –(A/I)rcAsInA

The morphology of this suffix is as follows:

(80)*-(A/I)r-cAsInA*
-AOR *-CONV*

-cAsInA derives manner adverbs from adjectives with a negative connotation: *aptalcasına* 'stupidly', *salakçasına* 'like a twit'.

The suffix -(A/I)rcAsInA can be a *simplex subordinator* or act as *manner*. Both tags are syntactically adverbials.

When it is a *simplex subordinator***,** *-(A/I)rcAsInA* form discourse relations with an EXPANSION sense (81). *-(A/I)rcAsInA* has the meaning of 'as if', 'like' such as *hissedercesine* 'as if feeling'.

(81) Genç kız, **bedenindeki yorgunluğu atmak** **iste**rcesine *kımıldayıp duruyordu.* (00045224)
The young girl was *fidgeting continuously* as if she **wanted** **to remove the fatigue from her body.**

Similar to the *–(y)ArAk* suffix, if -(A/I)rcAsInA does not denote a separate event, fact or state about the world, and only modifies the matrix verb as a sentential adverb and it's annotated as *Adverbial* rather than *Simplex subordinator*. For example, *taparcasına* 'as if worshiping' in (82) is not a discrete event from the main verb *sevmemden* 'my loving', instead, it only modifies it by 'excessively' meaning.

(82)İyi olmamdan, onu taparcasına sevmemden sıkıldı. (00002213)

He/she was tired of me loving her worshippingly.

Annotation Results and Disagreements

Of the 48 total instances, 36 were annotated as DC, 1 was annotated as manner and 1 was annotated as *other* by both annotators. The remaining 10 were cases of disagreement. The inter-annotator reliability for the annotators is Kappa = 0.303 ($p < 0.05$) in *–ArcAsInA* annotations. Similar to *–ArAk*, the ambiguity of 9 instances out of 10 is due to the *DC-Manner* distinction. The abstract object interpretation of the annotators for these converbs was different.

Similar to *–ArAk*, *–ArcAsInA* is highly ambiguous between two roles; *Simplex subordinator* and *Manner*. There are also some cases where the role of *–ArcAsInA* depends on the subjective interpretation of the reader. In such occasions, despite the converb*–ArcAsInA* denotes a discrete event; there is a strong semantic relation between the converb and the verb of the main claus,e so the line between the DC and NDC roles is blurred. Such ambiguities can be resolved by looking at the frequency of the compound verb forms. In (83), adverbial clause *koşarcasına* 'as if running'doesn't have its own object that makes also hard to interpret it as abstract object.

(83)Arkamı dönüp koşarcasına uzaklaşıyorum yanından. (00007121)
I turn back and 'run away/move away as if running' from him/her.

And in (84), the converb *yararcasına* 'as if splitting' becomes the predicate of the idiomatic expression *kılı kırk yarmak* 'splitting hairs', which is the manner of another adverbial clause *analiz ederek* 'by analyzing'. Consequently, the clause is not denoting an abstract object and it's annotated as *Manner*.

(84) Marx ve Engels, 2.5 aylık işçi iktidarı Paris Komünü deneyini kılı kırk yararcasına analiz ederek bazı sonuçlara varmışlardır: Sosyalizm, kapitalizm ile sınıfsız toplum (komünizm) arasında kısa bir geçiş aşamasıdır. (00012112)
Marx and Engels had come to some conclusions by analyzing the 2.5-month-long Paris commune and workers-in-power experiment very carefully: Socialism is a short transition phase between capitalism and classless society (communism).

5.6 –dIğIndA

The morphology of this suffix is as follows:

(85)*–dIK-(I)*-[m|n|mIZ|nIZ|lArI(n)]-*dA*
-FACT-AGR-LOC

The suffix *–dIğIndA* can be a simplex subordinator (DC), complement of verb phrase (NDC) or occur in a discourse adverbial (*Other* DC).

When it is a *simplex subordinator,* the characteristic function of *–dIğI(n/m)dA* is to indicate that the situation described by the superordinate clause is/was ongoing at the time of the event expressed by the adverbial clause (86). *–dIğI(n/m)dA* means 'when' and it forms discourse relations with TEMPORAL sense.

(86)**Remziye kapıyı açtı**ğında yoğun bir duman bulutuyla karşılaştım. (00045124)

When **Remziye opened the door**, I came across a very dense cloud of smoke.

The suffix *–dIğIndA* can be a complement for a limited list of verb phrases (87), (88), (89).

(87) Şimdi ileri sürülen Kıbrıs ve Ege'nin arkasında Patrikhane, Heybeliada Ruhban Okulu, Pontus gibi sorunların da olduğunda kuşku yoktur. (10250000)
There is no doubt that there are also problems like Patrikhane, Heybeliada Ruhban Okulu, Pontus behind the Cyprus and Ege issues that are put forward recently.

(88)Tanıklar ve kanıtlar katilin İbrahim Çiftçi olduğunda birleşmişti. (10510000)
The witnesses and the evidences converged on that the killer was İbrahim Çiftçi.

(89)Kar sporları programında kayak yapanların sıçrattığı karla üşüttüğünde ısrar ediyor… (00053123)
He/She insists on having caught the cold because of the snow that splashed from the skiers on the winter sports program.

This converb can also occur in discourse adverbials when followed by the appropriate pronouns (90).

(90)Böyle olduğunda da her zaman olduğu gibi yükselen enflasyon ve faizler reel faizi olduğundan daha şişirecek'' dedi. (20560000)

He/she said that "When it's like this, increasing inflation and interests will raise real interest…"

As in the *-(y)AcAk* instances, there are also non-converbial forms of *–dIğIndA*, such as *anlattıklarımda* 'those that I have *told'*, which is a headless relative clause (91). Such erroneous instances are due to the deficiency of extraction of the converb instancesof – *dIğIndA* . In this case, this error can be identified just by the plural marker preceding the first person possessive marker, because this case can only be interpreted as a headless relative clause. However, if the converb was *anlattıklarında*, we would need the context of the converb in order to distinguish its role since it can be *senin anlattıklarında* 'in those that you have told' or *onlar anlattıklarında* 'when they told'. This ambiguity arises because Turkish second singular and third plural possessive markers have the same morphology.

(91)Bütün bu anlattıklarımda kedice olmayan birşeyler olduğunu biliyorum. (00054223)

I know that there is something which not cat-like in all those that I had told.

Annotation Results and Disagreements

Of the total 150 instances annotated, 1 was annotated as *Complement*, 144 were annotated as DC, and 1 was annotated as Non-*Converb* by both annotators. The remaining 4 were disagreements. The inter-annotator reliability is Kappa = 0.656 (p <0.05) which is relatively lower than expected. This is mostly because of the *DC-NDC* distribution in the sample, where there is only one *Complement* instance agreed by the both annotators. Therefore, *-dIğIndA* is considered less Ambiguous since most of the instances are *DCs* rather than *Complements* despite the low agreement scores.

Non-Converb and discourse adverbial roles of *–dIğIndA* can be differentiated by morphology in most cases, so converbs with *–dIğIndA* are ambiguous between *Simplex subordinator* and *Complement* roles. *Complement* roles precede a very limited set of factive verbs such as *şüphesi olmak* 'to have a doubt about', *birleşmek* 'to agree on' etc., and – *dIğIndA* turns out to be a *Simplex subordinator* most of the time. Consequently, it's considered as less ambiguous.

5.7 –dIğIndAn

The morphology of this suffix is as follows:

(92)*–DIK-(I)*-[m|n|mIZ|nIZ|lArI(n)]-dAn
-FACT-AGR-ABL

The converb *-dIğIndAn* can be a simplex subordinator (DC), complement of verbs (NDC) or creates complex subordinators and headless relative clauses (*Other*).

As a *simplex subordinator, -dIğIndAn* means 'because of', 'since' and forms discourse relations with CONTINGENCY:Cause:reason sense (93).

(93)Yaralar **güneş ışığı görmedi**ğinden *iyice azmıştı.* (00001231)
The wounds have *gotten worse* since **they haven't been exposed to sun light.**

The suffix *-dIğIndAn* can also form complex subordinators with postpositions like *dolayı*, *ötürü* 'since' with CONTINGENCY sense or *bu yana* 'since' with purely TEMPORAL sense (94). According to the postposition, complex subordinator can have Contingency or Temporal sense. This sense ambiguity is similar to the sense ambiguity of *since* in PDTB (Prasad, et al., 2008) but the postpositions in complex subordinators usually prevent this ambiguity (Demirşahin, Sevdik-Çallı, Balaban, Çakıcı, & Zeyrek, 2012).

(94)İstem dışı bir bakıştı; işte sonunda geldim, anlamına da gelirdi, son bıraktığımızdan bu yana canımı sıkacak olumsuz bir şey olmuş mu anlamına da. (00032261)

It was an involuntary look; it meant both here I finally came, and has anything bad happened since I left.

Converb with *-dIğIndAn* can be complement of factive verbs (95).

(95)Son günlerde oldukça verimsiz olduğumdan yakınıyorum ona. (00005221)

Recently, I have been complaining to him about being very unproductive.

Similar to the *dIğIndA* case, there are also non-converbial forms of *–dIğIndAn*, such as *gördüklerimden* 'of what I saw' which is a headless relative clause (96). Such erroneous instances are again due to the errors in the extraction the converbs. Such errors can be identified just by looking for the plural marker preceding the first person possessive marker, because this case can only be interpreted as a headless relative clause.

(96)Neydi ki telâşım, gördüklerimden hoşnuttum; kendimi canlı, istekli ve amaçlı bulmuş olmaktan öyle hoşnuttum ki, ``Bütün zamanlar senin, niye tadını çıkarmıyorsun ki?''

