AN INTER-ANNOTATOR AGREEMENT MEASUREMENT METHODOLOGY FOR THE TURKISH DISCOURSE BANK (TDB)

A THESIS SUBMITED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ŞABAN İHSAN YALÇINKAYA

IN PARTIAL FULLFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

SEPTEMBER 2010

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. Deniz Zeyrek
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
Supervisor

Examining Committee Members

Assoc. Prof. Dr. Cem Bozşahin	(METU, CENG)	____________________
Prof. Dr. Deniz Zeyrek	(METU, COGS)	____________________
Dr. Ruket Çakıcı	(METU, CENG)	____________________
Dr. Ceylan Talu-Yozgatlıgil	(METU, STAT)	____________________
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 : Şaban İhsan Yalçınkaya

Signature : __________________

ABSTRACT

AN INTER-ANNOTATOR AGREEMENT MEASUREMENT METHODOLOGY
FOR THE TURKISH DISCOURSE BANK (TDB

Yalçınkaya, Şaban İhsan
MS, Department of Cognitive Science
Supervisor: Prof. Dr. Deniz Zeyrek

September 2010, 128 pages

In the TDB[1]-like corpora annotation efforts, which are constructed by the intuitions of the annotators, the reliability of the corpus can only be determined via correct inter-annotator agreement measurement methodology (Artstein, & Poesio, 2008). In this thesis, a methodology was defined to measure the *inter-annotator agreement* among the TDB annotators.

[1] Turkish Discourse Bank (TDB) is used to denote ODTÜ-MEDİD (Orta Doğu Teknik Üniversitesi Metin Düzeyinde İşaretlenmiş Derlem), a project supported by TÜBİTAK.

The *statistical tests* and the *agreement coefficients* that are widely used in scientific communities, including Cochran's Q test (1950), Fleiss' Kappa (1971), and Krippendorff's Alpha (1995), were examined in detail. The inter-annotator agreement measurement approaches of the various corpus annotation efforts were scrutinized in terms of the reported statistical results. It was seen that none of the reported inter-annotator agreement approaches were statistically appropriate for the TDB. Therefore, a comprehensive inter-annotator agreement measurement methodology was designed from scratch. A computer program, the Rater Agreement Tool (RAT), was developed in order to perform statistical measurements on the TDB with different corpus parameters and data handling approaches.

It was concluded that Krippendorff's Alpha is the most appropriate statistical method for the TDB. It was seen that the measurements are affected with data handling approach preferences, as well as the used agreement statistic methods. It was also seen that there is not only one correct approach but several approaches valid for different research considerations. For the TDB, the major data handling suggestions that emerged are: (1) considering the *words* as building blocks of the annotations and (2) using the *interval approach* when it is preferred to weigh the partial disagreements, and using the *boundary approach* when it is preferred to evaluate all disagreements in same way.

Keywords: Discourse, Discourse Bank, Inter-Annotator Agreement, Corpus Reliability, Text Span Annotation, Agreement Coefficients

ÖZ

TÜRKÇE SÖYLEM BANKASI İÇİN İŞARETÇİLER ARASI UYUM ÖLÇÜM METODOLOJİSİ

Yalçınkaya, Şaban İhsan
Yüksek Lisans, Bilişsel Bilimler Bölümü
Tez Yöneticisi: Prof. Dr. Deniz Zeyrek

Eylül 2010, 128 sayfa

TSB[2] benzeri derlem işaretleme çabaları, işaretçilerin sezgileriyle inşa edildiğinden, derlem güvenilirliği sadece doğru işaretçiler arası uyum ölçüm yöntemiyle ölçülebilir (Artstein, & Poesio, 2008). Bu tezde, Türkçe Söylem Bankası (TSB) işaretçileri arasındaki *işaretçiler arası uyumu* hesaplamak için bir yöntem tanımlanmıştır.

Bilimsel çevrelerde yaygın olarak kullanılan *istatistiksel testler* ve *uyum katsayıları*, Cochran'ın Q testi (1950), Fleiss'in Kappası (1971), ve Krippendorff'un Alphası (1995)

[2] Tükçe Söylem Bankası (TSB), TÜBİTAK tarafından desteklenen ODTÜ-MEDİD (Orta Doğu Teknik Üniversitesi Metin Düzeyinde İşaretlenmiş Derlem) projesini ifade etmektedir.

da dahil olmak üzere, detaylı bir şekilde incelenmiştir. Çeşitli derlem işaretleme çabalarının işaretçiler arası uyum ölçüm yaklaşımları istatistiksel açıdan irdelenmiştir. Görülmüştür ki bu çabaların bildirilmiş hiçbir işaretçiler arası uyum ölçüm yaklaşımı istatistiksel olarak TSB'ye uygun değildir. Bu nedenle, kapsamlı bir işaretçiler arası uyum ölçüm yöntemi baştan tasarlanmıştır. Tasarlama sürecinde, TSB üzerinde istatistiksel ölçümleri değişik derlem parametreleri ve veri işleme yaklaşımlarıyla gerçekleştirmek üzere, Derecelendirici Uyum Aracı (DUA) adı verilen, bir bilgisayar programı geliştirilmiştir.

TSB için en uygun istatistiksel yöntemin Krippendorff'un Alphası olduğu sonucuna varılmıştır. Görülmüştür ki ölçümler kullanılan uyum istatistiklerinden etkilendikleri kadar veri işleme yaklaşımı tercihlerinden de etkilenmektedirler. Yine görülmüştür ki bütün araştırma konuları için tek bir doğru yaklaşım yoktur, ancak çeşitli araştırma konuları için değişik doğru yaklaşımlar vardır. TSB için, bu tezde ortaya çıkan ana veri işleme yaklaşımları: (1) *kelimeleri* işaretlemelerin yapı taşı olarak değerlendirmek ve (2) *aralık yaklaşımının* kısmi uyumsuzlukları ağırlıklandırılmak istenildiğinde kullanılması, ve *sınır yaklaşımının* bütün uyumsuzlukları aynı şekilde değerlendirmek istenildiğinde kullanılmasıdır.

Anahtar kelimeler: Söylem, Söylem Bankası, İşaretçiler Arası Uyum, Derlem Güvenilirliği, Metin Kapsam İşaretlemesi, Uyum Katsayıları

To My Wife…

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest sincere gratitude to my supervisor Prof. Dr. Deniz Zeyrek for her guidance, support, criticism and encouragement she provided me during this study. Also, I would like to thank to Assoc. Prof. Dr. Cem Bozşahin for the initial idea and his continuous support.

I thank Dr. Annette Hohenberger and Dr. Ceyhan Temürcü for their valuable criticism and suggestions. I would also like to thank to Dr. Ceylan Talu Yozgatlıgil for making complicated statistical issues more tangible.

I am also grateful to my project fellows Dr. Ruket Çakıcı, Işın Demirşahin and Ayışığı Başak Sevdik Çallı for their precious contributions to this study and their patient replies to my endless questions.

I owe special thanks to my colleague Dr. Kerim Savran for his guidance and kindness. He had always excellent answers to my endless how-to questions. Also I am grateful to my colleagues: Oğuz Acartürk, Salih Aslan, Uğur Bilgi, Gökhan Biricik, Özgür Bora Ergün, Fatma Şirin Koru and Cevahir Turgut for listening patiently my complaints and motivating me always. I feel lucky for having such great friends.

I wish to express my gratitude to my parents Halise and Yalçın Yalçınkaya and my beloved sister Ezgi Yalçınkaya for their lifetime of support and encouragement. I thank them for always being with me.

I would like to express my deepest gratitude to my wonderful wife Sinem Sürmeli Yalçınkaya, without her love, understanding, patience and support this study would never be ended.

I also thank to TÜBİTAK (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu) for financially supporting me with domestic graduate study fellowship.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Corpus linguistics is a data-centric linguistic research method where linguistic theories are drawn from collected-data (McEnery, & Wilson, 2001). As Artstein, & Poesio (2008) mentioned, since mid 1990s, the number of corpus-based linguistic efforts has been increasing. Examples are the Brown Corpus (Francis, Kucera, & Mackie, 1982), the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993), the British National Corpus (Burnage, 1996), the Rhetorical Structure Theory (RST) based corpus (Lynn, Marcu, Okurowski, 2003), the Penn Discourse Tree Bank (Miltsakaki, Prasad, & Webber, 2004), the Chinese Discourse Tree Bank (Xue, 2005), the Turkish Discourse Bank (Zeyrek, & Webber, 2008) the Hindi Discourse Relation Bank (Oza, Prasad, Kolachina, Sharma, & Joshi, 2009b), and the Leeds Arabic Discourse Tree Bank (Al-Saif, & Markert, 2010). The major reason for this increase is the development of computer technology that makes processing, analyzing, storing, and distributing huge amounts of data possible. However huge amount of data, in addition to several research opportunities, bring some problems. The foremost problem is the reliability issue of the collected data. A corpus study is considered to be suspicious when it is based on a corpus whose reliability is arguable (Artstein, & Poesio, 2008).

The Turkish Discourse Bank (TDB) is a corpus annotation project that aims to reveal Turkish discourse structures by annotating METU Turkish Corpus, which is a sentence-level corpus that contains 2 million words from different genres (Say, Zeyrek, Oflazer,

& Özge, 2002; Zeyrek, Turan, Bozşahin, Çakıcı, Sevdik-Çallı, Yalçınkaya, et al., 2009). The data collection process of the TDB was performed by two or three annotators where all the annotators performed annotations individually in the whole corpus. The annotators performed text span annotations for explicit discourse connectives and their two arguments in order to establish the semantic relations between them. As in other corpus efforts, in the TDB project, the reliability of the collected (annotated) data needs to be proved. The reliability of the collected data is measured by inter-annotator agreement statistic methods, which rate the agreement of the annotators among the corpus. The primary objective of this thesis is to determine appropriate inter-annotator agreement statistics to measure the reliability of the TDB data, hence the TDB project.

In order to reach the objective of the thesis, first the agreement statistic methods were discussed in detail. Then, the corpus annotation efforts, which either follow the same principles with the TDB or which have used inter-annotator agreement measurement approaches, were examined. As a result of the examinations, the positive and negative statistical aspects of the corpus annotation efforts were determined. Also, it was seen that none of the reported inter-annotator agreement measurement approaches were adequate for the TDB.

As a result of the investigations, it was decided to design a comprehensive inter-annotator agreement measurement methodology for the TDB. In order to design an adequate inter-annotator agreement methodology, the annotation data must be fully examined. Then, the TDB was examined in following headings:

- Annotation data type (text span annotations of Turkish explicit discourse connectives)
- Annotation process
- Dependency analysis of annotations
- Annotation data handling challenges

- Huge amount of annotation data
- Possibility of corrupted annotations (annotation process problems and annotation storage problems)
- Evaluating annotation together, which were performed at different times by different annotators
- Converting text span annotations into discrete computational data units
- Computationally problematic dependencies

In the light of all these investigations, data handling approaches and the agreement statistics, which can be used with these data handling approaches, were proposed. As a result of these, several measurement combinations (data handling approaches X agreement statistics) emerged. In order to perform inter-annotator agreement measurements on this huge amount of data for various combinations, an original computer program was developed. The computer program is capable of performing measurements for different annotation categories (discourse connective, annotator number annotation repository, data handling approaches, agreement statistics, etc.). The measurement results produced by this computer program were analyzed and the most appropriate inter-annotator agreement measurement combinations were decided.

As a result of the work carried out, it was seen that there is not one correct way to measure the inter-annotator agreement of the TDB, but there are different ways for different research aspects which can be produced with the inter-annotator agreement measurement methodology, which is proposed in this thesis. Additionally, the inter-annotator agreement methodology and the computer program, which are presented in this thesis, are also important sources for the corpus annotation efforts other than the TDB. It is planned to make this program available to community with a user manual.

This thesis is organized in the following order: In Chapter 2, three classes of statistic methods that can be used to measure the inter-annotator (inter-rater) agreements will be discussed in detail. First, Cochran’s Q Test (Cochran, 1950) will be elaborated, and then the Benett's Sigma (Benett, Alpert, & Goldstein, 1954), the Scott's Pi (Scott, 1955), the Cohen's Kappa (Cohen, 1960), and Fleiss’ Kappa (Fleiss, 1971) will be discussed as the members of the Kappa family. Finally, Krippendorff’s Alpha (Krippendorff, 1995) will be presented.

In Chapter 3, the inter-annotator agreement approaches of several discourse annotation projects will be examined by using discussions in Chapter 2. In this section, the Penn Discourse Tree Bank (PDTB), the Rhetorical Structure Theory (RST) based annotation effort, the Hindi Discourse Relation Bank (HDRB), the Chinese Discourse Tree Bank (CDTB), and the Leeds Arabic Discourse Tree Bank (LADTB) projects are covered.

In Chapter 4, the TDB project will be presented. In the light of Chapter 2 and 3, some answers will be sought to the following questions: Which agreement statistics are appropriate for the TDB, and how the annotation data of the TDB should be used in the agreement statistics. Consequently, several agreement statistics and several data handling methods will be determined, which can be used to measure the inter-annotator agreement on the TDB.

In Chapter 5, the results of the TDB agreement measurements, which are calculated by a computer program developed during this thesis, will be presented. Also, some data handling methods will be eliminated upon the result analysis and, the usage areas of the rest of the methods will be explicitly defined. Finally, the computer program and possible future work opportunities are elaborated.

In Chapter 6, the most appropriate inter-annotator agreement methodologies based on the findings in Chapter 5 will be presented for the TDB, with the future work plans.

CHAPTER II

INTER-RATER AGREEMENT

Gwet (2001) defines inter-rater reliability as the amount of agreement between raters (data generators). A rater is a classifier that classifies subject items into predefined categories according to a particular classification rule set.

In general, inter-rater reliability coefficients are used as the statistical magnitudes to measure the quality of rated data. In the literature, the term is named in various ways, such as, inter-observer reliability (Hartmann, 1977), inter-coder reliability (Fleiss, 1971), inter-judge reliability (Dillon, 1984) and so on, but all of them aim to measure the quality of data collection process by measuring the agreement on the preferences (Gwet, 2001). In the scope of this thesis, the term *inter-annotator agreement coefficient* will be used, as it will be more accurate.

However, it is seen that some research studies prefer to use the statistical tests instead of agreement coefficients (Yöndem-Turhan, 2001). Statistical tests differ from agreement coefficients in that they are the methods that operate through the hypothesis to reject the null hypothesis (Siegel, & Castellan, 1988).

In this chapter, in order to cover both of these points of view about the inter-annotator agreement issue mentioned above, first Cochran's Q test will be elaborated as it is a well known and commonly used statistical test, also applicable to the TDB. Afterwards, a selection of agreement coefficients, including Fleiss' Kappa and Krippendorff's Alpha, will be discussed in detail.

2.1 COCHRAN'S Q TEST

There are various statistical tests each of which operates correctly for particular populations and data types. Deciding on an appropriate statistical test for the TDB was a challenge. Therefore, first, the reasons for selecting Cochran's Q test will be discussed, and then the mathematical basis of the test will be presented.