What was my hurry, I was happy with what I saw; I was so happy to find myself alive, keen and oriented, that (I thought)"all the time is yours, why don't you enjoy it?"

Annotation Results and Disagreements

Of the total 152 instances annotated, 31 were annotated as *Complement*, 91 were annotated as DC, 2 were annotated as *Non-Converb* and 19 were annotated as *Other* by both annotators. The remaining 9 instances were disagreements. The inter-annotator reliability for *–dIğIndAn* annotations is Kappa = 0.945 (p <0.05). Differences are due to misplacement of the first argument within sentence. Also *Non-Converbs* are *Complements* of a verb phrase at the same time so they result in disagreements although both annotations are technically not wrong. The converb *–dIğIndAn* is less Ambiguous in terms of their *DC-NDC* roles.

The Complex Subordinator and the *Non-Converb* cases can be identified by the morphology and the postpositions. *–dIğIndAn* is ambiguous between the *Simplex subordinator* and the *Complement* roles, and they can be disambiguated by checking the factive verbs of the *Complements* and other syntactic features. However, there are some Headless Relative Clauses that we should pay attention. In (97), *korktuğumdan* 'more than what I feared'is a headless relative clause and there is an ambiguity due to the same morphology with converbs. One difference is that the *da* particle that follows the converb is a focus particle and means 'too'. On the other hand, *da* following the headless relative clause is a modifier with 'even' meaning. But still, there is no apparent syntactic clue to differentiate them and the context knowledge is essential for disambiguation.

(97)Cezam, korktuğumdan [da] ağır oldu. (00008213)
My penalty was [even] more severe than I feared.

Consequently, the converb *–dIğIndA* is less ambiguous, since it can be disambiguated by syntactic features, and therefore annotated by human annotators with a high agreement (Kappa = 0,945).

5.8 –dIkçA

This suffix is a combination of –DIK, the factive subordinating suffix and –cA, a derivational suffix. One of the functions of the converbial suffix *-DIkçA* is to indicate that one event happens in proportion to the occurrence of another.

-dIkçA creates converbs that can be simplex subordinator (DC), form adverbial items and other lexicalized compound words (*Other*).

When it is a *Simplex subordinator,* the converb *–dIkçA* signifies a discourse relation with the meaning of 'as far as' or 'whenever'.

(98)**Ben yürü**dükçe *gökyüzünün rengi de değişiyordu.* (00007121)

The color of the sky changed **as I walked**,

The converb *–dIkçA* can also only modify a sentence or a clause and will not denote a discrete event from the verb phrase in which it modifies. In these cases, it is annotated as *Other*

(99) Başka deyişle, Kemalizm, baştan belirlenmiş bir düşün ve uygulamalar dizelgesi değil, mantık temelinde, olabildikçe değişik düşünce ve uygulamalara olanak veren bir açılımdır, diye düşünüyorum. (10510000)

In other words, I think that Kemalism is not a predetermined mentality and applications list; it's rather an expansion which permits different thoughts and applications as much as possible on a logical basis.

Some examples of *–dIkçA* appear in reduplication and they generally mean that the action happens repeatedly or continuously. In such cases, the converb *–dIkçA* doesn't denote a separate abstract object, so they aren't annotated as *Simplex subordinator*. Nevertheless reduplication can be an abstract object.

(100) Zıpladıkça zıplardım ve bir cambaz olmaya karar verirdim ansızın. (00010111)

I would keep jumping and jumping, and then suddenly I would decide to become an acrobat.

Annotation Results and Disagreements

Of the total 126 instances annotated, 110 were annotated as DC, and 14 were annotated as *Other* by both annotators. The remaining 2 were disagreements. The inter-annotator reliability for –dIkçA annotations is Kappa = 0.924 (p <0.05). One of the disagreements was due to fırsat buldukça 'on occassion' whose idiomatic meaning leads to a non-abstract object interpretation of it. The converbs –dIkçA is considered less Ambiguous between DC-*Other* roles as *Other* roles consist of mainly reduplication.

Except reduplications and idiomatic usages, *–dIkçA* always creates converbs that become *Simplex subordinators*. *–dIkçA* creates idiomatic expressions such as *gün geçtikçe* 'day by day', *fırsat buldukça* 'on occasions', *olabildikçe* 'as possible' etc. that are not taken as abstract objects. In general, reduplications are formed by the repetition of same verb such as *uzadıkça uzuyor* 'getting longer', *karadıkça kararıyor* 'getting dark' etc. where the converb adds a continuum meaning to verb phrase. Therefore, *–dIkçA* is less ambiguous.

5.9 –IncA

-IncA is an adverbial suffix that expresses a temporal relation between two clauses in general.

-IncA can be a *simplex subordinator* (DC), create complex subordinators, lexicalized and idiomatic expressions (*Other*).

When it is a *Simplex subordinator -IncA* signifies *t*emporal relations that specify the time of the situation expressed by the superordinate clause (Göksel & Kerslake , 2005). Therefore, it produces discourse relation with TEMPORAL sense

(101) **Dikkatlice bakı**nca, *kızın bir kukla olduğunu gördüm.*(00002113)
When I looked carefully, *I saw that the girl was a puppet.*

The converb *-IncA* also occurs in complex subordinators in the form of *-(y)IncAyA kadar/değin/dek* 'until' Since the postposition requires the dative suffix *–yA* after *–IncA*, their retrieval is taken as error, and they can be easily removed by checking for the dative suffix at the end of the converb.

(102) Yeni hükümet Cumhurbaşkanı'nca imzalanıncaya kadar eski başbakan yani Ecevit görevine devam eder. (20370000)
Until the new government is signed by the president, the ex-prime minister, namely Ecevit, will continue to serve.

There are lexicalized expressions in which *–IncA* occurs. One of these lexicalized expressions has the meaning of ‘as for…’ and it doesn’t denote an abstract object since there is no event, fact or state.

(103) [Cevat ağabeye gelince], o halam evlendiğinde beş yaşında bir çocukmuş. (00019131)

(104) [As for Brother Cevat], he was a five years old child when my aunt got married.

-(y)IncA also can take place in adverbial expressions. For instance, *zamanı gelince* ‘in due time’ is a sentential adverb in (105). Note that it can be *Simplex subordinator* in other contexts.

(105) **Zamanı geli**nce *bütün yeşillerin arasında yeşerecek, bütün sarıların içinde sararacaktır.* (00035220)
In due time, it will turn green smong all the green and then will grow yellow among all the yellow.

In annotation samples, there are also erroneous instances such as *uyarınca* ‘according to’ which is a postposition, not a converbial form.

(106) İlk hareketten sonra, her şey artık [doğa yasaları uyarınca] cereyan eder. (00016112)
After first movement, everything happens [according to the laws of the nature].

Annotation Results and Disagreements

Of the total 150 instances annotated, 119 were annotated as DC, 1 was annotated as *Non-Converb*, and 19 were annotated as *Other* by both annotators. The remaining 11 were disagreements. The inter-annotator reliability for *–IncA* annotations is Kappa = 0.892 (p <0.05). The converb *–IncA* is also less *Ambiguous*. Most of the NDC examples come from idiomatic usages of *gelince* ‘as for’ and *uyarınca* ‘according to’.

The *complex subordinator* role of *–IncA* can be identified by checking the *–yA* suffix and the postposition after converb. Otherwise, *-IncA* creates converbs which are *Simplex subordinators* or lexicalized items. The lexicalized forms of this converb can be spotted from the morpho-syntactic features at the left edge of the converb. In such cases, converb with *–IncA* follows a nominal phrase which ends with the dative suffix and doesn’t signify a discourse relation. Since *–IncA* cases can easily be disambiguated by annotators with high agreement (Kappa = 0,892), they are less ambiguous and can be automatically disambiguated by using morpho-syntactic features.

5.10 –Ip

-Ip is a converbial suffix which has a conjunctive function.

Except its lexicalized and idiomatic usages (*Other*), converbs with *-Ip* are *Simplex subordinators* (DC).

When the suffix *-Ip* is a *simplex subordinator*, it has the meaning of ‘and’ and creates discourse relation with EXPANSION sense.

(107) Hemen **hazırlan**ıp *arabaya bindi.* (00001231)
He/she got ready quickly and got into the car.

-Ip is a frequent conjunctive suffix which occurs as parts of a lexicalized expressions and verb phrases.

(108) Kırgınlıkların, korkuların eriyip gidecekti, hepsi benim olacak, bana geçecekti. (00001131)
All your resentments and fears would *melt away*, and all of them would pass to me and would be mine.

-ıp durmak in (109), is also a lexicalized form. In such cases *durmak* 'to stand' means that event of the preceding converb happens continuously.

(109) gece rüyamda Zübeyde'yi gördüm, yüzü yoktu ya da vardı; ama ben bir türlü seçemiyordum, beyaz bir duman vardı yüzünün olduğu yerde ve [kımıldanıp duruyordu], bedeni ise dolgun ve etliydi, onu görüp sarılabiliyordum, acı yeşil bir elbise giymişti, omuzları bembeyaz ve yuvarlaktı. (00047224)
That night Zübeyde was in my dream; there was or wasn't her face; but I couldn't perceive it somehow; there was a white smoke where her face was, and it [kept stirring], whereas her body was plump and fleshy, I could see her and hug her; she was wearing a green chilli dress; her shoulders were snow-white and round.

Converb in **Hata! Başvuru kaynağı bulunamadı.** is used in reduplication in which the event happens repeatedly. The converb *yolup* isn't annotated as *Simplex subordinator* even though the repeated converb creates an abstract object and in this case *attığı* 'those that they threw' becomes the second argument thus whole subordinate clause Onların dallardan **yolup yolup** *attığı* 'those that they plucking and throw from the branches' holds a discourse relation within itself. The first *yolup* in reduplication is not annotated as simplex subordinator.

(110) Onların dallardan yolup yolup attığı erikleri önlüğündeki torbaya dolduruyordu. (00032161)

She was filling the pocket on her apron with the plums they kept ripping and ripping and throwing of the branches.

Sometimes reduplication does not contain the same form of the verb twice, but the negative form follows the positive form, building an expression meaning "whether or not" as in **Hata! Başvuru kaynağı bulunamadı.** the first –Ip form is not an abstract object alone but reduplication with the factive inflection builds the expression *gönder-ip gönderemeyeceğimi* 'whether I could send or not' which is the object of the main verb *sormak* 'to ask'.

(111) Bir gün beni merak ettiğini, fotoğrafımı gönderip gönderemeyeceğimi sordu. (20420000)

One day he said that he was curious about me and asked whether I could send him a photo of him or not.

Annotation Results and Disagreements

Of the total 150 instances annotated, 106 were annotated as DC and 24 were annotated as *Other* by both annotators. The remaining 20 were disagreements. The inter-annotator reliability for *–Ip* annotations is Kappa= 0.646 (p <0.05) which is an acceptable agreement. The disagreements are due to the different interpretations of reduplication idioms such as *yanıp tutuşmak* 'to yearn for', *geçip gitmek* 'to go by', çekip çıkarmak 'to pull out' etc. These reduplications are agreed to be idiomatic expressions where the converb*–Ip* doesn't denote a separate event. Nevertheless, there are examples such as *al-ıp aktarmak* 'transfer' in which annotators cannot agree on any decision. Such cases are highly ambiguous since abstract object interpretation highly depends on the reader.

-Ip becomes *Simplex subordinators* and occurs in a variety of lexicalized items. *Simplex subordinators* can be distinguished by looking for reduplications made by *–Ip* and the collocations such as *yanıp tutuşmak* 'to yearn for', *dolup taşmak* 'to swarm', *çekip çıkarmak* 'to pull out', *sayıp dökmek* 'to recount' etc. Still there are hard cases in which annotators cannot agree on any decision.

> (112) Yüzlerce binlerce yıl öteden gelen türküleri olduğu gibi alıp aktarmaktan yana hiçbir zaman olmadım.
> I never stand up for transferring songs that comes from hundreds and thousands years just as they are.

In (112), *alıp aktarmak* can be interpreted as two district event like *almak* 'take' and *aktarmak* 'transfer' or as a single event 'transfer'. This interpretation is completely subjective and depends on the reader's perception of event. Apart from such exceptionally hard cases, *-Ip* is an ambiguous converb which can disambiguated by using syntactic features.

5.11 –(y)ken

-(y)ken is an adverbial suffix with 'while' meaning. The segment -(y) is the copula so *-(y)ken* attaches not directly to the verb stem, but instead to verbal suffix or to a nominal.

-(y)ken can be a *Simplex subordinators* (DC), or occur in other discourse markers or lexicalized expressions (*Other*).

Kurtul-ur-ken 'while surviving' in (113) is a simplex subordinator with the meaning 'while' and has a CONTINGENCY sense. The *simplex subordinators –(y)ken* is polysemous like *While* is a polysemous connective in English, as both may have both COMPARISON and TEMPORAL sense (114).

> (113) **Halim yaralı kurtulur**ken *Mesut orada can vermiş.* **(**00003221**)**
>
> **While Halim survived injured** *Mesut died there.*
>
> (114) *Bütün gece sizi* **uyur**ken *seyretti*. (00063160)
>
> *He/she watched you all night* **while you're sleeping**.

Der-ken 'while telling' is a lexicalized discourse adverbial in (115) which takes its second argument anaphorically. It means 'just then'.

(115) Derken, berber dükkânının on beş yirmi adım ötesinde, upuzun boyuyla camın arkasına dikilip köy alanını seyreden berbere bakarken buldu kendini. (00064211)
Just then, he found himself fifteen or twenty steps away from the barber shop, looking through the glass with his imploring height.

-(y)ken may appear in reduplications like other adverbial suffixes do. Similar to other reduplicated converbs, they create idiomatic expressions instead of simplex subordinators. For example, *durup dururken* in (116) means 'for no reason' that is quite different than its compositional meaning.

(116) Niçin durup dururken bir insanın kimliği, yaşamı, şu hayattaki konumu değiştirilsin? (00002113)
Why would the identity, life, and the status in this life of a be changed for no reason?

Note that *derken* can be used with its literal meaning 'while saying' as in (117), in which case it as annotated as a *simplex subordinator*.

(117) **"Tanrı bir matematikçi mi?" der**ken *bulduğu* ilişkilerle, kuramlarla Alman denizaltıların şifresini çözmeyi başarıyor. (10210000)

With the relations and the theories *he discovered* while **asking "is God a mathematician?"** he manages to solve the cipher of the German submarines.

Derken can also be used with the meaning of 'just after', 'just then' in a special context (118). These instances are annotated as *Other/Lexicalized* since it has no abstract objects as the first argument.

(118) Yemek, çay, kahve, tütün derken iyice samimi olmuşlardı. (00001231)

They become very friendly just after the food, tea, coffee, tobacco.

Annotation Results and Disagreements

Of the total 150 instances annotated, 130 were annotated as DC, and 8 were annotated as *Other* by both annotators. The remaining 12 instances were disagreements. The inter-annotator reliability for *–ken* annotations is Kappa= 0.536 (p <0.05) which shows relatively low agreement. The first reason for low agreement is the unbalanced distribution of DC-NDC in the samples: the number of DCs is 130 out of 150 instances. Secondly, the nominal with *–ken* such as *çocuk-ken* 'when … child' causes disagreement because of the different interpretations of the annotators.

Converbs with *–ken* becomes *Simplex subordinators* most of the time except when they are discourse adverbials or part of a reduplication. Reduplications and discourse adverbials can be eliminated by using syntactic features, thus, *-ken* is considered as *Ambiguous*.

5.12 –mAksIzIn

The morphology of this suffix is as follows:

(119) *-mAk-sIz-In*
–INF-NEG-PASS

-mAK forms verbal nouns and converbs and *-sIz* is the adjectival suffix meaning 'without'.– *mAksIzIn*, means 'without doing smtg', expressing manner negatively (G&K, 2008).

All instances of–*mAksIzIn* are *Manner* but some of them signify a discourse relation in addition to their manner roles.

If the converb joins an abstract object in a subordinate clause beyond the main clause, they are annotated as *simplex subordinators* (120).

(120) Bu şekilde **kamu maliyesine herhangi bir yük getiril**meksizin *devlet üniversitesi sayısının 53'ten 70'e çıkarılabileceği vurgulanıyor.* (10130000)
In this way, *it's emphasized that the number of the state universities can be increased from 53 to 70,* **without adding a burden to public finance**.

If the converb doesn't denote an abstract object then it's the manner for the verb of superordinate clause **Hata! Başvuru kaynağı bulunamadı.**.

(121) Işığın bütün vadiyi dalga dalga aşıp gitmesini, geçip gittiği yere hayatının bir parçasını verircesine silinip yok olmasını kıpırdamaksızın izlemiştim. (00030130)
I watched unmoving the light going over the whole valley, where it disappeared as if giving the part of its life to the places where it passed away.

Ol-maksızın 'without' is a special case of -mAksIzIn, where the converb follows any nominal but loses its event, fact or state meaning of 'being' (122)

(122) Gözlerini kısıp konuşmasını sürdürüyor: ``Belki de tutkunuzun kaynağını yardımım olmaksızın siz bulacaksınız.'' (00007121)
He squints and continues his speech: "Maybe you will find the source of your passion without my help."

Annotation Results and Disagreements

Of the total 48 instances annotated, 30 were annotated as DC, 3 were annotated as *Manner*, and 1 was annotated as *Other* by both annotators. The remaining 14 instances were disagreements. The inter-annotator reliability for *–mAksIzIn* annotations is Kappa= 0.278 (p <0.05) which shows quite low agreement. This is mostly due to the difficulty of *DC-Manner* disambiguation task and the small sample size.

5.13 –mAktAnsA

The morphology of this suffix is as follows:

(123) *-mAk-tAn-sA*
-INF-ABL-COND

Adverbial clauses marked with *-mAktAnsA* 'rather than' are used in sentences expressing preference (Göksel & Kerslake , 2005) and they become *Simplex subordinators*.

All of the instances of *-mAktAnsA* form discourse relations with EXPANSION:Alternative:chosen alternative.

(124) **Aylık al**maktansa *toplu para alıp holdinge yatırmış.* (20340000)

Instead of receiving monthly, *he/she took all the money and invested in holding company.*

Annotation Results and Disagreements

All 7 instances of *–mAktAnsA* were annotated as DC by both annotators. There is no disagreement. All the examples of converbs made by *–mAktAnsA* create Simplex subordinators hence they are considered as unambiguous.

5.14 –mIşcAsInA

The morphology of this suffix is as follows:

–mIşcAsInA 'as if' express manner by evoking similarity with another, purely imagined action by the same subject, or by suggesting an underlying motivation or emotion (G&K, 2008).

Mostly, *-mIşcAsInA* creates adverbial clauses which is tied to another subordinate clause or the main clause in which they become manners.