There are two main classes of statistical tests: parametric and non-parametric tests. The first challenge was deciding on the main class of the intended test. The parametric tests require well designated populations and they are suitable for numerical data, whereas, non-parametric tests do not need population assumptions and non-numerical data can be evaluated by them (Siegel, 1957). So, it was seen that non-parametric tests are suitable for corpus annotation efforts. The second challenge was deciding on the appropriate measurement theory. As Siegel et al. (1988) mentioned there are four measurement theories: nominal, ordinal, interval, and ratio scales. The TDB annotation process is a classification process of the text spans, and there is no relation between the assigned classes. This kind of data fits perfectly for the nominal scale definition. Then, as the final two challenges, annotations are performed by two or three annotators, and they just make dichotomous (binary) classifications (annotate or do not annotate). For that kind of data analysis, Siegel (1988) suggests Cochran's Q test.

This statistical test which aims to "test the significance of differences between ratios or percentages in two or more independent samples" and which is based on χ^2 test was presented by Cochran in 1950 (p. 256). This test is an extended version of Quin McNemar's (McNemar, 1949) test to handle more than two samples. Therefore, investigating the McNemar test first will ease the discussion, and then the Cochran's addition to the McNemar test can be elaborated.

The McNemar (1949) test is designed for observations that aim to measure the changes on two samples. An example research study is presented by McNemar, which was conducted on USA soldiers at the time of World War II. Soldiers were asked "whether they thought that the war against Japan would last more or less than a year". Afterwards, soldiers were lectured for the difficulties of the war against Japan and the same question was asked again. The concern of the experiment was to determine if the lecture was significant on changing the minds of the soldiers. An effective lecture would affect the soldiers' mind in the same direction, whereas the lecture that had no effect would not affect the soldiers mind in the same direction, and changes would be randomly distributed. McNemar used a 2x2 table to illustrate mind changes:

Table 1 2x2 table of McNemar

	After lecture: Less	After lecture: More
Before lecture: Less	A	b
After Lecture: More	C	d

According to Table 1, b is the number of less than a year to more than a year changes, and c is the number of more than a year to less than a year changes after the lecture. McNemar used b and c values to calculate χ^2 as follows:

$$n = (b + c), with\ probablity\ \tfrac{1}{2}$$ **(Equation 2.1)**

$$\chi^2 = \frac{(b-\frac{1}{2}n)^2}{\frac{1}{2}n} + \frac{(c-\frac{1}{2}n)^2}{\frac{1}{2}n} = \frac{(b-c)^2}{b+c}$$ **(Equation 2.2)**

As McNemar did, Cochran presented his test with a sample experiment, a diphtheria bacilli investigation that was conducted by the Communicable Disease Centre, U.S. Public Health (Cochran, 1950). In the investigation, researchers aimed to determine the different habitat effects on the growth of diphtheria bacilli. For this purpose, 69 samples were collected from suspicious cases. Each sample was equally divided into habitat A, B, C and D. Table 2 illustrates the investigation results. 1s represent growth, and 0s represent no growth in related habitats.

Table 2 The Cochran's Q test table for diphtheria bacilli experiment

	A	B	C	D	Num. Of Samples
	1	1	1	1	4
	1	1	0	1	2
	0	1	1	1	3
	0	1	0	1	1
	0	0	0	0	59
Totals(T_j)	6	10	7	10	

Let,

c be the column number of the table

u_i be the ith row total number of 1s

T_j be the jth column total number of 1s

$\overline{T}$ be the average of 1s for each c (Total 1s/c)

df be the degree of freedom

H_0 be the null hypothesis that the habitats has no effect on growth of diphtheria bacilli

the significance level that is selected as 0.05

The null hypothesis of the test claimed that, the success probability of diphtheria bacilli is same in all habitats. Mathematically that means, u_i is distributed among c columns in $\binom{c}{u_i}$ ways (Cochran, 1950). Cochran formulated his claim as is Equation 2.3 below:

$$Q = \frac{c(c-1)\sum(T_j-\bar{T})^2}{c(\sum u_i)-(\sum u_i^2)} \quad \textbf{(Equation 2.3)}$$

Siegel (1988, p. 174) explains how to evaluate the value found in Equation 2.3: “Q is distributed approximately as χ^2 with df = c-1. If the probability associated with the occurrence of when H_0 is true of a value as large as the observed value of Q is equal to or less than α, reject H_0”.

It will be accurate to statistically analyze the above investigation in order to make the Cochran’s test clearer. Analysis of the investigation is,

$$c = 4$$

$$\bar{T} = \frac{(6+10+7+10)}{4} = 8.25$$

$$\sum_j (T_j - \bar{T})^2 = (6-8.25)^2 + (10-8.25)^2 + (7-8.25)^2 + (10-8.25)^2 = 12.75$$

$$\sum_i u_i = 4+4+4+4+3+3+3+3+3+2 = 33$$

$$\sum_i u_i^2 = 4^2+4^2+4^2+4^2+3^2+3^2+3^2+3^2+3^2+2^2 = 113$$

$$Q = \frac{(4)(3)(12.75)}{(4)(33)-(113)} = 8.05$$

According to Appendix A, Table 21, Q≥8.05 has probability of occurrence when H_0 is true of p-value <0.05 when df is 3=4-1. If the significance level is selected as 0.05, H_0 is rejected.

Finally, the issue of the power of test should be investigated. According to Cohen "The power of a statistical test of a null hypothesis (H_0) is the probability that the H_0 will be rejected when it is false, that is, the probability of obtaining a statistically significant result" (1992, p. 98). In this respect, for Cochran's Q test, calculating the power of test is not necessary when the sample size is greater than 4 and the product of the sample size and the category size is greater than 24, because the exact distribution of Q is perfectly approximated by the χ^2 distribution. In the example above, the sample size is 69, and thus there is no need to calculate the power of test. Also, the TDB contains tens of annotations; therefore Cochran's Q Test can be applied and evaluated without the power of test notion. (See Appendix B for the java implementation of Cochran's Q Test.)

2.2. AGREEMENT COEFFICIENTS

Gwet (2001) defines agreement coefficients as the methods that are used to estimate the reliability of the rated data. These methods are used when rated data are qualitative, and true classification is hard to determine. Several agreement coefficients are used in the literature to calculate inter-annotator agreement, which all claim to eliminate agreements by chance and which all have different approaches to agreement issue. In this thesis, several well-known agreement coefficients will be investigated. In order to ease the understanding of the statistical concepts of these agreement coefficients, a set of terms and agreement coefficient basis will be presented in the below section.

2.2.1 A SET OF COMMON TERMS AND AGREEMENT COEFFICIENT BASIS

In this thesis, the definitions of Artstein & Poesio (2008) will be followed in principle, but their terms will be used with some fine-tuning. There are two reasons for this: First, the terminology does not fully cover the needs of all discussed agreement coefficients, so some definitions from Krippendorff (1995 and 2004b) will be added for full coverage. Secondly, it is preferred to use terms specific to an annotation study. The list of the agreement terms and abbreviations which are used in this thesis as follows:

- Annotating designates coding/rating
- Annotator number is always referred to as I
- Category number is always referred to as K
- Annotation number is always referred to as H
- Individual annotator is always referred to as i or j
- Each annotator may perform annotation to category c or k
- When annotations are performed on discrete data
    - Annotated items are referred to as g or h
- When annotations are performed on continuous data;
    - Units and gaps are represented by their beginning (b) and length (l)
    - Units and gaps are referred as g or h
    - As an attribute of continuum data, v is used to define gaps and units. v represents a unit when its value is 1, and a gap when its value is 0
    - δ^2_{cigjh} represents the difference function between unit g that is annotated by annotator i, and unit h, that is annotated by annotator j, for category c.
- Agreement and disagreement will be shown as follows;
    - Agr_h indicates the agreement on item h
    - Expected (chance) agreement: Ae
    - Expected (chance) disagreement: De
    - Observed agreement: Ao
    - Observed disagreement: Do

Basically, agreement on an annotation can be determined by means of the percentage of agreed annotations. Percent agreement can be found for a particular annotation by giving the value 1 to annotation when all of the annotators agree (Agr_h=1) and by giving the value 0 to annotation when at least one of the annotators does not agree with the rest. To find the agreement for an annotation effort the following Equation 2.4 formula can be used (Artstein, & Poesio, 2008):

$$A_o = \frac{1}{H} \sum_{h \in H} Agr_h \qquad \textbf{(Equation 2.4)}$$

This method is not wrong, but it is statistically weak and misleading. First of all, agreement by chance among annotators is not considered. The chance factor is a variable that affects the results unpredictably for different annotations, annotators, moments, and situations. Therefore, agreements cannot be comparable among different studies without removing the chance factor (Scott, 1955). Secondly, some annotation categories may be likely to be chosen more than others by their nature. In other words, determined categories do not have to be evenly distributed. As Hsu & Field (2003) indicated, in an artificial annotation effort where two annotators perform annotations for two categories and categorize 95% of annotations under the first category and the rest under the second category, a randomly picked first category choice shall be correct for 90.25% (0.95x0.95), and the second category choice 0.25% (0.05x0.05). In the example above, an agreement below 90.25% would not be reliable because it may have occurred by chance. Consequently, a reliability determination methodology without chance-correction would be unreliable and incomparable. Furthermore, percentage agreement only offers exact match evaluation. Exact match is inadequate when there are three or more annotators.

For these reasons, agreement coefficients are necessary. However, the starting points and the considerations of these coefficients vary. Therefore in the agreement measurements, appropriate agreement coefficient for the rating paradigm of the research and data handling mechanisms should be used. Therefore, in the subsequent sections, the most popular and well based agreement coefficients i.e., the Kappa family and Krippendorff's Alpha will be discussed and their characteristics will be explained.

2.2.2 KAPPA AND SIMILAR AGREEMENT COEFFICIENTS

The Kappa agreement coefficient was presented to computational linguists and cognitive scientist who study discourse by Jean Carletta in 1996. However, the first kappa implementation was introduced by Cohen (1960), long before Carletta (1996).

Cohen's Kappa (1960) can be used to measure the degree of agreement, by correcting the agreement by chance between two annotators, where each annotator annotates subjects on a nominal scale (Fleiss, 1971). Another version of kappa by Cohen is Cohen's Weighted Kappa (1968), which is very similar to the normal kappa, but additionally weights the different disagreement types.

Cohen's Kappa is a member of a chance-corrected agreement measurement family that measures agreement between two annotators, which were mostly introduced in the 1950s and 60s. The most known three examples are ς (Benett, 1954), π (Scott, 1955), and κ (Cohen, 1960) (Artstein, & Poesio 2008). All of them have the following formula:

$$\varsigma, \pi, \kappa = \frac{A_0 - A_e}{1 - A_e} \qquad \textbf{(Equation 2.5)}$$

$$\text{Where, } A_e^{\varsigma} = A_e^{\pi} = A_e^{\kappa} = \sum_{k \in K} P(k|i_1)\, P(k|i_2)$$

The difference between ς, π, and κ lies in the calculation of $P(k|i)$, where $P(k|i)$ is the probability that annotator i will assign an arbitrary item to category k (Zwick, 1988; Hsu, & Field, 2003; Artstein, & Poesio 2008). If annotators had made random picks:

- ς: Assumes uniform distribution. For any two annotator i_m, i_n, and any two categories k_j, k_l, $P(k_j|i_m) = P(k_l|i_n)$
- π: Assumes same distribution for each annotator. For any two annotators i_m, i_n, and any category k, $P(k|i_m) = P(k|i_n)$
- κ: Assumes separate distribution for each annotator.

The above statements lead to the following chance agreement formulas (Artstein, & Poesio 2008):

$$A_e^{\varsigma} = \sum_{k \in K} \frac{1}{k^2} = \frac{1}{k} \qquad \textbf{(Equation 2.6)}$$

$$A_e^{\pi} = \sum_{k \in K} \left(\frac{n_k}{2H}\right)^2 = \frac{1}{4H^2} \sum_{k \in K} n_k^2 \qquad \textbf{(Equation 2.7)}$$

$$A_e^{\kappa} = \sum_{k \in K} \frac{n_{i_1 k}}{H} . \frac{n_{i_2 k}}{H} = \frac{1}{H^2} \sum_{k \in K} n_{i_1 k} n_{ik} \qquad \textbf{(Equation 2.8)}$$

$A_e^{\pi} \geq A_e^{\varsigma}$ and $A_e^{\pi} \geq A_e^{\mathrm{k}}$ correlations are extracted from the above formulas, and the relation of A_e^{ς} *and* A_e^{k} cannot be exactly extracted. In order to stay on the safe side, the coefficients which tend to produce higher chance agreement (A_e) values shall be selected. π remains as the safest agreement measurement coefficient between two annotators.

Some TDB annotations are performed by three annotators. An agreement coefficient that handles three annotators is needed. Therefore, Fleiss' Kappa, which is an extended version of π for three or more coders and "defines the amount of agreement on a particular item as the proportion of agreeing judgment pairs out of the total number of judgment pairs for that item" (Artstein, & Poesio, 2008 p. 562), is more suitable for the TDB needs. Fleiss' Kappa is formulated as follows:

Let, n_{hk} be the number that the sample h is assigned to category k, and $\binom{i}{2}$ be the total number of judgment pairs per item. The division of $\binom{n_{hk}}{2}$ totals of all categories by $\binom{i}{2}$ would give the agreement on the item h. The total observed agreement is the arithmetic average of the agreements on the all samples (Artstein, & Poesio, 2008). Then the observed agreement can be given as:

$$agr_h = \frac{1}{\binom{i}{2}} \sum_{k \in K} \binom{n_{hk}}{2} \qquad \textbf{(Equation 2.9)}$$

$$A_o = \frac{1}{H}\sum_{h\in H} agr_h = \frac{1}{HI(I-1)}\sum_{h\in H}\sum_{k\in K} n_{hk}(n_{hk}-1) \qquad \textbf{(Equation 2.10)}$$

According to Fleiss (1971), expected agreement shall be calculated by considering all coder judgments together by assuming the same distribution for each coder, just like Scott (1955). Therefore, the expected agreement formula of Fleiss is very similar to Scott's (Artstein, & Poesio, 2008):

$$\hat{P}(k) = \frac{1}{HI} n_k \qquad \textbf{(Equation 2.11)}$$

$$A_e = \sum_{k\in K}\left(\hat{P}(k)\right)^2 = \sum_{k\in K}\left(\frac{1}{HI} n_k\right)^2 = \frac{1}{(HI)^2}\sum_{k\in K} n_k^2 \qquad \textbf{(Equation 2.12)}$$

Fleiss (1971) presents his claim with a diagnosis experiment, that 6 psychiatrists diagnose 30 subjects into 5 categories. The psychiatrists assign numbers from 1 to 5 for their diagnosis, which indicate depression, personality disorder, schizophrenia, neurosis, and other, respectively. Fleiss constructs the following agreement table (Di Eugenio, & Glass, 2004) in order to arrange the data:

Table 3 Agreement table of Fleiss' (1971) diagnosis experiment

	Category					
Subject	**1**	**2**	**3**	**4**	**5**	**agri**
1				6		1.000
2		3			3	0.400
3		1	4		1	0.400
4					6	1.000
5		3		3		0.400
6	2		4			0.467
7			4		2	0.467
8	2		3	1		0.267
9	2			4		0.467
10					6	1.000
11	1			5		0.667
12	1	1		4		0.400
13		3	3			0.400
14	1			5		0.667
15		2		3	1	0.267
16			5		1	0.667
17	3			1	2	0.267
18	5	1				0.667
19		2		4		0.467
20	1		2		3	0.267
21					6	1.000
22		1		5		0.667
23		2		1	3	0.267
24	2			4		0.467
25	1			4	1	0.400
26		5		1		0.667
27	4				2	0.467
28		2		4		0.467
29	1		5			0.667
30					6	1.000
Total	26	26	30	55	43	
Pk	0.144	0.144	0.167	0.306	0.239	

An agreement table is a table in which rows represent the samples (subjects) and columns represent the number of annotators who assigned the category to the samples. For example, subject 26 is categorized to category 2 (personality disorder) by 5 psychiatrists, and to category 4 (neurosis) by 1 psychiatrist.