All instances of *–mIşcAsInA* are *simplex subordinators* in our sample, and they depict distinct events from superordinate clauses as in (125).

(125) Ve bacaklarım **taş bağlan**mışçasına *ağırlaştı*; koşamaz, neredeyse yürüyemez oldum. (00007221)

And *my legs went heavy* **as if stones were tied to them**; I became unable to run, or even walk.

Annotation Results and Disagreements

All 20 instances of *–mIşcAsInA* were annotated as DC by both annotators. There is no disagreement. *–mIşcAsInA* depicts very similar morphological form to *–ArcAsInA,* but all the instances of it are *Simplex subordinators* and no *Manner* instance is found. *–mIşcAsInA* is expected to show same ambiguity cases with *–ArcAsInA* due to similar morphological structure, so a larger data set is necessary in order to capture all possibilities. Because of its similarity to *–ArcAsInA*, *–mIşcAsInA* is considered highly ambiguous.

5.15 –sA

-sA suffix can be used as like volitional modality, conditional suffix or in deliberative questions.

-sA is a frequent suffix which can also be seen in the independent word form of *ise*. We are interested in only *–sA* which is bound to a verb or nominal and construct adverbial clauses. *–sA* has multiple roles within discourse like other frequent suffixes.

When it is a *simplex subordinatior –sA* signifies discourse relations with CONTINGENCY sense (126)

(126) ``**Polis de, herhangi bir bilgi** <u>**ister**</u>se, *bunları söyleyebilirim* onlara. (00006231)
<u>If</u> **police** <u>**wants**</u> **any information too**, *I can tell these to them.*

When *–sA* attaches to some nominals, it may form discourse adverbials rather than a *simplex subordinator* (127).

(127) <u>Nedense</u> bir annen olduğunu hiç düşünmemişim... (00005221)
<u>Somehow</u>, I never thought you would have a mother…

-sA may occur in coordinating conjunctives when it is used as *yoksa* 'or' as in **Hata! Başvuru kaynağı bulunamadı.**

(128) Ben mi yanlış ya da yetersiz düşünüyorum, <u>yoksa</u> bu işte bir tuhaflık mı var, bilmem. (00054223)

I don't know whether I my thoughts are wrong or insufficient, or there is something fishy about this business.

There are idiomatic expressions in which *–sA* takes place. Since idioms have different meanings than the compositional meaning of its parts, converbs in such idioms are not considered as *Simplex subordinators*. For instance, *kısmet ol-ur-sa* is an idiomatic expression that means 'hopefully' in (129)

(129) Şimdi kısmet <u>olursa</u> ANAP kongresinden sonra 'üçüncü adım'ı atacağız. (20540000)
Now, hopefully, after the ANAP congress we will take the 'third step'.

If *–sA* attaches to pronouns, it may become a focus particle as in (130).

(130) <u>Bense</u> silaha karşıydım, beni öldürseler bile insanları konuşarak ikna etme taraftarıyım. (00057221)
<u>As for me</u>, I was against guns, I believe in convincing people through dialog even if they would kill me.

Converbs with *–sA* may appear in lexicalized expressions where a discourse relation is established by the lexical items. In **Hata! Başvuru kaynağı bulunamadı.**, *Nerede … -sA* is the lexical item which holds relation. Since converb with *–sA* doesn't signify the discourse relations by itself, instances like (131), (132), and (133)are annotated as *Other DC*

(131) Nerede bir kurtarma kazısı <u>varsa</u>, oraya gittim. (00013112)
<u>Wherever there was</u> a rescue excavation, I went there.

(132) Islık kime <u>çalınmışsa</u> o koşardı. (00032161)
Whoever the whistle was blown for, that one would run.

(133) **Nasıl ki inananlar, Allah'ın hikmetinden sual** <u>**edemez**</u>se, *parti üyeleri de liderlerinin tasarruf ve takdirlerini sorgulayamazlar, ona boyun eğerler!..*
Just as the believers do not question the Judgment of God, *the members of the party cannot question the will and the way of their leader, and just submit to him.*

Annotation Results and Disagreements

Of the total 150 instances annotated, 88 were annotated as DC, 2 were annotated as *Non-Converb*, and 41 were annotated as *Other* by both annotators. The remaining 19 were disagreements. The calculcated inter-annotator agreement of *–sA* is acceptable with Kappa = 0,769 (p <0,05). The disagreement is mostly between *DC-Other* annotations and these are mostly lexicalized discourse relation markers.

–sA suffix can be focus particle and there are such instances within samples set due to morphological disambiguator errors. Such cases can be eliminated by looking the root of the converb since focus particles attach to nominal.

Additionally, *–sA* can be *Simplex subordinators* or creates other kinds of discourse relations. Other discourse relations can be discourse adverbials or coordinating conjunctions such as *yoksa* 'otherwise', 'or' and conventional constructions such as nerede … *-sA* in (131).

The ambiguity between *Simplex subordinators* and other discourse connectives can be disambiguated by the syntactic class of verb and looking for conventional constructions. Therefore, *–sA* can be disambiguated by human annotators with acceptable agreement and can be automatically annotated by using syntactic features.

CHAPTER 6

Disambiguation Studies in Discourse

In this chapter, two types of ambiguity in discourse are explained and their resolution methods are surveyed. These are the identification of discourse connectives in discourse and the automatic disambiguation of the connectives' senses. Then, some of these methods are used in a case study at the end of this chapter.

6.1 Identification of Discourse Relations

In PDTB, the explicit discourse connectives are largely unambiguous, such as *although* and *additionally*, which are almost always used as discourse connectives and the senses of the relations they signal are unambiguously identified as COMPARISON and EXPANSION, respectively. However, not all discourse connectives have these desirable properties. A discourse relation marker can be ambiguous between its discourse and non-discourse use. For example, *once* can be either a temporal discourse connective or simply a sentential adverb meaning "formerly" (Pitler & Nenkova, 2009).

Only 11 of the 100 connectives in the PDTB appear as a discourse connective more than 90% of the time. These connectives are *although*, *in turn*, *afterward*, *consequently*, *additionally*, *alternatively*, *whereas*, *on the contrary, if* and *when*, *lest*, and *on the one hand...on the other hand*. For example, *although* acts as a discourse connective 91.4% of the time while *or* only serves a discourse function 2.8% of the time (Pitler & Nenkova, 2009).

Emily Pitler and Ani Nenkova demonstrate that even using the string of the connective as the only feature creates a reasonably high baseline, with an f-score of 75.33% and an accuracy of 85.86% (Pitler & Nenkova, 2009). In order to train a maximum entropy classifier to differentiate the discourse vs. non-discourse use, the explicit discourse connectives annotated in the PDTB are used as positive examples and occurrences of the same strings in the PDTB texts that were not annotated as explicit connectives are used as negative examples. They report that using only the syntactic features and ignoring the identity of the connective results in an f-score of 88.19% and accuracy of 92.25%. Using both the connective and syntactic features is better than using either individually, with an f-score of 92.28% and accuracy of 95.04%. In this study, the syntactic features used are *Self Category* of the connective, which can be part of speech tag of the word; *Left Sibling Category,* which is immediately to the left of the Self Category; *Right Sibling Category,* which is immediately to the right of the Self Category; and *Parent Category,* which is the immediate parent of the Self Category. And the results for this Discourse vs. Non-discourse Disambiguation task are given in Table 6-1.

Features	Accuracy	f-score
(1) Connective Only	85.86	75.33
(2) Syntax Only	92.25	88.19
(3) Connective+Syntax	95.04	92.28
(3)+Conn-Syn Interaction	95.99	93.63
(3)+Conn-Syn+Syn-Syn Interaction	**96.26**	**94.19**

Table 6-1 Discourse versus Non-discourse Usage **(Pitler & Nenkova, 2009, p. 15)**

As the table illustrates, using the connective and the syntactic features together results better than their individual uses. They also argue that different connectives have different syntactic contexts for their discourse use, for example, features like “connective=also - RightSibling=SBAR” raised the f-score about 1.5%, to 93.63%. Last raw of the table shows the slight increase of the f-score to 94.19% after adding the interaction terms between pairs of syntactic features.

6.2 Disambiguation of Senses

PDTB provides sense annotations for all discourse connectives in PDTB 2.0 since discourse connectives can have more than one sense, just like verbs, depending on the context. Despite the fact that some of the discourse connectives always occur with just one of the senses (for example, *because* is almost always a CONTINGENCY), some others are quite ambiguous. For example, *since* appears with three different senses; one purely TEMPORAL in Example (134), another purely CONTINGENCY:Causal in Example (135) and a third both CONTINGENCY:CAusal and TEMPORAL in Example (136).

(134) *The Mountain View, Calif., company has been receiving 1,000 calls a day about the product* since **it was demonstrated at a computer publishing conference several weeks ago**. (Prasad, et al., 2008, p. 4)

(135) *It was a far safer deal for lenders* since **NWA had a healthier cash flow and more collateral on hand.** (Prasad, et al., 2008, p. 4)

(136) *Domestic car sales have plunged 19%* since **the Big Three ended many of their programs Sept. 30.** (Prasad, et al., 2008, p. 4)

Below, Table 6-2 shows the list of top polysemous connectives with their multiple senses in PDTB (Prasad, et al., 2008).