The solution of the example is:

$$A_o = \frac{1.000 + 0.400 + 0.400 + \cdots + 0.467 + 0.667 + 1.00}{30} = \frac{16.667}{30} = 0.556$$

$$A_e = (0.144)^2 + (0.144)^2 + (0.167)^2 + (0.306)^2 + (0.239)^2 = 0.220$$

$$K = \frac{0.556 - 0.220}{1.000 - 0.220} = 0.430$$

Fleiss' Kappa's sampling distribution is approximately normally distributed for large sample numbers. However, the significance of the result should be tested when the sample size is not large enough. When it is concluded that the test is not significant, it will mean that a positive value is, nevertheless, as a result of random coding (Artstein, & Poesio, 2008). The significance of the result is tested as below.

The mean approximates to 0 and variance approximates to the following formula (Siegel, & Castellan, 1988):

$$var(\kappa) \approx \frac{2}{HM(M-1)} \frac{A_e - (2M-3)A_e^2 + 2M(M-2)\sum p_k^3}{(1-A_e)^2} \quad \textbf{(Equation 2.13)}$$

Z statistic can be used to test the hypothesis of absence of agreement against hypothesis of existence of agreement.

$$z = \frac{\kappa}{\sqrt{var(\kappa)}} \quad \textbf{(Equation 2.14)}$$

For the above example, z value is,

$$\sum p_k^3 = (0.144)^3 + (0.144)^3 + (0.167)^3 + (0.306)^3 + (0.239)^3 = 0.048$$

$$var(\kappa) = \frac{2}{(40)(6)(5)} \frac{0.220 - [(2)(6) - 3](0.220^2) + (2)(6-2)(0.048)}{(1 - 0.220)^2}$$

$$= \frac{2}{1200} \frac{0.1684}{0.6084} = 0.00046$$

$$z = \frac{0.430}{\sqrt{0.00046}} = 20.02$$

According to Appendix E, Table 22, the resulting z value exceeds the significance level (0.05 where z = 1.64). It is concluded that this test is statistically significant. Because the number of annotations in the TDB is large enough for normal distribution, there will be no need to measure the significance of result for the TDB annotations. (See Appendix C for the java implementation of the Fleiss Kappa.)

2.2.3 KRIPPENDORFF'S ALPHA

Krippendorff (1995) presents several chance-corrected agreement measurement methodologies, which constitute Krippendorff's Alpha agreement coefficient family. The Alpha family and the Kappa family have very similar claims, just like Fleiss' Kappa, Alpha agreement coefficients "are calculated by looking at the overall distribution of judgments without regard to which coders produced these judgments" (Artstein, 2008, p. 564).

Unlike Kappa, Krippendorff does not consider observed and expected agreements but considers observed and expected disagreements (Krippendorff, 1995). α = 1.0 represents exact agreement, α = 0.0 represents exact disagreement:

$$\alpha = 1 - \frac{D_o}{D_e} \qquad \textbf{(Equation 2.15)}$$

However, the distinguishing difference is that Krippendorff's Alpha family enables several agreement measurement approaches by taking into account researcher considerations. The diversity is obtained via several disagreement weighting functions, such as interval, ordinal, ratio, and unit (Krippendorff, 2004a & Krippendorff, 2004c).

The last one of these disagreement weighting functions, Krippendorff's Alpha for unitization is the appropriate agreement coefficient for corpus annotation efforts like TDB in which, annotators perform text span annotations without any previously defined structure (Krippendorff, 2004b; Artstein, & Poesio, 2008). In alpha for unitization, the concern is restricted “to one dimensionally extending continua and assume that agreements are functions of the intersection of these units and that disagreements are a function of the differences between them, all measured by their length” (Krippendorff, 1995, p. 49). Alpha for unitization can be used to measure the following annotation agreement/disagreements: (Diamonds represent annotated segments and regular lines represent gaps. Note that filled diamonds and empty diamonds are just used to differentiate consecutive annotation units.)

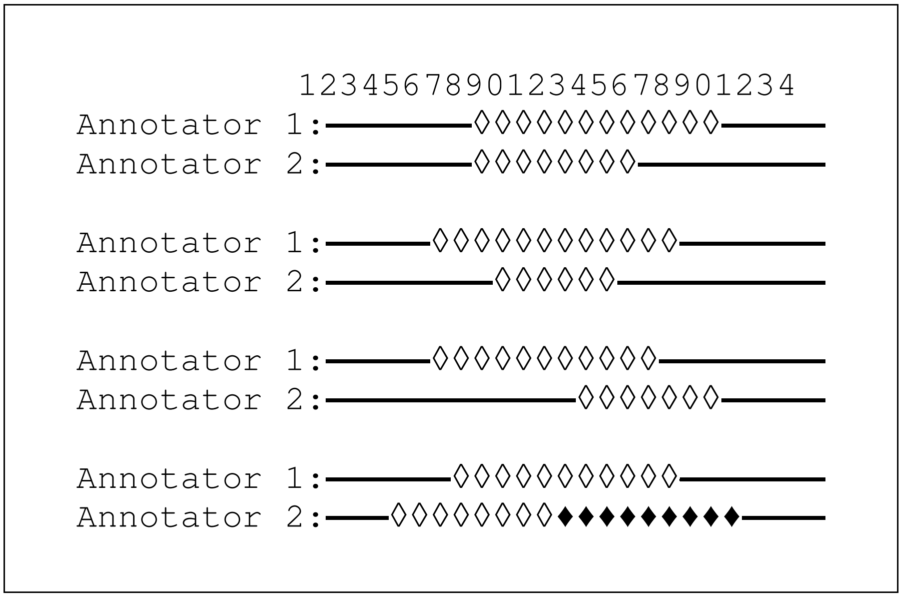


Figure 1 Sample annotation representations that Krippendorff’s Alpha can be used

The weight (distance) between two annotators is calculated with the following function (Krippendorff, 2004b):

$$v_{cig} = \begin{cases} 0\ iff\ section\ \langle cig \rangle\ is\ gap \\ 1\ iff\ section\ \langle cig \rangle\ is\ unit \end{cases} \qquad \textbf{(Equation 2.16)}$$

$$\delta^2_{cigjh} = \begin{cases} \left(b_{cig} + b_{cjh}\right)^2 + \left(b_{cig} + l_{cig} - b_{cjh} - l_{cjh}\right)^2 & iff\ v_{cig} = v_{cjh} = 1\ and - l_{cig} < b_{cig} - b_{cjh} < l_{cjh} \\ l^2_{cig} & iff\ v_{cig} = 1, v_{cjh} = 0\ and\ l_{cjh} - l_{cig} \geq b_{cig} - b_{cjh} \geq 0 \\ l^2_{cjh} & iff\ v_{cig} = 0, v_{cjh} = 1\ and\ l_{cjh} - l_{cig} \leq b_{cig} - b_{cjh} \leq 0 \\ 0 & otherwise \end{cases}$$

(Equation 2.17)

And the weight function for unitization is used as following:

$$D_{oc} = \frac{\sum_{i=1}^{I} \sum_g \sum_{j=1|j \neq i}^{I} \sum_h \delta^2_{cigjh}}{I(I-1)L^2}$$ **(Equation 2.18)**

$$D_{ec} = \frac{\frac{2}{L} \sum_{i=1}^{I} \sum_g v_{cig} \left[\frac{N_c - 1}{3}\left(2l^3_{cig} - 3l^2_{cig} + l_{cig}\right) + l^2_{cig} \sum_{j=1}^{I} \sum_h \left(1 - v_{cjh}\right)\left(l_{cjh} - l_{cig} + 1\right)\ iff\ l_{cjh} \geq l_{cig}\right]}{IL(IL - 1) \sum_{i=1}^{I} \sum_g v_{cig} l_{cig} \left(l_{cig} - 1\right)}$$

(Equation 2.19)

$$*N_c = \sum_{i=1}^{I} \sum_g v_{cig} = the\ total\ number\ of\ units\ of\ category\ c\ identified\ by\ all\ I\ annotators$$

(Equation 2.20)

Above defined Doc and Dec are the observed and expected disagreements for a particular category. In order to calculate the overall Alpha, the sum of each category's Doc and Dec shall be used:

$$\alpha = 1 - \frac{\sum_c D_{oc}}{\sum_c D_{ec}}$$ **(Equation 2.21)**

In the following examples (Krippendorff, 1995), two representative annotation efforts are discussed. Sample texts are digitized into 24 representative segments. Note that the segments shall be the smallest identifiable item. For text span annotation, segments may be characters or words and for audio and video context annotations segments may be seconds. In the first annotation example, Ashley performed one annotation that spans over 18 segments and Arvin performed 7 annotations, where the lengths of annotations are 2, 1, 1, 1, 3, and 1, respectively. Then, in the second annotation example, Bertha performed 4 annotations, where the lengths of annotations are 10, 4, 6, and 4, respectively and Bill performed 5 annotations, where the lengths of annotations are 4, 4, 7, 2 and 7.

```
           123456789012345678901234
Ashley:    ◊◊◊◊◊◊◊◊◊◊◊◊◊◊◊◊◊◊──────
Arvin:     ◊◊♦◊♦◊♦♦♦◊──────────────

Bertha:    ◊◊◊◊◊◊◊◊◊◊♦♦♦♦◊◊◊◊◊◊♦♦♦♦
Bill:      ◊◊◊◊♦♦♦♦◊◊◊◊◊◊◊♦♦◊◊◊◊◊◊◊
```

Figure 2 Two representative annotations for Krippendorff's Alpha

Units and gaps are digitized according to rule of v_{cig} as follows. Underlining is used in order to make explicit the beginning and ending of each annotation:

```
         123456789012345678901234
Ashley:  111111111111111111000000
Arvin:   111111111100000000000000

Bertha:  111111111111111111111111
Bill:    111111111111111111111111
```

Figure 3 Digitized version of the annotations in Figure 2

Alpha agreement coefficient for Ashley-Arvin example is as follows:

$$D_o = \frac{2612.0}{1152.0} = 2.267$$

$$D_e = \frac{\frac{2}{24}25390.0}{1942.0} = 1.089$$

$$\alpha_A = 1 - \frac{2.267}{1.089} = -1.081$$

Alpha agreement coefficient for Bertha-Bill example is as follows:

$$D_o = \frac{510.0}{1152.0} = 0.442$$

$$D_e = \frac{\frac{2}{24}9264.0}{2002.0} = 0.386$$

$$\alpha_B = 1 - \frac{0.442}{0.386} = -0.014$$

Both α are negative, which indicate systematic disagreement, which exceeds chance by disagreement (Krippendorf, 1995). On the other hand, the Fleiss Kappa coefficients of the same examples just indicate a disagreement for the Ashley-Arvin example and indicates an exact agreement for the Bertha-Bill example; $\kappa_A = 0.314$ and $\kappa_B = 1.000$. The main reason for the differences is that the α coefficient considers unitization by using distance function but κ does not have such considerations. Note that the difference between α and κ coefficients does not always change dramatically and they tend to produce similar results when there are no predominantly unitization problems in annotations.

As the last step of inter-annotator agreement measurement, significance of the test can be measured by Z test or an appropriate bootstrapping algorithm can be applied to find the confidence interval (Hayes, & Krippendorff, 2007). But such a measurement is not necessary for the TDB annotations, because of the large number of annotations/samples, as mentioned in § 2.2.2 z value calculation. (See Appendix D for the java implementation of the Krippendorff Alpha for unitization.)

To conclude, the significant properties of the statistical methods, introduced in this Chapter, namely Cochran's Q Test (Q), Fleiss' Kappa (K), and Krippendorff's Alpha (A), are summarized in Table 4:

Table 4 Summary of the statistical methods

	Yes	No
Agreement coefficient	K, A	Q
Statistical test	Q	K, A
Chance-corrected	Q, K, A	-
More than two raters	Q, K, A	-
Nominal scale	Q, K, A	-
Only binary data	Q	K, A
Assumes same distribution for each annotator	K, A	Q
Disagreement weighting function	A	Q, K

CHAPTER III

INTER-ANNOTATOR AGREEMENT APPROACHES OF OTHER CORPUS STUDIES

In this Chapter, five discourse annotation efforts will be elaborated to accurately determine the basis of the inter-annotator agreement approach of a TDB-like text span annotation based discourse corpus that will be devised as the end-product of this study. The following is a list of these five annotation efforts with brief reasons for inclusion in this thesis:

- The Penn Discourse Tree Bank (PDTB):
    - The TDB follows the principles of the PDTB, like annotation efforts from Hindi, Arabic, and Chinese languages.
    - It has a means of inter-annotator agreement measurement method, yet it is deficient.
- The Rhetorical Structure Theory (RST) based annotation effort:
    - It is an effort which commits to inter-annotator agreement issue.
    - It shows the measurement methods can be and should be tailored according to the needs of the specific task.
    - It brings some phenomena which are vital to cross-linguistic studies such as reproducibility and reliability

- The Hindi Discourse Relation Bank (HDRB) and the Chinese Discourse Tree Bank (CDTB):
    - They follow the PDTB, and they both have some diversifications from the PDTB. These differences point out that a cross-linguistic inter-annotator agreement approach shall be able to handle minor aspect changes.
- The Leeds Arabic Discourse Tree Bank (LADTB):
    - It follows the PDTB, with minor language specific differences.
    - The LADTB project group has reported some incomplete and case specific agreement statistics.

The PDTB, the HDRB, the LADTB, and the CDTB, each of a resource for which is different languages sharing the same main principles with the TDB as they all follow the PDTB. Therefore, a common comprehensive inter-annotator agreement measurement approach, which will be clear with the outputs of this thesis, would open cross-linguistic research areas among these efforts.

In the following sections, the PDTB, the RST, the HDRB, the LADTB, and the CDTB will be explained to clarify the inter-annotator agreement approaches.

3.1 THE PENN DISCOURSE TREE BANK (PDTB)

The Penn Discourse Tree Bank (PDTB) is the largest annotated discourse corpus (Prasad, Dinesh, Lee, Joshi, & Webber, 2008a) and the TDB follows the PDTB principles (Zeyrek, & Webber, 2008). Therefore, the PDTB is an essential source for this thesis. In this section, the PDTB will be introduced, discourse annotations will be explained, reported annotation efforts of the PDTB will be discussed, the inter-annotator agreement results and the agreement calculation methodology of the PDTB will be elaborated according to statistical elaborations in Chapter 2, and finally a sound inter-annotator agreement coefficient will be proposed for the PDTB.

3.1.1 INTRODUCTION

The PDTB corpus is the collection of more than 1 million words of Wall Street Journal. This corpus has the sentence-level syntactic annotations of the Penn Tree Bank (Marcus et al., 1993) and the predicate-argument annotations of the Prop-bank (Palmer, Guildea, & Kingsbury, 2005). Hence, the PDTB corpus allows syntactic, semantic and discourse studies by using a single source.