Connective	**Senses**
after	succession (523), succession-reason (50), other (4)
since	reason (94), succession (78), succession-reason (10), other (2)
when	Synchrony (477), succession (157), general (100), succession-reason (65), Synchrony-general (50), Synchrony-reason (39), hypothetical (11), implicit assertion (11), Synchrony-hypothetical (10), other (69)
while	juxtaposition (182), Synchrony (154), Contrast (120), expectation (79), opposition (78), Conjunction (39), Synchrony-juxtaposition (26), Synchrony-Conjunction (21), Synchrony-Contrast(22), COMPARISON (18), Synchrony-opposition (11), other (31)
meanwhile	Synchrony-Conjunction (92), Synchrony (26), Conjunction (25), Synchrony-juxtaposition (15), other(35)
but	Contrast (1609), juxtaposition (636), contra-expectation (494), COMPARISTON (260), opposition (174), Conjunction (63), Conjunction-Pragmatic contrast (14), Pragmatic-contrast (14), other (32)
however	Contrast (254), juxtaposition (89), contra-expectation (70), COMPARISON (49), opposition (31), other (12)
although	expectation (132), Contrast (114) juxtaposition (34), contra-expectation (21), COMPARISON (16), opposition (9), other (2) and Conjunction (2543), List (210), result-Conjunction (138), result (38), precedence-Conjunction (30), juxtaposition (11), other(30)
if	hypothetical (682), general (175), unreal present (122), factual present (73), unreal past (53), expectation (34), implicit assertion (29), relevance (20), other (31)

Table 6-2 Top ten polysemous connectives (Explicit) **(Prasad, et al., 2008, p. 6)**

Miltsakaki et al. show that using syntactic features and a simple Maximum Entropy (MaxEnt) model can achieve some success in automatically disambiguating among the connective's senses. They used three polysemous connectives: *since* which has temporal, causal, temporal/causal senses; *while* which has temporal, as well as all three contrastive senses – comparison, opposition and concession; and *when* with a purely temporal sense, a simultaneously temporal and causal sense, a conditional sense and a concessive sense. In order to train MaxEnt model, they give a four-dimensional vector with the following features (Miltsakaki, Dinesh, Prasad, Joshi, & Webber, 2005):

1. *Form of auxiliary have - Has, Have, Had or Not Found.*
2. *Form of auxiliary be - Present(am, is, are), Past (was, were), Been, or Not Found.*
3. *Form of the head - Present (part-of-speech VBP or VBZ), Past (VBD), Past Participal (VBN), Present Participal (VBG).*
4. *Presence of a modal - Found or Not Found. The number of instances with a modal tense was few, so distinguishing between the various kinds of modals did not aid in increasing accuracy.*

Accordingly, a sentence like "*He has been going to the mall."* would be assigned the vector [Has, Been, HeadPresentParticipal, ModalNotFound], and the sentence "*He had gone to the mall."* would be assigned the vector [Had, BeNotFound, HeadPastParticipal, ModalNotFound]. The results show that these features aid in distinguishing the temporal from the causal sense for the connective ***Since*** (Table 6-3).

Experiment	Accuracy
(T,C,T/C)	75.5% (53.6%)
({T,T/C}, C)	90.1% (53.6%)
(T,{C,T/C})	74.2% (65.6%)
(T,C)	89.5% (60.9%)

Table 6-3 Average accuracy of sense disambiguation in 10-fold cross validation for *since*. T stands for Temporal, C for Causal, and T/C for Temporal/Causal. Accuracy of the baseline (predict most frequent sense) is parenthesized. **(Miltsakaki, Dinesh, Prasad, Joshi, & Webber, 2005, p. 9)**

In addition to the features described above, a few additional features were added specific to ***while*** such as the relative position of Arg2 to Arg1. The two other features were the presence of same verb in both arguments and the adverb not present in the head verb phrase of a single argument. These are used to distinguish between the comparative and concessive senses. Table 6-4 shows such a correlation of these features with senses.

Feature	T	Con	Comp	Opp
Preposed	0.1%	37.4%	0%	62.5%
Interposed	0%	75%	0%	25%
Arg2 Non-Finite Participal	73.3%	6.7%	0%	20%
Same verb	2.5%	0%	62.5%	25%
Single not Arg	0%	62.5%	0%	27.5%

Table 6-4 Co-occurrence of a feature with a sense for while (Miltsakaki, Dinesh, Prasad, Joshi, & Webber, 2005, p. 11)

The same tense vector and the explicit time feature are used for the connective ***when***. The classifier was able to differentiate the *temporal* senses from *conditional* senses, but not good at distinguishing between the *temporal* and *temporal/causal* senses. The results for *when* are given in Table 6-5.

Experiment	Accuracy
(T,T/C,Cond)	61.6% (47.6%)
(T,{T/C,Cond})	50% (52.3%)
({T,T/C},Cond)	82.6% (69.1%)

Table 6-5 Average accuracy of sense disambiguation in 10-fold cross validation for *when*. T stands for Temporal, T/C for Temporal/Causal, and Cond for Conditional. Accuracy of the baseline (predict most frequent sense) is parenthesized. **(Miltsakaki, Dinesh, Prasad, Joshi, & Webber, 2005, p. 11)**

The features used in this study may not be applicable across genres yet an improvement of 15-20% over the baseline was seen across the board (Miltsakaki, Dinesh, Prasad, Joshi, & Webber, 2005).

Pitler et al. shows that while there is a large degree of ambiguity in temporal explicit discourse connectives in PDTB, overall connectives are mostly unambiguous and allow high-accuracy prediction of discourse relation type (Pitler, Raghupathy, Mehta, Nenkova, Lee, & Joshi, 2008). According to the study, most of the comparison and temporal relations are explicitly marked and discourse connectives are mostly unambiguous. Based on these facts, they suggest that even based only on the connective, classification of discourse relations could be done for all data, particularly well for explicit examples alone.

Class	Explicit (%)	Implicit (%)	Total
Comparison	5590 (69.05%)	2505 (30.95%)	8095
Contingency	3741 (46.75%)	4261 (53.25%)	8002
Temporal	3696 (79.55%)	950 (20.45%)	4646

Expansion	6431 (42.04%)	8868 (57.96%)	15299

Table 6-6 Discourse relation distribution in semantic and explicit/implicit classes in the PDTB (Pitler, Raghupathy, Mehta, Nenkova, Lee, & Joshi, 2008, p. 2)

According to Table 6-6, temporal and comparison relations are predominantly explicit, the contingency relations are almost evenly distributed between explicit and implicit, and the expansion relations are implicit generally. Just based on this analysis, they build a decision tree classifier by using connective itself as binary features. Results are given in Table 6-7.

Task	All relations	Explicit relations only
Comparison	91.28% (76.54%)	97.23% (69.72%)
Contingency	84.44% (76.81%)	93.99% (79.73%)
Temporal	94.79% (86.54%)	95.4% (79.98%)
Expansion	77.51% (55.67%)	97.61% (65.16%)

Table 6-7 Decision tree classification accuracy using only the presence of connectives as binary features. The majority class is given in brackets. (Pitler, Raghupathy, Mehta, Nenkova, Lee, & Joshi, 2008, p. 2)

Four disambiguation task settings are prepared while training the classifier so that each type of relation is distinguished from all others. For example, comparison relations can be distinguished from all other relations in the corpus with overall accuracy of 91.28%, based only on the discourse connective.

They additionally suggest that global sequence classification of the relations in text can lead to better results, especially for implicit relations. For instance, explicit comparison and implicit contingency co-occur much more often than expected thus when there is an explicit comparisons relation it is more likely to find an implicit contingency relations in the text (Pitler, Raghupathy, Mehta, Nenkova, Lee, & Joshi, 2008).

In the study of Emily Pitler and Ani Nenkova (2009), syntactic features such as *Self Category* of the connective, *Left Sibling Category*; *Right Sibling Category*; and *Parent Category* are used to identify discourse relations. In this study, they also demonstrate that the same syntactic features improve performance in disambiguation among the senses of a discourse connective. In their experiments they consider only the top level categories: Expansion, Comparison, Contingency, and Temporal because the top-level senses are general enough to be annotated with high inter-annotator agreement and they are common to most theories of discourse (Pitler & Nenkova, 2009).

They use syntactic features and string of the connective to train Naïve Bayes classifier and report 94% accuracy which is the human inter-annotator agreement on the top level sense class also. Results for this Naïve Bayes classifier are given in Table 6-8:

Features	**Accuracy**
Connective Only	93.67
Connective+Syntax+Conn-Syn	**94.15**
Interannotator agreement on sense class (Prasad et al., 2008)	94

Table 6-8 Four-way sense classification of explicits (Pitler & Nenkova, 2009, p. 16)

In addition to Pitler and Nenkova's study, which reports results only for the topmost (Class) level of the PDTB's senses, Versley proposes that it's possible to build classifiers to disambiguate senses of discourse connectives with finer distinction namely at types and subtypes levels defined in PDTB (Yannick, 2011). He proposes a hierarchical classification that will allow the forecast of finer classes while making use of the taxonomical information contained in the PDTB's hierarchical label set. For instance, the topmost classifier would classify the relation as *Temporal*, and then the second-level classifier for *Temporal* would determine that the relation is *Temporal.Asynchronous*, and the third-level classifier for *Temporal. Asynchronous* would choose *Temporal.Asynchronous.Precedence* as the finest-level relation (Yannick, 2011).