In the PDTB, three discourse aspects are annotated:

1- **The discourse relation annotations** are the annotations that aim to reveal the discourse relations of English. According to the PDTB, discourse relations are aggregation of discourse connective and its two arguments (Arg1 and Arg2), where a discourse connective is a predicate of the relations that establish association between its two arguments. The discourse annotations are performed without commitment to any high-level theory. Thus, the discourse annotations of the PDTB can be used for extensive studies (Prasad et al., 2008a).

2- **The sense annotations** are the annotations that are performed according to a pre-defined hierarchical classification schema, in order to enable sense disambiguation studies.

3- **The attribution annotations** are the annotations that state the relation between discourse elements (discourse connectives and discourse arguments) and the agent (the writer or some other individual). These annotations are used to show discourse level and sentence level correlation (Dinesh, Alan, Miltsakaki, Prasad, Joshi, & Webber, 2005) and to establish subject-related analysis (Prasad, Dinesh, Lee, Joshi, & Webber, 2006).

The main concern of this thesis is measuring agreement on Arg1 and Arg2 (discourse relation annotations). Therefore, the discussions about the PDTB will be limited to these categories.

3.1.2 ANNOTATIONS OF THE DISCOURSE RELATIONS: ARG1AND ARG2

There are two kinds of realization of the discourse relations in the PDTB. The first type is realized by explicit connectives which are lexically identifiable (Webber, Joshi, Miltsakaki, Prasad, Dinesh, Lee, et al., 2006; Prasad, et al., 2008a). Explicit connectives are classified into three syntactical classes: subordinating conjunctions (e.g., when, because, although, so that, if, etc.), coordinating conjunctions (e.g., and, but, or, etc.), and discourse adverbials (e.g., however, for example, then, otherwise, etc.). All explicit connectives bound two arguments which are labeled as Arg1 and Arg2. There is no limitation for the location and span length of the Arg1 and Arg2, and thus, they can be anywhere in the text and they can be single or multiple sentences or clauses. Identifying the Arg2 is relatively easier than identifying Arg1 because Arg2 is mostly adjacent to the discourse connective. Additionally, annotation of the arguments is restricted bys the Minimality principle. The Minimality principle imposes the annotation to be as short as possible to define the discourse relation adequately.

The second type of discourse relations is realized by implicit connectives. Implicit connectives are not identifiable by an overt lexical item such as a discourse connective. These relations are inferred relations between adjacent sentences. The annotation of implicit connectives is performed by selecting the explicit connective that best illuminates the discourse relation. The TDB project group did not annotate implicit relations.

3.1.3 Annotation Guidelines and The Current Version of The PDTB

As Webber et al. (2006) mentioned, the PDTB puts forth important assets. For instance, discourse structure definitions of the PDTB are comprehensive and traceable. Also, the PDTB corpus enables syntactic, semantic and discourse level studies at the same time, which let researchers to conceive the relation between these levels. In addition to these, challenging NLP questions can be investigated by the PDTB. Finally, automatic discourse structure identifications can be performed via the PDTB. Undoubtedly, a corpus which has these important claims shall be reliable. Reliability of a corpus is measured by the quality of its annotations. The annotation quality is directly related with the annotation guidelines, which should be understandable, sound, and well-formed for conscientious annotations.

To ensure high quality annotation, the PDTB presented two annotation guideline versions which annotators followed during their annotations. The first annotation guideline was presented in 2006 by the PDTB research group. The second version (PDTB Manual 2.0) was presented in 2008 and is still in use. In the current annotation guideline, the background theory, the annotation style, the explicit connectives and their arguments, the implicit connectives and their arguments, sense annotation, attribution annotation, and representation format of the PDTB are explained in detail. The PDTB 2.0 was annotated according to the PDTB Manual 2.0. In this latest version of the PDTB, of the 40600 relations annotated; 18459 are explicit connective annotations and 16224 are implicit connective annotations (Prasad et al., 2008a).

Annotations of the PDTB-2.0 were performed in two stages. At the first stage, discourse relations and their arguments were annotated by two annotators. As a result of the first stage, 40600 relations were defined in the entire corpus. At the second stage, annotations that the two annotators did not match were re-annotated by a team. The team consisted of four experts (Prasad et al., 2008a).

3.1.4 Reported Inter-Annotator Agreement Measurement

To the best of knowledge, there is only one inter-annotator agreement measurement reported for the PDTB which is Miltsakaki et al. (2004), where the agreement is measured for a subset of corpus.

The sub corpus mentioned in § 3.1.3 contains 2717 explicit connective annotations and 368 implicit connective annotations. Each connective was annotated by two individual annotators. Then, the inter-annotator agreement was measured by exact match criterion, according to which, 1 was assigned for the annotations that two annotators exactly agree and 0 was assigned for the annotations that the annotators partially or totally disagree. The percentage of assigned 1s to total annotations was used to measure the agreement. The PDTB project group had, first, decided to use the kappa agreement coefficient (Siegel, & Castellan, 1988), because of the chance-correction property of the kappa. However, then, they decided that the discourse annotations were not suitable for the kappa. This decision will be discussed in detail in § 3.1.5, but first the annotations and inter-annotator agreement measurements of Miltsakaki et al. (2004) will be presented.

There are different aspects of explicit and implicit connectives, and thus different inter-annotator agreement measurements. The annotation styles of explicit and implicit connectives are also different. In the sub corpus, 10 kinds of explicit connectives were annotated from two classes; subordinating conjunctions and adverbials. The annotated subordinating conjunctions are *when, because, even, though,* and *so that*. From the adverbial class, *nevertheless, otherwise, instead, therefore,* and *as a result* are annotated. There were two types of inter-annotator agreement measurement methods applied to these explicit connectives. In the first method, agreement was calculated separately for the Arg1 and Arg2 annotations, and in the second method agreement is calculated by counting the Arg1 and Arg2 together. By the first method, the overall agreement was measured as 90.2%, where subordinating conjunction agreement was 92.4% and adverbial agreement was 71.8%. The agreement measurements of the second method were 82.8% for overall, 86.0% for subordinating conjunctions and 57.0% for adverbials.

The inter-annotator agreement among annotators for implicit connectives was measured a little differently from explicit connectives. The argument annotations were measured by the first method of explicit connectives and Arg1 and Arg2 annotations were counted separately. Additionally, 5 groups of explicit connectives were provided to classify implicit connectives into. There were 85.1% agreement measured for the argument annotations, and 72.0% agreement measured for the classification of implicit connectives.

All the presented agreement measurements, except for the adverbial annotation agreements, seemed convincing enough. Moreover, the majority of disagreements (79%) were determined as a result of partial overlap, which was treated as total disagreement according to exact match criteria. In addition, disagreement rate for the adverbials were explained as arguments of adverbials have greater tendency to be non-adjacent to the connective. These claims were used to point out the reliability and plausibility of the PDTB annotations.

3.1.5 A CRITICAL EVALUATION OF THE REPORTED INTER-ANNOTATOR AGREEMENT APPROACH

The PDTB's inter-annotator agreement approach can be evaluated for both its positive and negative aspects, however, eventually, the essential point is whether any improvements can be suggested.

As for the positive side, first of all, the PDTB's inter-annotator agreement was measured on a corpus that was annotated according to an annotation guideline. Without a proper annotation guideline it is infeasible to reproduce the same annotations with different individuals and even with same individuals within different times (Hayes, & Krippendorff, 2007). This is a plus to the reliability tally sheet of inter-annotator agreement measurement approach of the PDTB. Annotator number is also a plus; the annotations were not performed by only one annotator but they were performed by two individual annotators separately. Also, after the annotations, four experts worked over for disagreeing annotations. Finally, the inter-annotator agreement results and methodology were clearly described, which makes their findings re-analyzable.

However, there are some possible negative sides of this approach. First, the exact match criterion is not a chance-corrected agreement method, thus the agreement by chance was not discarded in the PDTB's results. This is the most important question on the soundness of the methodology. The specified reasons of not using the kappa agreement statistic are that the annotated data are not categorized, the text span sizes of the annotations are undetermined for argument annotations, and there are unrestricted expression options for implicit connectives (Miltsakaki et al., 2004).

In fact, categorization is not a problem for argument annotations. All argument annotations are categorized into two categories for both explicit and implicit connectives: the text spans that are arguments and that are not arguments. Here the problem is not categorization but unitization; categories are explicit but it is not clear how to handle and measure the data that are categorized.

Certainly, the kappa statistic (Carletta, 1996) has no solution to offer for this unitization problem. On the other hand, Krippendorff (1995) suggests a chance-corrected agreement coefficient which is based on using the smallest identifiable units to measure inter-annotator agreement when the annotations are performed by selecting text spans. As comprehensively mentioned in § 2.2.3, Krippendorff's Alpha for unitization takes unlimited text spans as the series of characters and measures agreement among these series by using a distance function.

The third drawback of the PDTB's approach to inter-annotator agreement is about the implicit connective annotations. Miltsakaki et al. (2004) indicate that the expression classes that are used to replace implicit connectives are open-ended, thus, these annotations are not suitable for the kappa statistic. On the other hand, in the same paper of Miltsakaki et al. (2004), it is stated that the expressions are restricted to five classes, while analyzing the annotations: *additional information, cause-effect relations, temporal relations, contrastive relations,* and *restatement or summarization*. Basically the expression set may seem unlimited for annotators, but eventually the annotations are categorized into just five categories. Therefore, inter-annotator agreement of implicit connective annotations can be measured by the kappa statistic or by another chance-corrected agreement coefficient such as Krippendorff's Alpha.

That is to say, with some tailoring, the chance-corrected agreement coefficients can be used to measure the agreement on the explicit and implicit connective annotations:

1- Krippendorff's Alpha for unitization is appropriate for evaluating text span annotations.
2- Fleiss' Kappa or Krippendorff's Alpha can be used for the evaluation of implicit connective replacements.

A final comment on the PDTB's approach to inter-annotator agreement is about the future studies. The annotations of only two annotators' agreement can be measured by the PDTB's current approach. A corpus annotation effort performed by three or more annotators cannot be measured by the exact match criterion. Certainly, another measurement method can be used to measure agreement among three or more annotators, but these results cannot be compared with the results obtained from two annotators. Moreover, the PDTB's leading position for discourse corpus studies is indisputable as there are many followers of the PDTB, like the TDB (Zeyrek et al., 2009), the Hindi Discourse Tree Bank (Oza et al., 2009), the Leeds Arabic Tree Bank (Al-Saif, & Markert, 2010), and like the Chinese Discourse Tree Bank (Xue, 2005). These research groups may want to compare their agreement measurement among each other to universally test their theories. Using the same chance-corrected agreement coefficient allows comparison among such corpora.

In conclusion, there are important properties of the PDTB as reported in Miltsakaki et al. (2004): it is annotated by two annotators according to an annotation guideline, the final check of the PDTB annotations is performed via four experts, and the inter-annotator agreement methodologies and the measurements are described clearly. On the other hand, the reported agreement method is not chance-corrected and the used methodology limits the annotator number to two. In the light of the observed positive and negative properties, it can be stated that the PDTB's reported inter-annotator agreement measurement methodology may be improved.

3.2 AN ANNOTATION WORK IN THE FRAMEWORK OF RHETORICAL STRUCTURE THEORY (RST)

The RST is a comprehensive discourse theory that describes discourse relations in a hierarchical structure which was firstly proposed and studied in 80s (Taboada, & Mann, 2006). As Mann & Thompson (1987, 1988) mentioned, the RST framework was successfully used to study several linguistic contentions.

However, beside the place of the RST in discourse linguistics, the main reason for the RST to be discussed in this thesis is the study of Lynn et al. (2003) that emphasize the reliability of corpus annotations. For this thesis, the most important property of the aforementioned corpus is the concern about the reliability. In their work, it was declared that one of the prior concerns is the annotation reliability of the corpus.

Before elaborating on the reliability issues of the corpus, it will be appropriate to introduce the RST, and the annotation process of Lynn et al.'s corpus. Afterwards, Lynn et al.'s agreement measurement methodologies and the measurement results will be elaborated. Finally, the positive and negative sides of this reliability approach will be discussed.

3.2.1 INTRODUCTION

Lynn et al. (2003) summarizes the features of the RST tree format (discourse structure) as follows:

- Text spans, which are the leaves of the tree, shall be minimum blocks and the tree leaves are called *elementary discourse units.*
- Internal nodes of the tree shall constitute continuous text spans.
- There are two categories defined to designate the informative degree of the nodes. A more informative node shall be labeled as *nucleus*, and an ancillary node shall be labeled as *satellite*.
- Nodes shall be connected to each other via a *relation.*

It is relatively easier to perform corpus annotations with the guidance of such a well defined framework. In fact, the framework itself automatically defines the borders of the annotation process. In this manner, first, the annotators shall determine the elementary discourse units, and finally relations and nuclearity of elementary discourse units shall be determined accordingly.

3.2.2 Annotation Process

Before Lynn et al. (2003), Marcu, Amorrortu, & Romera (1999) performed a RST corpus annotation which led to the development of an annotation protocol that Lynn et al. (2003) followed. First this work will be summarized.

3.2.2.1 Annotation Protocol of Marcu et al. (1999)

In their protocol, Marcu et al. (1999) mentioned three constituents of a RST corpus annotation, which they also performed a small size corpus annotation of 90 texts from various sources.

First of all, they suggested using clauses as elementary discourse units. Secondly, they defined 70 rhetoric relations to link elementary discourse units. In some cases, more than one relation may be plausible to annotators. In order to solve these confusions, relations were grouped by considering their meanings, and the relation groups were rated according to their specificity. The annotators were directed to use less specific relations when they hesitated to select the correct relation. Finally, Marcu et al. (1999) proposed a training session apart from the final annotations. In their work, each annotator individually performed the annotation of a same small portion of the target corpus to build experimental discourse structures, as the first stage of the training season. As the second stage, they considered again the elementary discourse structure definition, some rhetorical relations definitions, and the ratings of the relation groups.

3.2.2.2 Annotation Process of Lynn et al. (2003)

In the study of Lynn et al. (2003), annotators were trained before the final annotation work, and performed final annotations with respect to Marcu et al.'s (1999) study. The annotators were linguists who had annotation experiences in different branches of linguistics to increase the success of trainings. During the trainings, the annotators were familiarized with the annotation tool, and they performed individual annotations according to a draft annotation rule set. After the individual work, the annotations were compared. As a result, some examples for elementary discourse unit segmentation, nuclearity determination, and relation selection issues were gathered to guide annotators.

The whole annotation work was performed in about ten months. The 25% (100/385) of corpus annotations were performed in four months according to the training phase examples. After four months, annotation consistency was measured. The measurement results led to some rule enhancements and an annotation approach change. Lynn et al. (2003) decided to reconsider the elementary discourse unit annotations and to pre-segment elementary discourse units by two annotators prior to annotations. In the remaining six months, annotators re-annotated the first portion of the corpus including the remaining of corpus. During this last phase, nucleartiy and relation annotations were performed by using pre-segmented elementary discourse units. Only one illustrative annotation study for 5 documents was performed without any pre-segmented elementary discourse units.

3.2.3 INTER-ANNOTATOR AGREEMENT ON ANNOTATION EFFORT OF LYNN ET AL. (2003)

Lynn et al. (2003) used Marcu et al.'s (1999) kappa statistic methodologies to measure various aspects of the inter-annotator agreement on their RST based corpus. Five topics were presented to fully cover the typical agreement issue of those kinds of corpora. The first topic deals with unit segmentation and the rest of them suggest methodologies for the issues emerging with the hierarchical structure of the corpora. Essentially, in all the methodologies for hierarchical aspects, hierarchical structure was flattened to a linear table by considering each possible segment pairs as units which constitute the source data to compute the kappa statistic. The following is a suitable example, which is a modified portion of a sample annotation from the study of Marcu et al. (1999), to clarify the claim above. In Figure 4, there are two nuclearity segmentation examples for two levels that represent two hierarchical discourse structures of the same text:

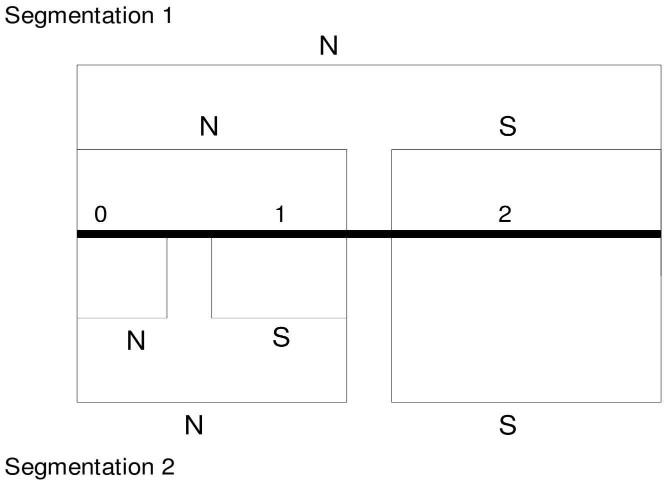


Figure 4 Two sample hierarchical RST discourse structures for the same text. (N=Nucleus, S=Satellite)

As a result of flattening, the following data table is constructed from the discourse structure above:

Table 5 Data table of Figure 4

Segment	Segmentation 1	Segmentation 2
[0,0]	none	N
[0,1]	N	N
[0,2]	N	None
[1,1]	none	S
[1,2]	none	None
[2,2]	S	S

The constructed agreement table is used as the input to the kappa statistic. For this sample the attributes of the kappa statistic are 2 annotators (Segmentation 1, Segmentation 2), 3 categories (N, S, none), and 9 samples (segment pairs).

In the light of this explanation, five inter-annotator agreement aspects are as follows:

1. Unit Level (k_w and k_u): Marcu et al. (1999) present two kinds of kappa statistics to measure agreement on elementary discourse units which are calculated considering two different approaches. In the first case (k_w), it is assumed that the unit boundaries can be the end of any word. The second case (k_u) suggests taking the unit boundaries as the locations that at least one annotator annotated as boundary. The two approaches have different chance factors because units and unit numbers those are included in measurement changes. The change of chance factor directly affects the results. In Marcu et al.'s (1999) sample corpus, measurements of k_w are around 0.90 while k_u measurements are around 0.75. This is a nice example that illustrates that the results depend on not only on the selected statistical methodologies but also on their application manner.

2. Spans Level (k_s): This statistic suggests measuring the hierarchical discourse segment annotations.

3. Nuclearity Level (k_n): This statistic suggests measuring the hierarchical nuclearity annotations.

4. Relation Level (k_r): This statistic suggests measuring the hierarchical rhetorical relation annotations.

5. Group of Relations Level (k_{rr}): This statistic suggests measuring the hierarchical rhetorical relation annotations when the relations are grouped according to their rhetorical similarity.

Lynn et al. (2003) presented agreement results for all of Marcu et al.'s (1999) inter-annotator agreement statistics. There were two kinds of results: in the first result set, the evolution of agreement among raters by time was illustrated, and in the second result set, the final annotation agreements in various corpus subset annotations that were performed by different annotator pairs were presented. Both of the results were the agreements among two to three annotators for about 30 documents. As mentioned before, Lyn et al. (2003) preferred to use pre-segmented elementary discourse units during their final annotation, except the training session annotations and the illustrative final annotation. They presented agreement results for unit levels (for pre-segmented and not pre-segmented elementary discourse units). However, they did not indicate whether they had used k_w or k_u to calculate. For the spans level, nuclearity level, relation level and group of relations level they closely followed Marcu et al. (1999).

In Lynn et al's (2003) measurements, the inter-annotator agreement that changed among time indicates that there were 0.10 to 0.15 increase during the annotation for all levels. Unit level agreement results that were performed on not pre-segmented text increased from 0.87 to 0.97. For pre-segmented data, unit level agreement was expected 1.00, but the measurements are between 0.95 and 1.00. The reason of the lesser agreement than expected agreement was explained as the annotators' misusing of the annotation tool. As seen, exterior or human factors may change the reliability of annotated corpus even if the task was clearly defined. At the spans level, agreement increased from 0.77 to 0.89. At the nuclearity level, agreement increased from 0.70 to 0.85. At the relation level, agreement increased from 0.60 to 0.75. Finally, agreement increased from 0.64 to 0.78 at the group of relations level. These results indicate that the reliability of this corpus can be sustained for all levels (unit, spans, nuclearity, relation, and group of relations) by training annotators. Another important inference from the results is that annotating grouped relations or annotating individual relations does not significantly affect the reliability of the annotations.

The inter-annotator agreements of the final annotations were measured by the various subsets of the corpus. In these subsets, there were 6 annotation sessions which were performed by two annotators. During each annotation session, 4 to 7 distinct texts were annotated. The average kappa value of each level was as follows:

Table 6 Summary of Lynn et al.'s (2003) agreement measurements

Level	Average Kappa
Unit	0.97
Spans	0.86
Nuclearity	0.80
Relation	0.72
Group of Relations	0.75

These results point out that the reliability of this corpus is repeatable under varying settings. Again, these results show that there is no significant difference between annotating relations by grouping them or not.

After this comprehensive review of the properties of Lynn et al.'s (2003) corpus, the positive and negative aspects of Lynn et al.'s (2003) inter-annotator agreement approach can be examined.

3.2.4 A CRITICAL EVALUATION OF THE LYNN ET AL.'S (2003) INTER-ANNOTATOR AGREEMENT APPROACH

There are lots of properties that make Lynn et al's (2003) approach re-usable. Firstly, their approach is supported with well-developed annotation guidelines which are based on a well-defined discourse theory. The annotation guidelines are brought to maturity by the training sessions. The training sessions not only advanced the annotation guidelines but also the annotators. Therefore, including training sessions in the annotation protocol is a double win. As another note worthy aspect, Lynn et al. (2003) conducted all their annotations with at least two individual annotators in order to apply chance-corrected inter-annotator agreement measurement methodologies. They presented adjusted chance-corrected statistic (the kappa statistic) according to the needs of their theoretical framework. The kappa statistic is implemented in four different ways to meet the needs of the RST-based hierarchical discourse structure. Moreover, they explicitly presented the results and the corresponding methodologies with the negligible exception that they did not mention the exact kappa statistic used to measure the unit level agreement. Finally, they presented two sets of results, one of which is useful to monitor the effect of trainings, and the other is useful to monitor the repeatability of the annotations among different texts and annotators.

Besides such important suggestions and implementations, Lynn et al's approach is not flawless. The problems can be classified into two groups: which are inherited from Marcu et al.'s (1999) study, and those which emerged in the study of Lynn et al. (2003). The problems inherited from Marcu et al. (1999) are also discussed in the study of Marcu et al. (1999). All the problems raised by Marcu et al. (1999) will not be presented here, only the problems that will contribute to the ultimate goal of this thesis will be presented. The most noticeable problem is that the agreement levels (units, spans, nuclearity, relation, and group of relations) are independent from each other. With this approach, it is impossible to assign the source level of disagreement. For instance, wrong unit segmentation may lead to false nuclearity annotations, but may not affect the other levels' annotation. On the other hand, in another wrong unit segmentation case, all levels' annotation may be affected. Secondly, because of the hierarchical nature of the annotations, there exist lots of *none* annotations where in fact annotators do not perform annotations. As Marcu et al. (1999) explained, for a hierarchical discourse tree with *n* leaves, there will be *n(n+1)/2* nodes, and *2n-1* of these nodes will be different than *none*. Such a big amount can artificially affect the agreement in the positive manner. The final problem which is inherited form Marc et al.'s (1999) study is that all annotations in all levels of the tree equally affect the agreement results.

The second group of problems in the study of Lynn et al. (2003) may be summarized as follows. First, segmentation is not considered as an annotation problem so some disagreement types that can occur during segmentation are suppressed. Second, pre-segmentation of the text may affect the annotators' decision.

In conclusion, Lynn et al.'s approach is not flawless but it is obvious that they present a well defined inter-annotator agreement measurement approach with different measurement approaches for different annotation issues.

3.3 THE HINDI DISCOURSE RELATION BANK

In this section, the current and projected HDRB corpus will be described, the differences of the HDRB and the PDTB, and how the HDRB can be used for cross-linguistic comparisons will be discussed. Afterwards, the reason of the need for a common inter-annotator agreement approach in all these efforts will be mentioned. Finally, the effects of the diversity between the HDTB and the PDTB on the possible inter-annotator agreement approach will be discussed.

The Hindi Discourse Relation Bank (HDRB) project group is one of the followers of the PDTB. The HDRB project group aims to develop a large discourse corpus which is based on the PDTB's lexical approach (Oza, Prasad, Kolachina, Meena, Sharma, & Joshi 2009a; Oza et al. 2009b; Prasad, Husain, Sharma, & Joshi, 2008b).

The HDRB is a 200K word corpus where the texts are taken from 400K word Hindi dependency corpus. The texts of the corpus belong to the newspaper articles from several genres such as: politics, sports, films, and so on. As the future goal, the group aims to extent the HDRB to the whole 400K corpus in order to conduct cross level (discourse-syntactic) research (Begum, Husain, Dhwaj, Sharma, Bai, & Sangal, 2008).

Currently, the HDRB has annotations of explicit and implicit connectives with their arguments, and sense of discourse relations. They have several differences from the PDTB. For explicit connectives, the HDRB uses *sentential relative, subordinator,* and *particle* grammatical classes in addition to *subordinating conjunction, coordinating conjunction*, and *adverbial* grammatical classes of the PDTB. As another difference, implicit relation annotations are paragraph internal in the PDTB, however in the HDRB there is no such a restriction. The sense classes also differ. The difference is not at the top level of sense classes but at the second and third levels. The HDRB project group preferred to re-organize the lower level sense classes in order to capture senses more accurately. Except for these mentioned divergences, the HDRB completely follows the PDTB's discourse approach and the PDTB's annotation guidelines (Begum et al., 2008).

Such parallel discourse implementations of different languages provide important opportunities for cross-linguistic discourse research. The HDRB project group declares that their work will contribute to this research area. Currently, they have two cross-linguistic claims which are exhibited as a result of an initial annotation experiment. In the annotation experiment, a sub-corpus, which has 35 texts from the HDRB corpus, is annotated where only explicit connective annotations are performed. The experiment shows that there are no significant differences between the HDRB and the PDTB annotations in the distribution of the discourse relations and sense annotations. In their first claim, they argue that the morphological properties of a language do not affect the connective usage. In contrast to English, Hindi language is a morphologically affluent language. However, the distribution of discourse relations (*explicit, implicit, AltLex, EntRel*, and *NoRel* relations) among the total annotations in the HDRB is very similar to the PDTB annotations. Secondly, the HDBR's sense distributions are very similar to the PDTB's sense distributions at the top level sense classes as mentioned above.

One would also want to conduct cross-linguistic research not only in the light of annotation distributions among classes but also in the light of inter-annotator agreement measurement results. First of all, without the reliability analysis of a corpus, a comparison of any corpus data, as the presented distribution data, is dubious. Furthermore, cross-linguistic research can be diversified and refined with a shared inter-annotator agreement approach. A shared approach would serve to analyze the agreement/disagreement characteristics among Hindi and English languages which can lead to comprehensive super-language theorizations.

On the other hand, the annotation decisions in different corpus annotation efforts are not exactly identical, e.g. due to linguistic properties of different languages. Thus, the inter-annotator agreement results cannot exactly be the same, even with a shared approach because the annotation process and annotations themselves are directly affected by annotation guidelines and annotators educated for the purpose. It is obvious that these will limit the comparability of the results, but it is also obvious that this will not completely remove comparability of results. With a shared agreement measurement approach, there would still be a wide area to conduct research, because the different annotation decisions are not major distinctions but minor adjustment decisions.

The HDRB and the PDTB can enable wide ranged cross-linguistic research on discourse when they are supported by shared inter-annotator agreement aspects. However, currently the HDRB has no reported inter-annotator agreement measurement approach.

3.4 THE LEEDS ARABIC DISCOURSE TREE BANK (LADTB)

The Leeds Arabic Discourse Tree Bank (LADTB) is the first effort which aims to produce an annotated discourse level Arabic corpus. Like the HDRB (see § 3.4.), the LADTB follows the principles of the PDTB (see § 3.3.) with some language specific modifications to the discourse approach of the PDTB. Different from the HDRB, the

LADTB presents some agreement results and methods to measure agreement. That is to say, the LADTB is discussed in this thesis for two reasons, first it follows the principles of the PDTB, and so its discourse considerations are almost same with the TDB's considerations. Secondly, the LADTB project group addresses inter-annotator agreement as a research issue (Al-Saif, & Markert, 2010).

In this section, the LADTB will be introduced, the annotation methodology of the LADTB will be presented, and the inter-annotator agreement measurement methodology will be discussed, respectively. Finally, the future work plan of the LADTB will be presented with some comments on the proposed inter-annotator agreement measurements.

The LADTB is a corpus which is built on Arabic Penn Tree Bank v.2 (Maamouri, & Bies, 2004). Currently, The LADTB contains a portion of Arabic Penn Tree Bank texts (537 news texts) which enable cross-level (syntactic-discourse level) research. In the LADTB, explicit connectives were annotated with their arguments. Also, the senses of discourse relations were annotated for these connectives. However, the LADTB does not contain any implicit discourse relation nor attribution annotations.

As mentioned before, there are several different annotation decisions from the PDTB as a result of language specific properties. First of all, Arabic is a morphologically abundant language where morphemes can function like words (clitics). Therefore, some clitics are needed to be considered as discourse connectives in the LADTB. Secondly, the LADTB project group prefers to re-organize the lower classes of discourse sense relations as the HDRB does. They do not change the first level classes (*Expansion, Contingency, Temporal,* and, *Comparison*) but they shrink the lower level classes and add two new lower level classes (*Similarity,* and *Background*) according to needs of current version of corpus.

The LADTB annotations were performed according to the PDTB’s annotation guideline. The annotations were performed by two independent trained native speaker annotators, who were not involved in any other preparation work of the corpus. The LADTB group developed an annotation tool to perform annotations and to handle their work specific requirements, such as right to left writing order and clitics annotations. The annotation tool also serves to collect research specific inter-annotator agreement data.