Versley uses similar syntactic features to the ones Pitler and Nenkova used. After the correct identification of the arguments in PDTB, he extracts the following indicators:

- *the part-of-speech of the first non-modal verb in the sentence (descending from the argument clause node into further VP and S nodes to cover both nesting of VPs and coordinated sentences)*
- *the presence (and word form) of modals and negation in the clause*
- *a tuple of (have-form, be-form, head-POS, modal present) as proposed by Miltsakaki et al. (2005).*

The results of the classifiers with different features are given in the Table 6-9. Versley reports that syntactic features, including the function tags and the inclusion of Arg1-related verb features, yield improvements over the version in which the connective itself is the only feature.

	d=1	d=2	d=3
hierarchical			
connective only	0.946	0.839	0.790
conn+syntaxA	0.954	0.847	0.796
conn+syntaxB	0.945	0.840	0.788
w/traces	0.948	0.843	0.792
w/function tags	0.954	0.847	0.796
conn+verb(arg1)	0.952	0.845	0.798
conn+synB+pos(arg1)	0.949	0.843	0.794
conn+pos(both)	0.949	0.843	0.794
conn+synB+pos(both)	0.947	0.839	0.788
greedy			
connective only	0.946	0.840	0.792
conn+syntaxA	0.955	0.847	0.798
conn+verb(arg1)	0.953	0.845	0.800

Table 6-9 Different versions of syntactic and tense/mood features **(Yannick, 2011, p. 152)**

On the other hand, the inclusion of tense information cannot improve over the information contained in the function tags but the incorporation of tense/mood information on the heuristically determined ARG1 yields useful results by itself (Yannick, 2011).

All studies mentioned so far use strings of connectives as the most reliable feature of the semantic sense of the discourse relation. However, in the absence of explicit connective words, Pitler et al. seek other features from the words of two arguments since they expect some relationship between the words in the two spans. For example, in the following example:

> (137) *The recent explosion of country funds mirrors the "closed end fund mania" of the 1920s, Mr. Foot says, when narrowly focused funds grew wildly* ***popular****. They fell into* ***oblivion*** *after the 1929 crash.* (Pitler, Louis, & Nenkova, 2009, p. 685)

The words *popular* and *oblivion* are almost antonyms and can trigger the contrast relation between the sentences. They use a large collection of automatically extracted explicit examples to find useful features from word pairs. As a result, their study finds the following features as informative (Pitler, Louis, & Nenkova, 2009, pp. 686-688):

- **Polarity Tags:** In this resource, each sentiment word is annotated as positive, negative, both, or neutral. The number of negated and non-negated positive, negative, and neutral sentiment words in the two text spans as taken as features.
- **Inquirer Tags:** To get at the meanings of the spans, they look up what semantic categories each word falls into according to the General Inquirer lexicon (Stone et al., 1966). Inquirer Tags have more fine-grained categories such as virtue or vice.
- **Money/Percent/Num:** If two adjacent sentences both contain numbers, dollar amounts, or percentages, it is likely that a comparison relation might hold between the sentences.
- **Verbs:** These features include the number of pairs of verbs in Arg1 and Arg2 from the same verb class. Two verbs are from the same verb class if each of their highest Levin verb class (Levin, 1993) levels (in the LCS Database (Dorr, 2001)) are the same.
- **First-Last:** The first and last words of a relation's arguments have been found to be particularly useful for predicting its sense.
- **Modality:** Modal words, such as "can", "should", and "may", are often used to express conditional statements.
- **Context:** Some implicit relations appear immediately before or immediately after certain explicit relations far more often than one would expect due to chance.

The results of the Naïve Bayes classifier with different features are reported in Table 6-10, where f-scores and accuracies are given in parenthesis and they run four binary classification tasks to identify each of the main relations from the rest (other).

Features	**Comp. vs. Not**	**Cont. vs. Other**	**Exp. vs. Other**	**Temp. vs. Other**	**Four-way**
Money/Percent/Num	19.04 (43.60)	18.78 (56.27)	22.01 (41.37)	10.40 (23.05)	(63.38)
Polarity Tags	16.63 (55.22)	19.82 (76.63)	71.29 (59.23)	11.12 (18.12)	(65.19)
WSJ-LM	18.04 (9.91)	0.00 (80.89)	0.00 (35.26)	10.22 (5.38)	(65.26)
Expl-LM	18.04 (9.91)	0.00 (80.89)	0.00 (35.26)	10.22 (5.38)	(65.26)
Verbs	18.55 (26.19)	36.59 (62.44)	59.36 (52.53)	12.61 (41.63)	(65.33)
First-Last, First3	21.01 (52.59)	36.75 (59.09)	63.22 (56.99)	15.93 (61.20)	(65.40)
Inquirer tags	17.37 (43.8)	15.76 (77.54)	70.21 (58.04)	11.56 (37.69)	(62.21)
Modality	17.70 (17.6)	21.83 (76.95)	15.38 (37.89)	11.17 (27.91)	(65.33)

Context	19.32 (56.66)	29.55 (67.42)	67.77 (57.85)	12.34 (55.22)	(64.01)
Random	9.91	19.11	64.74	5.38	

Table 6-10 f-score accuracy using different features; Naive Bayes (Pitler, Louis, & Nenkova, 2009, p. 689)

Pitler et al. report that word pair features supply 6% to 18% improvements in f-score over the baseline for each of the four tasks. The best improvement is in the *Contingency versus Other* prediction task yet the least improvement is in distinguishing Expansion versus Other prediction. One interesting result is that polarity tags are actually one of the worst classes of features for Comparison, achieving an f-score of 16.33. In contrast to common expectation, ***Comparison*** *relations* do not tend to have more opposite polarity pairs. The first, last and first three words in the sentence is the two most useful features for recognizing Comparison relations. For ***Contingency*** relations, verb information is the best predictor. Polarity tags, Inquirer tags and context were the best features for identifying ***Expansion*** relations with f-scores around 70%. Since the temporal implicit relation often contain words like "yesterday" or "Monday" at the end of the sentence, the first and last words of the sentence is a useful feature for ***Temporal*** relations. Therefore, the study affirms that different features fit best for different senses (Pitler, Louis, & Nenkova, 2009).

6.3 An Experiment with Simplex subordinators

Most of the subordinating conjunctives in Turkish take nominalized clauses as their second arguments and these nominalizations can have a variety of morphological features. Preliminary studies of TDB show that the morphological properties of the nominalized arguments allow a further degree of disambiguation for the sense of the connective (Demirşahin, Sevdik-Çallı, Balaban, Çakıcı, & Zeyrek, 2012). For example, *için '*for*'* can express relations with goal or cause sense and the sense of the relation can be disambiguated by simply looking at the morphology of the second argument. In (138),*–mek için* results in a goal driven relation by taking an infinitival clause as argument, and in (139) *-dığım için* results in a cause driven relation by taking a factive.

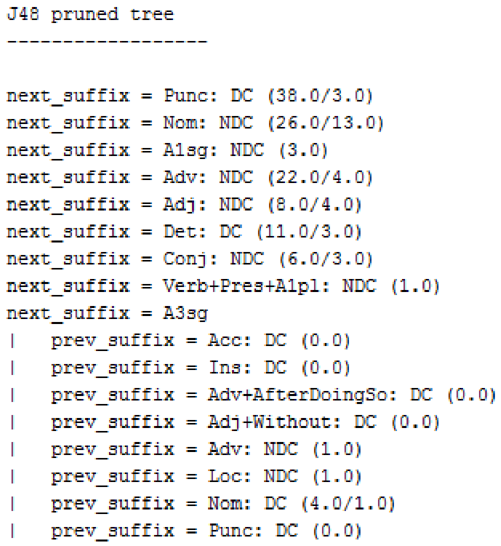
(138) **Onu görmek** <u>için</u> *tüm zamanınızı o parkta geçirmeye başlarsınız.*
<u>In order to</u> **see her** *you start to spend all your time in that park.*

(139) **Üvey babamı görmek istemediğim** <u>için</u> *yıllardır o eve gitmiyorum.*
<u>Since</u> **I don't want to see my step father**, *I haven't been to that house for years.*

Starting from these preliminary analyses on complex subordinators, we used machine learning algorithms to automatically disambiguate the different roles of the converbs by looking to the left and right morpho-syntactic context of the converb. The ambiguity of the converbs varies greatly, so we selected the converbs *–dIğIndAn*, *–dIğIndA* and *–ken* as the experimental candidates for the automatic annotation attempt, since they can be easily disambiguated by human annotators. Samples of these converbs were annotated manually to train the Decision Tree algorithm in Weka.

1. *-dIğIndAn* has 426 instances in the sentences but only 152 of them were appropriate for simple subordination, for the rest of the instances are followed by postpositions, building complex subordinators. 152 samples were annotated with DC (Discourse Connective) vs. NDC (Non-Discourse Connective) tags. The root and the rightmost suffixes of the words at the left and right of the converbs were taken as features. This feature set was used to train the Decision Tree J48 algorithm in Weka. Part of the decision tree output is given as Figure

6-1. It shows that converbs followed by **Punc** (Punctuation) appears to be DC generally, on the other side; it appears to be NDC if it's followed by a **Nom** (Nominalization) tag.



```
J48 pruned tree
------------------

next_suffix = Punc: DC (38.0/3.0)
next_suffix = Nom: NDC (26.0/13.0)
next_suffix = A1sg: NDC (3.0)
next_suffix = Adv: NDC (22.0/4.0)
next_suffix = Adj: NDC (8.0/4.0)
next_suffix = Det: DC (11.0/3.0)
next_suffix = Conj: NDC (6.0/3.0)
next_suffix = Verb+Pres+A1pl: NDC (1.0)
next_suffix = A3sg
|   prev_suffix = Acc: DC (0.0)
|   prev_suffix = Ins: DC (0.0)
|   prev_suffix = Adv+AfterDoingSo: DC (0.0)
|   prev_suffix = Adj+Without: DC (0.0)
|   prev_suffix = Adv: NDC (1.0)
|   prev_suffix = Loc: NDC (1.0)
|   prev_suffix = Nom: DC (4.0/1.0)
|   prev_suffix = Punc: DC (0.0)
```

Figure 6-1 Decision tree output of weka

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.362	0.188	0.538	0.362	0.433	0.715	NDC
	0.813	0.638	0.678	0.813	0.739	0.715	DC
Weighted Avg.	**0.643**	**0.468**	**0.626**	**0.643**	**0.624**	**0.715**	

Table 6-11 Weka output for decision tree

Table 6-11 gives the detailed accuracy scores by class names. F-measure is small and other accuracy values are not satisfying either. Scrambled sentences, errors in disambiguation results, and the small sample size are possible reasons for the low scores. Scrambled sentences were frequent in the samples, and as Eryiğit indicates, even the Perceptron is reported to have 96% accuracy, with their calculated accuracy on METU-Sabancı Treebank at 84% (Eryiğit, 2012). We believe that this experiment could to be repeated with some refinements such as using gold standard morphologic parses, finding solutions for scrambled sentences and using more samples from Metu Turkish Corpus for better results.