There are two presented inter-annotator agreement in the LADTB. First concern is mostly Arabic specific; the agreement on assigning clitics to discourse connective labels is measured. Secondly, agreement on the sense annotations of the discourse relations is measured. The kappa results are available for both measurements. The kappa results show that annotations are very reliable (0.88) and almost reliable (0.57) for the first and the second concerns, respectively. Additionally, Al-Saif and Markert (2010) present Krippendorff’s Alpha (0.58) for the second concern. It is mentioned that a variation of Krippendorff’s Alpha which enables partial agreement on annotations of sets is used.

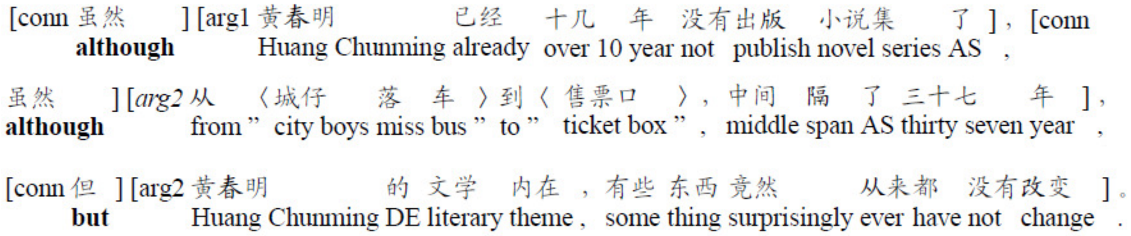
Finally, the LADTB project group declares that they are planning to measure inter-annotator agreement among the connective arguments. Yet, there is no proposed agreement statistic supporting for this intention.

As seen, the LADTB project is a potential cross-linguistic source because it shares same principle with discourse studies from various languages (the PDTB-like efforts). The LADTB’s agreement approach is very positive. First of all, the LADTB project group present chance-corrected inter-annotator agreement results for two goal specific concerns where the annotations are performed by two independent individual annotators according to a guideline. However, the mentioned methodologies to measure agreement are not clear to re-perform measurements. In addition to this, so far there is no proposed method to measure agreement on argument location and span annotations, which is the most challenging agreement issue of the PDTB-like efforts.

3.5 The Chinese Discourse Tree Bank (CDTB)

In 2005, Xue introduced Chinese Discourse Tree Bank (CDTB) project. The CDTB is an explicit discourse annotation project which follows the PDTB's principles. Yet, the CDTB project is not completed. The study of Xue (2005) is a prelude which emerges as a response to the challenge of argument annotation in Chinese.

As most other discourse annotation efforts, the CDTB aims to add a discourse annotation level to already the syntactically annotated corpus (Xue, Xia, Chiou, & Palmer, 2005). The CDTB promises cross-linguistic research opportunities. The CDTB follows the PDTB's broadly accepted principles. Inevitably, there are some minor language specific differences from the PDTB. The most striking difference is in the realization of the subordinating conjunctions. In Chinese, subordinating conjunctions can be conjoined, where there are two instances of a subordinating conjunction:



[conn 虽然] [arg1 黄春明 已经 十几 年 没有出版 小说集 了]，[conn
although Huang Chunming already over 10 year not publish novel series AS ,

虽然] [*arg2* 从 〈城仔 落 车 〉到〈 售票口 〉，中间 隔 了 三十七 年]，
although from " city boys miss bus " to " ticket box " , middle span AS thirty seven year ,

[conn 但] [arg2 黄春明 的 文学 内在 ，有些 东西 竟然 从来都 没有改变]。
but Huang Chunming DE literary theme , some thing surprisingly ever have not change .

The CDTB project group has an annotation guideline which is mostly the same as the PDTB's and also they discuss preliminary decisions, for example, the list of discourse connectives in scope, distribution of the discourse connectives in Chinese and sense disambiguation. The CDTB is a potential well-formed discourse study; however there is no reported methodology to measure inter-annotator agreement.

CHAPTER IV

THE TURKISH DISCOURSE BANK

The Turkish Discourse Bank (TDB) is a discourse annotation project which follows the PDTB's (Prasad, Dinesh, Lee, Joshi, & Webber, 2007) principles. The TDB aims to expand the METU Turkish Corpus to discourse level. In the beginning, 500k words sub-corpus of the METU Turkish corpus will be annotated, which is a sub-corpus with texts in various genres such as, fiction, interview, memories, and news articles (Zeyrek et al., 2009). The discourse annotations will include annotations of discourse connectives with their two arguments, modifiers, and supplementary text spans when the TDB is released by the end of 2010. Like the PDTB, the TDB takes discourse connectives as the discourse level predicates that relate two arguments, Arg1 and Arg2. Currently, 60 discourse connective types are annotated with approximately 7000 argument annotations (Zeyrek, Demirşahin, Sevdik-Çallı, Ögel-Balaban, Yalçınkaya, & Turan, 2010). Such an annotation effort that is performed according to a nonrestrictive discourse theory brings about reliability issues, because of the annotated data amount, annotator number, and text span annotations.

In this Chapter, the TDB and its reliability issue will be elaborated in depth as follows.

1- The types of Turkish discourse connectives, which are annotated in the TDB, will be presented.

2- The annotation process of the TDB will be examined along with the annotation guidelines and the annotation tool.

3- The dependency analysis of the TDB annotations will be presented.

4- The data representation of the TDB annotations will be explained.

5- The challenges in handling the TDB annotation data will be elaborated on.

6- The ideas to handle data challenges will be discussed in detail, where the ideas are adapted from context analysis (Krippendorff, 2004a) to the TDB during the study of this thesis.

7- How to convert annotation data as input to the inter-annotator agreement methods that was mentioned in Chapter 2 will be described, with the help of presented data handling ideas.

8- Finally, already reported agreement results on the TDB will be discussed.

4.1 Turkish Discourse Connectives

The TDB takes discourse connectives as discourse level predicates that relate two arguments, Arg1 and Arg2, just like the PDTB. In the TDB, there are two kinds of discourse connective realizations: explicit and implicit connectives. Explicit connectives are lexically identifiable discourse items. On the other hand, implicit connectives can only be identified via the arguments that they relate. In the current scope of the TDB project, explicit connectives are annotated with their arguments, and the implicit connectives are left for subsequent studies (Zeyrek et al., 2010). In this section, the Turkish discourse connective classes will be briefly introduced with sample annotations. (In the samples, discourse connective annotations are illustrated via underlining the items, Arg1 annotations are illustrated via *italicizing* the items, and Arg2 annotations are illustrated via **bolding** the items.)

In the PDTB, explicit discourse connectives are grouped as coordinating conjunctions, subordinating conjunctions, and discourse adverbials. These classes are not completely covering syntactic aspects of the Turkish discourse connectives. Therefore, Zeyrek and Webber (2008) classified Turkish discourse connectives into five groups by considering the Turkish morpho-syntactic properties:

(1) *Simple Coordinating Conjunctions*: They relate the arguments which belong to the same syntactic type. (çünkü ‘because’, ve ‘and’, ya da ‘or’, ama ‘but’).
Ex.: *Yapılarını kerpiçten yapıyorlar, ama sonra taşı kullanmayı ögreniyorlar. Mimarlık açısından çok önemli,* çünkü **bu yapı malzemesini baska 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**.’

(2) Paired Coordinating Conjunctions: They are the discourse connectives which constitute two lexical items. (ya … ya ‘either … or’, hem … hem ‘both … and’, gerek … gerek(se) ‘either … or’).
Ex.: *Birilerinin* ya *işi vardır, aceleyle yürürler,* ya **koşarlar.**
‘*Some people are* either *busy and walk hurriedly*, or **they run.**’

(3) Simplex Subordinators: They are in the form of suffixes, and they take place at the end of main verbs. (-yArAk, ‘by means of’, -Ip ‘and’, -kEn ‘while, wherease’).
Ex.: Elektrik enerjisi üretiminde kömür kullanımı **Yunanistan'da yüzde 71** iken, *Türkiye'de yüzde 29*.
‘Use of coal in electricity generation is *29 percent in Turkey*, while **71 percent in Greece**.’

(4) Complex Subordinators: They are combination of a lexical item, and a suffix which conjoin the preceding verb. (-dIğI için ‘since’, -dAn once ‘before’, -Ir gibi ‘as if, as though’).

Ex.: **Herkes çoktan pazara çıkTIGI** için *kentin o dar, eğri büğrü arka sokaklarını boşalmış ve sessiz bulurduk.*
Since **everyone has gone to the bazaar long time ago**, *we would find the narrow and curved back streets of the town empty and quiet.*'

(5) Anaphoric Connectives (Discourse Adverbials): They are mostly misinterpreted as clausal adverbials; however in addition to matrix sentence anaphoric connectives they also need abstract object to relate with matrix sentence. (sonuç olarak 'consequently', aksine 'on the contrary', mesela 'for example').
Ex.: *Ali hiç spor yapmaz*. Sonuç olarak çok istedigi halde **kilo veremiyor**.
'*Ali never exercises*. Consequently, **he can't lose weight** although he wants to very much.'

As another difference from the PDTB, the TDB project group annotates non-finite clauses as arguments, because in Turkish, all non-finite clauses are related with an abstract object via suffixes.

In Turkish, connectives can be sentence-medial, sentence-initial, and sentence-final while in English connectives are sentence-initial or sentence-medial. For instance coordinating conjunctions may appear sentence-initially or sentence finally, whereas subordinators are always at the end of its Arg2 but the position of the Arg2 in a sentence can change. Therefore, subordinators may be sentence-medial or sentence-final. No doubt, this property of Turkish makes annotation process more challenging than English (Zeyrek et al., 2008; Zeyrek et al., 2009).

4.2 ANNOTATION CYCLE

All annotations in the TDB are performed according to an annotation guideline, which is a vital instrument for a reliable annotation effort (TDB-Group, 2010). Also, all annotators used an annotation tool (Aktaş, Bozşahin, & Zeyrek 2010) which enables utilities, such as browsing prior annotations and searching connectives with regular expressions. The following section introduces the annotation cycle of the TDB by detailing the above issues.

4.2.1 Annotation Guidelines

The current revised version of the annotation guidelines of the TDB was distributed to the project members in March 2010. In the guidelines, general annotation principles, and exceptional annotation issues are described mostly following the PDTB annotation guidelines (PDTB-Group, 2006 & 2008).

Major principles include what Turkish discourse connectives are, and where the arguments (ARG1 and ARG2) shall be searched. Annotation issues concerning the following exceptional cases of the TDB are also described in the annotation guidelines (TDB-Group, 2010):

- Minimality principle (annotations shall contain minimum text spans enough to fully cover the discourse)
- Annotation connectives with a listing function
- Annotation of texts whose sentences/clauses are interrupted by punctuation marks
- Annotation of shared arguments
- Structures that shall not be annotated because of the absence of abstract object interpretation

In brief, the TDB annotation guidelines form a sound basis for annotation efforts.

4.2.2 ANNOTATION TOOL

The TDB annotations are performed by a computer program which is named as DATT (Discourse Annotation Tool for Turkish). The DATT is a XML based stand-off annotation tool (Aktaş et al., 2010), which eases the work of researchers both at the annotation phase and at the post annotation research with the following properties:

- Unlike in-line annotation tools, the DATT enables the annotation of shared, nested, and crossing arguments (see § 4.3).
- The annotations are kept in XML files which enable distribution of annotations without the source text files. XML files are also handy files for post annotation research, such as inter-annotator agreement measurements.
- The Stand-off annotation is useful to investigate annotations layer by layer (Zeyrek et al., 2009).
- The DATT is a user friendly tool which provides regular expression searches and traceability of annotations. These properties also increase the quality of the annotation process.

4.2.3 ANNOTATION PROCESS

Prior to the annotation process, the annotators were trained with the annotation guideline. The pursued annotation process of the TDB can be explained in the following four steps (Zeyrek et al., 2009):

1- Annotations were performed for a particular connective by three or two annotators. The annotators performed a particular discourse connective and its arguments' annotations file by file for the whole corpus files. In this step each annotator worked individually.

2- Afterwards, individual annotations were compared. Then, the disagreements found were discussed, and solved by the project group.

3- Annotation guideline was revised according to the discussions.

4- The agreed annotations were checked if they completely obey the annotation guideline.

The above annotation process was cycled for all discourse connectives. However, in later phases of the annotation effort, the TDB group decided that the inter-annotator reliability has stabilized, and they switched to a more rapid annotation strategy. In the new strategy, the TDB group kept the annotation processes same, except the annotators. According to the new strategy, a pair of annotators and an individual annotator (practically two annotator teams) performed the annotations (Demirşahin, Yalçınkaya, & Zeyrek, 2010).

In the process defined above, the inter-annotator reliability shall be measured right after the first step because by the second step, annotator's individual decisions are judged and corrected. Thanks to the version control software that the TDB group uses, the annotation data, which were produced at each step of the annotation process, can be retrieved easily.

4.3 DEPENDENCY ANALYSIS OF DISCOURSE RELATIONS

Various annotation cases, such as overlapping and nested arguments, can occur in the TDB. Aktaş et al. (2010) introduced the configurations that seen in the TDB by using terminology of Lee, Prasad, Joshi, Dinesh, & Webber (2006):

- Independent Relations: The discourse connectives and their arguments are unrelated:

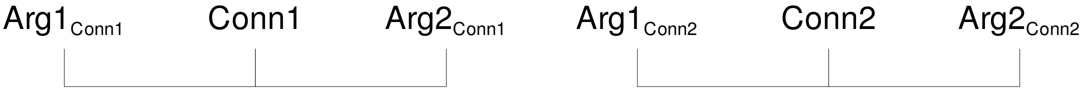


Figure 5 Independent relation case of Turkish discourse structure

- Full Embedding: A discourse structure constitutes argument of an another discourse structure (relation):

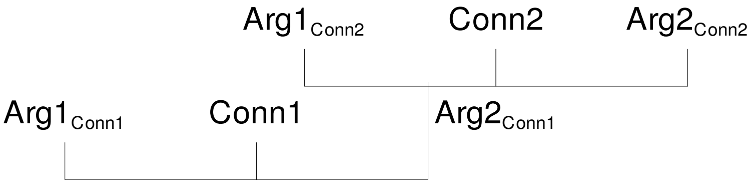


Figure 6 Full embedding case of Turkish discourse structure

- Shared Argument: Two different discourse structures can share the same argument:

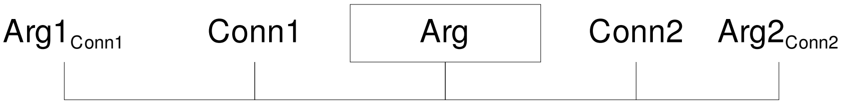


Figure 7 Shared argument case of Turkish discourse structure

- Properly Contained Argument: The argument of a discourse structure can include the argument of another discourse structure with extra text:

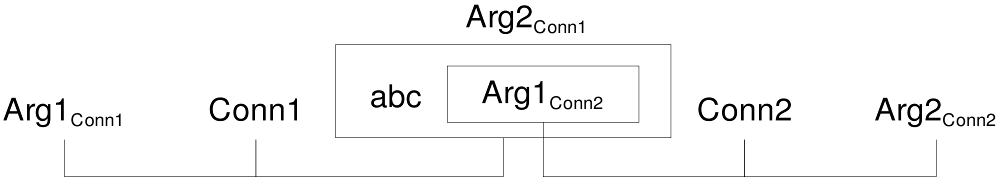


Figure 8 Properly contained argument case of Turkish discourse structure

- Properly Contained Relation: The argument of a discourse structure can include another discourse structure with extra text: (This discourse structure realization is special to Turkish.)