2.–*dIğIndA* has 502 instances. Only 6 of them have non-discourse roles, 2 of them are noun phrases rather than a converb, and the rest are discourse connectives. Therefore, *–dIğIndA* was found to be inappropriate for automatic annotation with supervised learning.

3. -*ken* samples were annotated for their discourse vs. non-discourse roles. 236 instances out of 255 are *simplex subordinators* and all non-discourse instances are idiomatic uses such as *derken* 'at that moment', *durup dururken* 'out of nowhere' etc. Therefore, *–ken* was also considered unambiguous in terms of its discourse role, so was inappropriate for automatic annotation with supervised learning.

CHAPTER 7

Conclusion and Discussions

This thesis examined the discourse connective role of converbs along with its other roles such as *Complement*, *Adverbial*, *Other* DC, *Lexicalized Expressions* etc. In order to shed light on the problem, an annotation procedure was performed with two annotators and agreement statistics were calculated. Some inferences about the disambiguation task could be made using the agreement statistics and the disagreement analyses; and possible implications of the study were presented. This chapter summarizes and discusses the findings from the annotation results. Then it explains the contributions of the thesis and possibilities for future studies.

7.1 Discussion

After ınterpreting the results of our analyses, we conclude that we can categorize the converbs presented in this study into three in terms of their degree of ambiguity.

The first category is the **Unambiguous Converbs.** These converbs, including *–mAktAnsA* and *–AlI,* are unambiguous, since all instance of them in TDB texts build only *simplex subordinators*. Obviously this claim should be tested with larger and more representative, and preferably multimodal data to achieve a more generalized conclusion about the converb; however, we can confidently say that all instances of these converbs in TDB are simplex subordinators.

The next category is the **Ambiguous Converbs.** The converbs which include *–AcAğInA, –AcAğIndAn, –dIğIndA,* and *–dIğIndAn*, can take place in *complements* for factive verbs as well as *simplex subordinators*. They can be disambiguated by using syntactic features which is also used by Pitler and Nenkova (Pitler & Nenkova, 2009).

Another group of ambiguous converbs including *–dIkçA, –IncA, –Ip, –ken,* and *–sA* are ambiguous since they can be either *simplex subordinators* or lexicalized expressions. They can be disambiguated by the syntactic features and by checking the degree of conventionality of the lexicalized expressions. Once again, larger and more representative data would be of use to extract the conventionality of these lexicalized expressions.

The final category is the **Hard Cases.** The *Manner* converbs such as *–ArAk, –ArcAsInA, –mAksIzIn,* and *–mIşcAsInA* comprise most of the hard cases because of the fact that the *DC* and the *Manner* roles have syntactically similar structures. Nevertheless, the syntactic components of the subordinate clause and the semantic relation between the converb and the matrix verb can help the disambiguation of these converbs to a degree.

The studies conducted in the scope of this thesis have also revealed some information about the nature of converbs which are likely to contribute to further studies.

Firstly, we found out that the factive verbs and syntactic trees are essential for automatic disambiguation task for all ambiguous converbs.

We observed that the converbs that act as *Complements* and *Manners* depict similar features and ambiguities among themselves. This is an indicator of how syntax is vital in disambiguation of converbs.

We also discovered that the frequent converbial suffixes occur in a variety of reduplications and other lexicalized constructions such as collocations, idioms and conventional constructions. In order to differentiate abstract object interpretations of the converbs from other cases, lexical semantics terms could be employed efficiently in order to measure compositionality and the degree of conventionality. The adverbial clauses denote abstract objects, as long as they keep the compositional meaning from their components. Therefore, free composition, collocations and fixed expressions are more likely to be interpreted as abstract objects, whereas idioms and lexicalized combinations are less likely.

We believe that the verb classes of Beth Levin and the Eventuality Types of Zeno Vendler can and should be utilized for Turkish to check semantic relations between the converb and the matrix verb as a feature to disambiguate *Manners* and *Simplex subordinators*.

During the annotation and the following analyses, we realized that the *simplex subordinators* can take anaphoric arguments, especially in anacoluthon and incomplete sentences. This may makes the automatic argument annotation task for *simplex subordinators* harder than it seems.

Thought the study, we came to believe that the ambiguity of the converbs is an important issue for machine translation. Translations of the converb examples depict the effect of the different roles of the converbs on translation of a sentence.

Finally, we discovered that the conventional construction of *–sA* examples can signify a variety of types of discourse relations. This seems to support the idea that the Discourse Relation Markers (DRMs) are a lexically open-ended class, which may or may not belong to well-defined syntactic classes such as conjunctions, prepositional phrases, subordinators etc. (Prasad, Joshi, & Webber, 2010)

7.2 Contributions

1. As part of the thesis study, TDB files were optimized: the relation attributes were separated into proper tags that allow making finer grained search within TDB; and the files are reorganized so that the annotation files of TDB are now parallel to the raw text files of TDB extracted from MTC. This makes TDB becomes more portable, and also more compatible with MTC, which will benefit researchers who's would like to conduct related studies on both corpora.

2. During the study for this thesis, some converbs that were not mentioned in the preliminary studies about the converbs in TDB such as *–sA, -AcAğInA, -AcAğIndAn* were brought to light and examined in detail.
3. This thesis promotes future studies of converbs by setting down certain principles and methods for the annotation of ambiguous converbs by human annotators.
4. This thesis also promotes automatic disambiguation studies by examining ambiguous cases and implementing some basis in terms of the methodologies and the tools for the disambiguation task.

7.3 Future Studies

As a future work, a comprehensive study can be conducted to examine sense ambiguity of all discourse connectives along with their DC-NDC ambiguity in Turkish.

Psycholinguistic experiments can be set up in order to understand a number of research questions such as:

- How are the reduplications interpreted and perceived by the native speakers? Specifically, do they perceive reduplications as a single event or two discrete events?
- Does the distance between the converb and the matrix predicate affect the abstract object interpretation of the converbs?
- How does the degree of conventionality change the perception of the abstract objects by the native speakers? Is this phenomenon specific to some of the converbs we examined, such as *–ArAk*, or is it a widespread phenomenon?

And finally, a lexicon consisting of lexicalized items build by converbs or an idiom bank can be created for Turkish, thus creating a valuable input for a variety of machine learning tasks, including but not limited to further disambiguation studies for converbs.

REFERENCES

Asher, N. (1993). Reference to Abstract Objects in Discourse. Dordrecht, Netherlands: Kluwer.

Demirşahin, I., Sevdik-Çallı, A., Balaban, H., Çakıcı, R., & Zeyrek, D. (2012, May 21). Turkish Discourse Bank: Ongoing Developments. *Proceedings of First Workshop on Language Resources and Technologies for Turkic Languages. (LREC'12)*. İstanbul, Turkey.

Eryiğit, G. (2012). The Impact of Automatic Morphological Analysis & Disambiguation on Dependency Parsing of Turkish. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012.* Istanbul.

Göksel, A., & Kerslake , C. (2005). *Turkish: a comprehensive grammar.* London; New York: Routledge.

Kamsties, E. (2001). Surfacing Ambiguity in Natural . Germany: FraunhoferInstitue für Experimentelles Software Engineering.

Kamsties, E. (2001). Surfacing Ambiguity in Natural Language Requirements. Kaiserslautern, Germany: Fraunhofer-Institue für Experimentelles Software Engineering.

Kingsbury, P., & Palmer, M. (2002). From Treebank to Propbank. *In: Third International Conference on Language Resources and Evaluation, LREC-2002*. Las Palmas, Canary Islands, Spain.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics 19, 313-330*. Cambridge, U.K: MIT Press.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics 19*, 313-330.

McEnery, T., & Wilson, A. (1996). Corpus Linguistics. Edinburgh : Edinburgh University Press.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A., & Webber, B. (2005). Experiments on Sense Annotations and Sense Disambiguation of Discourse Connectives. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005),.* Barcelona.

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2004). Annotating Discourse Connectives and their Arguments. *In Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*. Boston, MA.

Miltsakaki, E., Robaldo, L., Lee, A., & Joshi, A. (2008). Sense Annotation in the Penn Discourse Treebank. *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Vol 4919 pp 275-286.*

Pitler, E., & Nenkova, A. (2009). Using Syntax to Disambiguate Explicit Discourse Connectives in Text. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing.* Singapore.