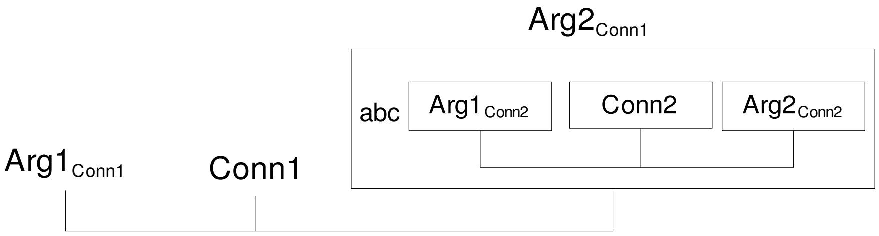


Figure 9 Properly contained relation case of Turkish discourse structure

- Nested Relations: A discourse structure can be embedded into another discourse structure: (This discourse structure realization is special to Turkish.)

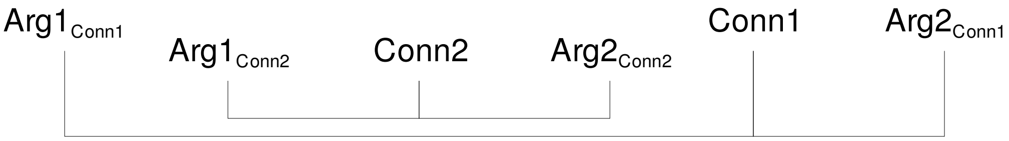


Figure 10 Nested relation case of Turkish discourse structure

- Pure Crossing: A discourse structure can cross over with another discourse structure:

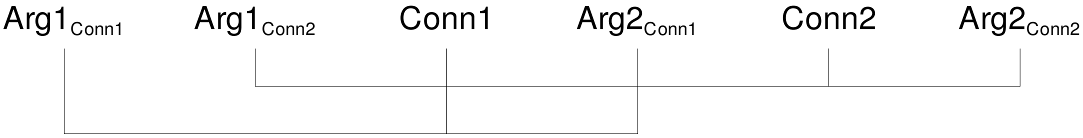


Figure 11 Pure crossing case of Turkish discourse structure

Not considering the cases above shall lead to a misinterpretation of the annotation data. Therefore, while measuring the inter-annotator agreement on the TDB, one should consider the data handling challenges due to the above cases.

4.4 DATA REPRESENTATION

As mentioned in § 4.2.2, the TDB annotations are performed via the DATT, where annotations are stored as stand-off XML files. Each stored annotation file (XML file), contains one annotator's annotations for a particular connective. For an example corpus, where there are 5 source text files, and where 4 types of connectives are annotated by 3 annotators, there would be 60 (5x4x3) annotation files.

As seen in example above, the annotation file number can grow in three dimensions: annotator number, connective number, and source file number of the corpus. Such an expansion tendency necessitates well defined annotation files for inter-annotator agreement measurements. Parallel to this need, the DATT produces well formed and easily usable XML files. In the produced annotation files, each annotation is kept in a record where the discourse elements are delineated by the character offsets from the beginning of the source files (Aktaş et al., 2010). The annotation record that is created for the following sample annotation will be used to explain the data representation in the TDB:

Akıntıya kapılıp umulmadık bir geceyi bölüştü benimle ve **bu kadarla kalsın istedi belki**.

'*She was drifted with a current and shared an unexpected night with me* and **perhaps she wanted to keep it this much only**.'

The following annotation record is produced in the corresponding annotation file (In the record, the <Conn> tag identifies the connective, and <Arg1> and <Arg2> tags identify the arguments. The <Span> tag represents each selected text span for a discourse element. More than one <Span> tag can be inserted to represent the cases which are mentioned in § 4.3. In each <Span> tag, annotated text (<Text>), and its beginning and end offsets (<BeginOffset>, <EndOffset>) are presented. All these tags are enclosed by the <Relation> tag which also identifies the type of relation.):

```
<Relation type="EXPLICIT">
    <Conn>
        <Span>
```

```
            <Text>ve</Text>
            <BeginOffset>260</BeginOffset>
            <EndOffset>262</EndOffset>
        </Span>
    </Conn>
    <Arg1>
        <Span>
            <Text>
                Akıntıya kapılıp umulmadık bir geceyi bölüştü benimle
            </Text>
            <BeginOffset>206</BeginOffset>
            <EndOffset>259</EndOffset>
        </Span>
    </Arg1>
    <Arg2>
        <Span>
            <Text>
                bu kadarla kalsin istedi belki
            </Text>
            <BeginOffset>263</BeginOffset>
            <EndOffset>293</EndOffset>
        </Span>
    </Arg2>
</Relation>
```

In addition to the XML attributes above, there are several attributes that contribute to complete the definition of the annotations, such as <Sup1>, <Sup2>, <Mod>, and <Shared> tags (Zeyrek et al., 2010). The information represented with these tags is out of the scope of this thesis.

4.5 Data Handling Challenges

The TDB project group currently aims to annotate 197 source files with at least 2 annotators for about 60 discourse connectives. When it is assumed that each connective was annotated 4 times on average in a source file by 2 annotators, there would be about 100k Arg1 annotations and 100k Arg2 annotations. The biggest data handling challenge in the TDB is the amount of the data. However, this is not the only challenge. The following is the list of data handling challenges concerning inter-annotator agreement measurement:

Challenge 1: The annotation data are too much to handle manually.

Challenge 2: There maybe data corruptions because of the annotation process or the annotation data storage process, such as missing files and missing connective annotations.

Challenge 3: One file is produced for each annotator's each connective annotation, which means annotations span over many files. In order to measure the inter-annotator agreement among all annotations, these files must be combined by a computationally valid way.

Challenge 4: The annotation data of the TDB are text spans. In order to measure inter-annotator agreement, these text spans must be transformed into discrete computational data units.

Challenge 5: The dependencies in the annotations (see § 4.3) may be computationally problematic for inter-annotator agreement measurement.

All of the above challenges are handled in different stages of this thesis, and the solution approaches of these challenges are discussed in different sections, as shown below:

Challenge 1, 2, & 3: The inter-annotator agreement measurements in this thesis are performed via a computer program, which is an original final product of this thesis. The computer program is capable of performing automated measurements by combining annotation files. Also, it is capable to take preventive actions for data corruptions, which means the program performs a validation procedure before including an annotation file in the measurement. The details of the computer program are presented in § 5.4

Challenge 4: The data unitization issue for the TDB is elaborated in § 4.6 in detail.

Challenge 5: The computational problems, which occur as a result of discourse dependencies, are handled after unitization of the annotation data. How this challenge is handled is in § 4.7.

4.6 DATA UNITIZATION

The collected data shall be represented in purpose-specific qualitative or quantitative units to enable computational analyses and inter-annotator agreement measurement. The TDB is a corpus which is a collection of annotated written texts. Therefore, data units shall be the elements of written texts. Besides, the following questions arise: which elements shall be used, how they should be represented, and where they should start and end. These are indirectly answered by Krippendorff (2004a, Chapter 5) in a content analysis perspective.

The content analysis is a form of textual analysis which aims to reveal the lingual patterns. Just like discourse analysis, it is based on corpus annotation. However, content annotation efforts differ from discourse annotation efforts in two main ways. First of all, in content annotation efforts, there are deliberate content variable units (lingual patterns), which the annotators are supposed to annotate. In contrast, in discourse annotation efforts, like the TDB, the content variable units (annotation subjects) are not previously determined, but they are determined as a result of the annotation process, which is performed by annotators who act according to an annotation guideline and the opinions of the annotators are still decisive. Secondly, in the content annotation efforts, the units are limited by the structure definitions of the patterns. However, in discourse annotation efforts, units are not limited, and they can span from utterances to the whole text (Truex, 2006). On the other hand, both content and discourse annotation efforts use units as research handles. Therefore, at first, both need to define types of units and ways of defining units. Afterwards, context annotation efforts use these definitions to find patterns, and the other discourse annotation efforts use the definitions to guide and train the annotators.

In brief, with minor approach differences, Krippendorff's (2004a, Chapter 5) elaborations on data unitization perfectly fit into discourse annotation efforts, and thus into the TDB. Krippendorff classifies units that are used in content annotation efforts into three types:

1- Sampling Units: An object-specific subset of units that are selected to represent the whole.
2- Coding Units: The smallest units that are enough to define all the information needed in content analysis.
3- Context Units: The units that limit the text contextually. Context units use coding units to specify their spans.

According to Krippendorff, the types of units are defined by using one or more of the following five distinction criteria:

1- Physical Distinctions: In content annotation efforts, physical distinctions are performed according to physical quantities such as time, length or volume. Generally, physical distinctions do not directly provide information for content analysis.
2- Syntactical Distinctions: These distinctions are based on syntactic definitions such as, sentences, paragraphs, chapters or journal articles.
3- Categorical Distinctions: These distinctions are performed according to membership in a class or category.
4- Propositional Distinctions: The proposed structures are used for distinctions. For example, proposing clauses as units.
5- Thematic Distinctions: These are motif-based distinctions that enable narrative comparability.

When the above classifications are applied according to the mentioned distinction criteria to the TDB, fundamental data unitization problems of the TDB are solved. In the following sections, unit types of the TDB will be discussed with their distinction criteria.

4.6.1 Types of Units in the TDB

Krippendorff (2004a, Chapter 5) defines three types of units for content annotation efforts: sampling, coding, and context units. In this section, the adaptation of the content annotation unit types to the TDB will be elaborated. In the elaborations, the coding and context units will be covered in more detail because the sampling units are not directly used in the inter-annotator agreement measurements. Therefore, this unit type is out of the scope of this thesis. On the other hand, the coding and context unit type decisions directly affect the result of inter-annotator agreement measurements.

4.6.1.1 Sampling Units

In the TDB, the discourse connective types can be named as sampling units because five discourse connective types, namely simple coordinating conjunctions, paired coordinating conjunctions, simplex subordinators, complex subordinators, and discourse adverbials are used to classify the entire discourse connectives. In the TDB, sampling units can be used to investigate annotations according to a particular syntactic or categorical property of the discourse connectives. In other words, sampling data unitization is not related to the inter-annotator agreement measurement process, therefore this kind of data unitization is out of scope of this thesis.

4.6.1.2 Coding Units

The coding units are the atomic units which are used to define the smallest element of the annotation efforts. This element should be defined by both using the distinction criteria, which are defined by Krippendorff (2004a, Chapter 5), and the theory behind the annotation effort.

In the TDB, the theory of annotation does not sanction any coding unit directly, because it is a low-level theory which does not limit the annotators by unit or structure definitions. Therefore, the criteria, which are used while determining the coding unit, should also not commit themselves to any high-level proposal. Therefore, the coding unit of the TDB should be the smallest piece of text. When all these are considered, there are two strong candidates; one of them is selecting characters as coding units as a result of a physical distinction criteria evaluation, and the other is selecting words as coding unit as a result of syntactical distinction.

As mentioned before, the TDB annotations are performed by using a computer program which is called as the DATT (Aktaş et al., 2010). In the DATT, the annotations are performed by selecting text spans character by character. Also, the annotated text spans are represented by character offsets relative to the starting character of the text. The offsets of starting and ending characters are held for each discourse element (Arg1, Arg2, and Conn) annotation. That is to say those, characters are physical distinctions that are automatically generated by the TDB annotation effort. When it is considered from the statistical perspective, a character is an atomic unit for texts that cannot be split up any into smaller pieces, which is an important property that is necessary to sustain the statistical representation capability of the unit.

However, one would assume that in the TDB annotations, annotators do not decide to include or exclude a character into a discourse element, but they decide if they should annotate a word or not. An annotation, which a part of word is included into the annotation and the other part is not, is mis-annotation except the subordinator connective annotations. A part or whole of the subordinator can be morphologically embedded into the last word of the Arg2, i.e., the verb. Yet, this is not a serious obstacle for taking words as coding units as this situation can be handled with a preprocess phase prior to agreement measurement. For the statistical perspective using words as coding units are also plausible since a word is an obvious syntactic element of a discourse structure which has enough granularities to represent annotations without over-generalization. On the other hand, taking sentences as coding units would be an over-generalization because when the TDB is investigated, it is seen that there are lots of agreed annotations which are not complete sentences (e.g., clauses and elliptic structures):

> Ex.: Elektrik enerjisi üretiminde kömür kullanımı **Yunanistan'da yüzde 71** iken, *Türkiye'de yüzde 29*.
>
> 'Use of coal in electricity generation is *29 percent in Turkey*, while **71 percent in Greece**.'

In brief, there are two candidates to take as coding units; characters (physical distinction), and words (physical and syntactic distinction) as there are several convincing proposals behind both of them. Therefore, in this thesis, agreement results will be presented for both coding unit candidates. A proposal for the TDB' coding unit will be made after the analysis of the measurements.

4.6.1.2 Context Units

In the TDB each discourse structure annotation corresponds to a context unit. Unlike content analysis, in discourse analysis, there are no predefined structures. The annotation process is a context unit determination process itself and inter-annotator agreement measurements are conducted to be sure that the context unitization is performed successfully. Hence, context unitization, and in general data unitization is vital for the reliability of annotation process.

As mentioned before, context units use coding units as building structures to indicate their limits. For the TDB, there are two possible usages of coding units to indicate context units:

1- Taking the starting and ending of the annotated text as context unit (**the boundaries**)

2- Taking all the coding units that span between the starting and ending of the annotations (**the interval**)

In the rest of this thesis, the first approach will be named as **the boundary approach**, and the second approach will be named as **the interval approach**.

Both the boundary and the interval approaches have advantages, depending on one's purpose. In the boundary approach, annotations that exactly match produce higher agreement results than the interval approach because, in the boundary approach only start and end offsets are considered while calculating the agreement. However, in the interval approach spans are used to calculate agreement, which makes the method more prone to produce positive agreement results. For not exactly matching but overlapping annotations, the boundary approach is more suitable. Therefore, one would select the appropriate context unitization approach according to one's research of interest.

As Krippendorff stated, context units can intersect with each other. Parallel to Krippendorff's statement, there are intersection cases of the discourse structures in the TDB, which are elaborated in § 4.3.

4.6.2 Realization of Data Unitization

In the previous sections (§ 4.6.1 and its sub-sections), three unit types for the TDB are elaborated according to Krippendorff's (2004a, Chapter 5) proposal for context annotation efforts: sampling, coding and context units. Except the first unitization type, the realizations of these definitions directly affect the inter-annotator agreement measurements on the TDB, because the coding unit type and context unit approach determines the digitized version of the annotations, and with different unitization preferences, the same annotations' digitized versions can be computationally very different. Therefore, it can be said that the data unitization preferences are the second important factor which affects the inter-annotator agreement measurement results.

In order to show the change of the digitized data according to the data unitization preferences, a sample annotation, which is also presented in the data representation section (§ 4.4), will be used in this section:

Akıntıya kapılıp umulmadık bir geceyi bölüştü benimle <u>ve</u> **bu kadarla kalsın istedi belki**.

'*She was drifted with a current and shared an unexpected night with me* <u>and</u> **perhaps she wanted to keep it this much only**.'