Pitler, E., Louis, A., & Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing.* Singapore.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., & Joshi, A. (2008). Easily identifiable discourse relations. *Proceedings of COLING.* Manchester.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., et al. (2008). The Penn Discourse Treebank 2.0. *In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Prasad, R., Joshi, A., & Webber, B. (2010, August). Realization of Discourse Relations by Other Means: Alternative Lexicalizations. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

Prasad, R., Joshi, A., Dinesh, N., Lee, A., Miltsakaki, E., & Webber, B. (2005, July). The Penn Discourse TreeBank as a Resource for Natural Language Generation. *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*. Birmingham, U.K.

Sak, H., Güngör, T., & Saraçlar, M. (2008). Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus. *GoTAL* (pp. 417-427). Springer.

Say, B., Zeyrek, D., Oflazer, K., & Özge, U. (2002). Development of a Corpus and a Treebank for Present-day Written Turkish. *Paper presented at the 11th International Conference on Turkish Linguistics*.

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. Amsterdam/Philadelphia: John Benjamins.

Webber, B. (2004, September). D-LTAG: Extending Lexicalized TAG to Discourse. *Cognitive Science*.

Webber, B., & Joshi, A. (1998, August). Anchoring a Lexicalized Tree-Adjoining grammar for Discourse. *ACL/COLING Workshop on Discourse Relations and Discourse Markers*. Montreal, Canada.

Webber, B., Joshi, A., Miltsakaki, E., Prasad, R., Dinesh, N., Lee, A., et al. (2005). A Short Introduction to the Penn Discourse TreeBank. *Copenhagen Working Papers in Language and Speech Processing*.

Yannick, V. (2011). Towards finer-grained tagging of discourse connectives. *DGfS Workshop Beyond Semantics*. Göttingen, Germany.

Zeyrek, D., & Webber, B. (2008). A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. *The 6th Workshop on Asian Language Resources, The Third International Joint Conference on Natural Language Processing, (IJNLP).*

Zeyrek, D., Turan, Ü., Bozşahin, C., Çakıcı, R., Sevdik-Çallı, A., Demirşahin, I., et al. (2009). Annotating Subordinators in the Turkish Discourse Bank. *ACL-IJCNLP, Linguistic Annotation Workshop III. 44–48.*

APPENDICES

Appendix A: Summary Table

Suffix	**Roles**	**Possible Features**	**Ambiguity**
-AcAğInA	• Simplex subordinator • Complement of Verb Phrase	• Syntactic features • Factive verb list	Ambiguous
-AcAğIndAn	• Simplex subordinator • Complement of Verb Phrase	• Syntactic features • Factive verb list	Ambiguous
-AlI	• Simplex subordinator		Unambiguous
-ArAk	• Simplex subordinators • Manner • Adverbials	• Semantic convenience • Shared object with main clause • Presence of *ol-arak*	Hard case
-ArcAsInA	• Simplex subordinators • Manner	• Semantic convenience • Shared object with main clause • Presence of an idiomatic expression	Hard case
–dIğIndA	• Simplex subordinators • Complement of Verb Phrase	• Syntactic features • Factive verbs	Ambiguous
–dIğındAn	• Simplex subordinators • Complement of Verb Phrase • Headless Relative Clause	• Syntactic features • Factive verbs • Context knowledge	Ambiguous
–dIkçA	• Simplex subordinators • Lexicalized items	• Syntactic features • Presence of	Ambiguous

		reduplication and lexicalized items	
–IncA	• Simplex subordinators • Lexicalized items	• Syntactic features • Presence of lexicalized items	Ambiguous
-Ip	• Simplex subordinators • Lexicalized items	• Syntactic features • Presence of reduplications and collocations	Ambiguous
–ken	• Simplex subordinators • Discourse adverbial • Lexicalized Items	• Syntactic features • Presence of reduplications	Ambiguous
–mAksIzIn	• Simplex subordinators • Manner	• Syntactic features • Shared object with main clause • Presence of *ol-maksızın*	Hard case
–mAktAnsA	• Simplex subordinators		Unambiguous
–mIşcAsInA	• Simplex subordinators • Manner	• Semantic convenience • Shared object with main clause • Presence of an idiomatic expression	Hard case
–sA	• Simplex subordinators • Other DC • Focus particle	• Syntactic features • Presence of Conventional Constructions	Ambiguous

Appendix B: List of Discourse Connectives in Turkish

connective	**type**
(ve)yahut	
diye	
mesela	
örneğin	
yani	
zira	
ama	conjoiner
çünkü	conjoiner
fakat	conjoiner
ve	conjoiner
veya	conjoiner
ya da	conjoiner
zaten	conjoiner
dA	conjoiner particle
ki	conjoiner particle
-AlI	simple subordinator
-ArAk	simple subordinator
-DAn	simple subordinator
-DığIndA	simple subordinator
-DIkçA	simple subordinator
-IncA	simple subordinator
-Ip	simple subordinator

connective	type
-ken	simple subordinator
-sA	simple subordinator
(ya/yok eğer) ... –sA	paired subordinator
-A dayanarak	paired subordinator
-A dek	paired subordinator
-A ek olarak	paired subordinator
-A ilaveten	paired subordinator
-A kadar	paired subordinator
-A karşılık	paired subordinator
-A karşın	paired subordinator
-A rağmen	paired subordinator
-AlI beri	paired subordinator
-DAn başka	paired subordinator
-DAn beri	paired subordinator
-DAn bu yana	paired subordinator
-DAn dolayı	paired subordinator
-DAn itibaren	paired subordinator
-DAn önce	paired subordinator
-DAn ötürü	paired subordinator
-DAn sonra	paired subordinator
-DIğI biçimde	paired subordinator
-dIğI gibi	paired subordinator
-DIğI şekilde	paired subordinator
-DIğI takdirde	paired subordinator
-DIğI/-AcAğI anda	paired subordinator
-DIğI/-AcAğI halde	paired subordinator
-DIğI/-AcAğI için	paired subordinator

connective	**type**
-DIğI/-AcAğI kadar	paired subordinator
-DIğI/-AcAğI sırada	paired subordinator
-DIğI/-AcAğI üzere	paired subordinator
-DIğI/-AcAğI zaman	paired subordinator
-DIğInA göre	paired subordinator
-Ir gibi	paired subordinator
-lA birlikte	paired subordinator
-mAk üzere	paired subordinator
-mAk yerine	paired subordinator
-mAk/-mAsI açısından	paired subordinator
-mAk/-mAsI amacıyla	paired subordinator
-mAk/-mAsI için	paired subordinator
-mAk/-mAsI üzerine	paired subordinator
-mAsI durumunda	paired subordinator
-mAsI halinde	paired subordinator
-mAsI nedeniyle	paired subordinator
-mAsI/-IşI yüzünden	paired subordinator
-nIn ardından	paired subordinator
-nIn yanısıra	paired subordinator
-sA dA/bile	paired subordinator
(bir) başka deyişle	discourse adverbial
(en) sonunda	discourse adverbial
(her) neyse	discourse adverbial
aksi halde	discourse adverbial (anaphoric?)
aksi takdirde	discourse adverbial (anaphoric?)
aksine	discourse adverbial
ardından	discourse adverbial

connective	**type**
ayrıca	discourse adverbial
benzer şekilde	discourse adverbial (anaphoric?)
bir de	discourse adverbial
ek olarak	discourse adverbial
en azından	discourse adverbial
halbuki	discourse adverbial
ilaveten	discourse adverbial
nihayet	discourse adverbial
o halde	discourse adverbial (anaphoric?)
o zaman	discourse adverbial (anaphoric?)
oysa	discourse adverbial
önce	discourse adverbial
öte yandan	discourse adverbial
sonra	discourse adverbial
sonrasında	discourse adverbial
sonuç olarak	discourse adverbial
sonuçta	discourse adverbial
yoksa	discourse adverbial
(ya/yok eğer) böyleyse	anaphoric discourse adverbial ?
böylece	anaphoric discourse adverbial ?
böyleyse de	anaphoric discourse adverbial ?
bu	anaphoric discourse adverbial
bu açıdan	anaphoric discourse adverbial
bu amaçla	anaphoric discourse adverbial
bu ana dek	anaphoric discourse adverbial
bu ana kadar	anaphoric discourse adverbial
bu anda	anaphoric discourse adverbial

connective	**type**
bundan itibaren *	anaphoric discourse adverbial
bu arada	anaphoric discourse adverbial
bu bağlamda	anaphoric discourse adverbial
bu biçimde	anaphoric discourse adverbial
bu durumda	anaphoric discourse adverbial
bu nedenle	anaphoric discourse adverbial
bu sırada	anaphoric discourse adverbial
bu şekilde	anaphoric discourse adverbial
bu takdirde	anaphoric discourse adverbial
bu yüzden	anaphoric discourse adverbial
buna dayanarak	anaphoric discourse adverbial
buna ek olarak	anaphoric discourse adverbial
buna göre	anaphoric discourse adverbial
buna ilaveten	anaphoric discourse adverbial
buna karşılık	anaphoric discourse adverbial
buna karşın	anaphoric discourse adverbial
buna rağmen	anaphoric discourse adverbial
bundan başka	anaphoric discourse adverbial
bundan dolayı	anaphoric discourse adverbial
bundan önce	anaphoric discourse adverbial
bundan ötürü	anaphoric discourse adverbial
bundan sonra	anaphoric discourse adverbial
bunun ardından	anaphoric discourse adverbial
bunun gibi	anaphoric discourse adverbial
bunun için	anaphoric discourse adverbial
bunun üzerine	anaphoric discourse adverbial
bunun yanısıra	anaphoric discourse adverbial

connective	**type**
bunun yerine	anaphoric discourse adverbial
bununla birlikte	anaphoric discourse adverbial
ondan beri	anaphoric discourse adverbial
ondan bu yana	anaphoric discourse adverbial