If an annotated discourse structure is named as a context unit, then the Arg1 and Arg2 would be named as mutually exclusive sub-context units that constitute the context unit with the discourse connective (Conn). (In the TDB, Arg1 and Arg2 are considered as annotation elements, however the Conn is not considered as an annotation element. Therefore, in the following illustrations only the digitized versions of the Arg1 and Arg2 annotations will be presented.)

In the following table relative character offsets of the discourse elements and the numeric values that represent the annotation of each discourse element is presented (see § 4.4 for the Xml version):

Table 7 Relative character offsets of a sample annotation's discourse elements

	Begin Offset	End Offset	Number that Represents Annotation
Conn	260	262	0 (none)
Arg1	206	259	1
Arg2	263	293	2

These four illustrations show all realizations for all coding preferences:

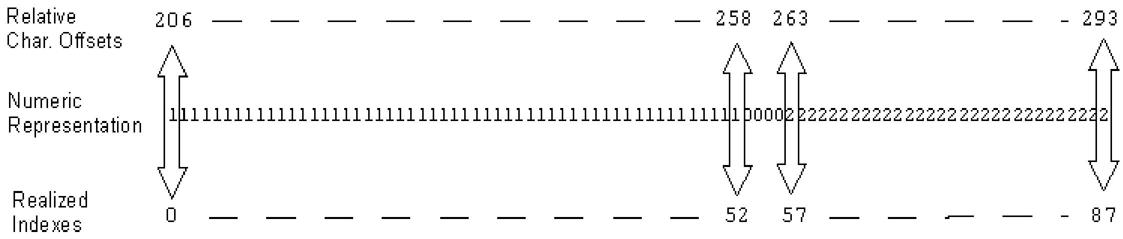


Figure 12 Taking characters as coding units, and using the interval approach to define context units

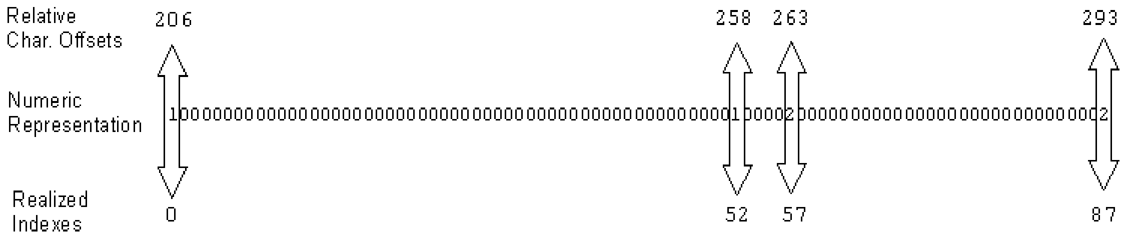


Figure 13 Taking character as coding unit, and using boundary approach to define context units

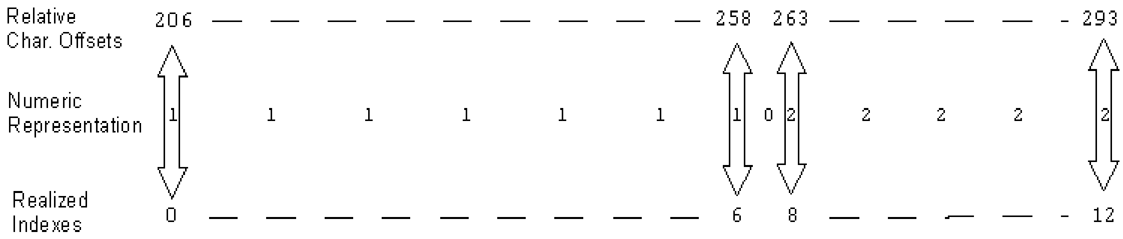


Figure 14 Taking words as coding units, and using the interval approach to define context units

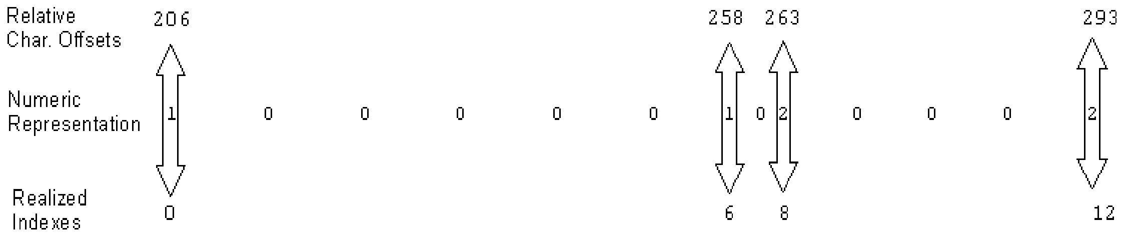


Figure 15 Taking words as coding units, and using the boundary approach to define context units

4.7 Agreement Measurement Methods

The agreement statistics that will be used in this thesis to measure the inter-annotator agreement among the annotators (namely Cochran's Q Test, Fleiss' Kappa, and Krippendorff's Alpha) have been comprehensively discussed in Chapter 2 and, in § 4.6, the ways of unitizing annotation data as inputs to the above mentioned agreement statistics have been presented. In this section, first, how to remove the computational problems, which occur in the annotation because of dependencies among discourse structures, will be illustrated with a sample annotation. Then, the usage of digitized annotations in the agreement statistics, which are mentioned in Chapter 2, will be explained.

Figure 16 is a schematic illustration of a sample annotation where two discourse structure annotations are performed by two annotators on the same text span. 1s represent coding units of Arg1, 2s represent coding units of Arg2, 3s represent *shared argument* (see § 4.3), and 0s represent not annotated text spans or the discourse connectives:

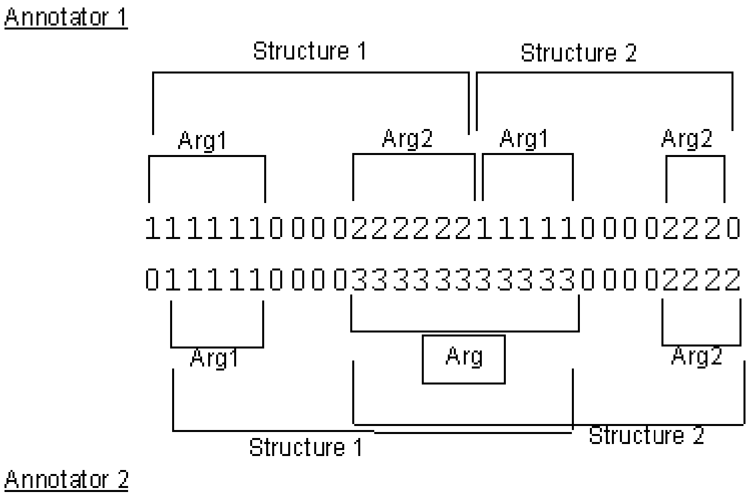


Figure 16 A schematic illustration of a sample annotation where two discourse structure annotations are performed by two annotators

This sample is a problem-ridden illustration which contains the following computation problems:

1. There is partial disagreement on Arg1 and Arg2 annotations among two annotators for both structures.
2. The second annotator's Arg2 annotation of the first structure and Arg1 annotation of the second structure is a *shared argument* case (see § 4.3).
3. There are two problems which are results of relative char offset approach:
    a. *Arg2 annotation in the first structure of the second annotator* encompasses both *Arg2 annotation in the first structure of the first annotator* and *Arg1 annotation in the second structure of the first annotator*.
    b. *Arg1 annotation in the second structure of the second annotator* encompasses both *Arg2 annotation in the first structure of the first annotator* and *Arg1 annotation in the second structure of the first annotator*.

The above mentioned partial disagreement case (1) is a pure agreement/disagreement determination problem which will be graded by the inter-annotator agreement measurement methods. However, case (2) and (3) are not such kind of problems. They are annotation approach-specific data representation problems. Therefore, these cases shall be handled after data unitization, and prior to the agreement calculation. First, the problem that occurs with case (2) shall be removed, and then the problem that occurs with case (3). Both removals shall be performed by replacing problematic annotation part with a computationally equivalent annotation.

In order to solve case (2), Arg1 and Arg2 annotations shall be handled separately. In fact, this is compulsory, because Arg1 and Arg2 annotations are semantically and syntactically separate annotations. As a result, separating Arg1 and Arg2 makes agreement measurements more valid and solves the problems that occur by the intersection or crossover of arguments of different discourse structures. The following figures (Figure 17 & Figure 18) are the illustrations of separated Arg1 and Arg2 annotations:

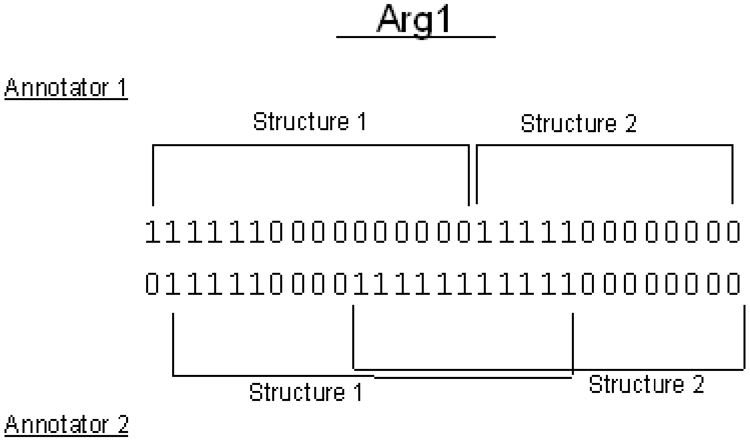


Figure 17 The illustration of separated Arg1 annotations

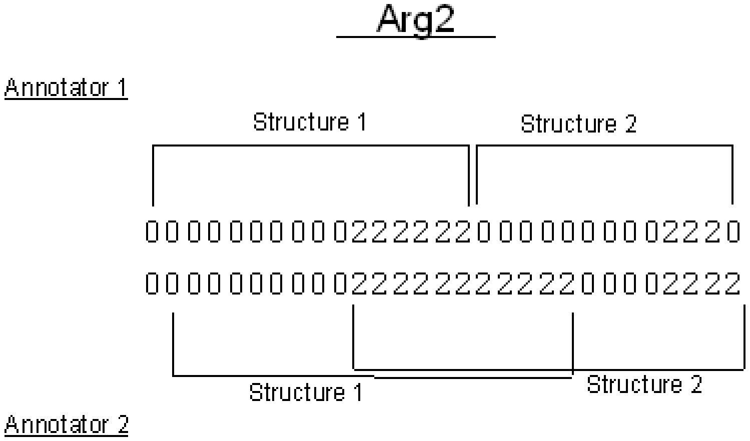


Figure 18 The illustration of separated Arg2 annotations

In order to solve case (3), both Arg1 and Arg2 annotations shall be extended till the structures do not intersect. Also, this approach handles discontinuous annotations. The following figures (Figure 19 & Figure 20) are the illustrations of extended versions of

Arg1 and Arg2 annotations (In order to make extensions distinguishable, * is used instead of 0):

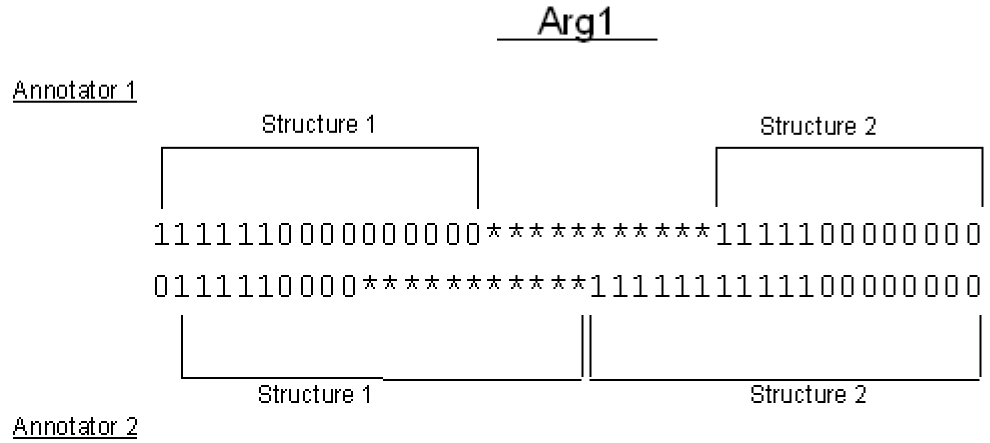


Figure 19 The illustrations of the extended versions of Arg1 annotations

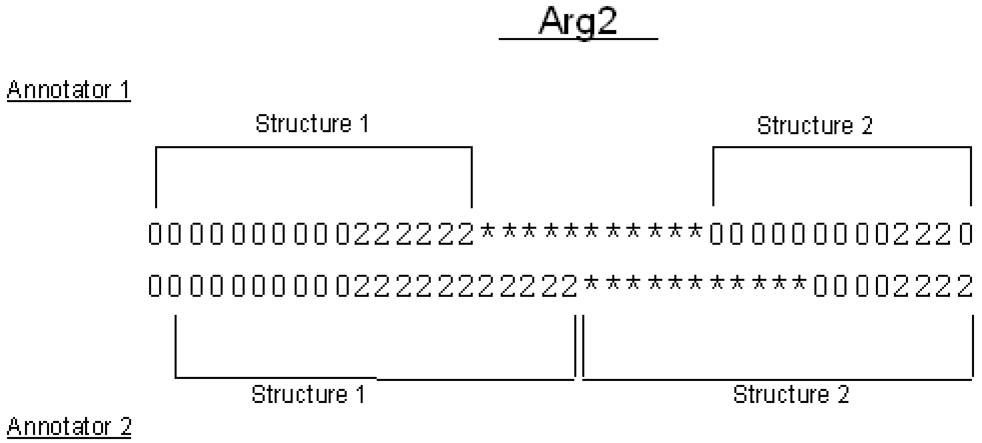


Figure 20 The illustrations of the extended versions of Arg2 annotations

Now the sample annotations are ready to be given as input to the inter-annotator agreement calculations. In the following sub-sections, the unitized and processed version of the sample annotation above will be placed into agreement tables of each method in order to clarify how the processed data are used in the agreement statistics.

4.7.1 COCHRAN’S Q TEST

Cochran’s Q test is a statistical test that can be used to evaluate the difference of the annotations of the two or more annotators. Here, the key point is that Cochran’s Q test operates on dichotomous data. Thanks to the proposed data unitization and handling methods, the annotations are converted to binary arrays where each element represents annotated or not annotated units. As a result, there are two individual tables for each Arg1 and Arg2 annotations.

In Tables 7 & 8, 1s represent annotated coding units and 0s represent not annotated coding units.

Table 8 The agreement table of Cochran’s Q test for Arg1

	Annotator 1	Annotator 2	Num. Of Samples
	1	1	10
	1	0	1
	0	1	6
	0	0	23
Totals	11	16	

Table 9 The agreement table of Cochran’s Q test for Arg2

	Annotator 1	Annotator 2	Num. Of Samples
	1	1	9
	1	0	0
	0	1	6
	0	0	25
Totals	6	10	

(How an agreement table is used to calculate agreement by using Cochran's Q test is described in § 2.1.)

4.7.2 Fleiss' Kappa

The representation of sample annotation in the agreement table of Fleiss' Kappa is as follows:

Table 10 The agreement table of Fleiss' Kappa for Arg1

	Category		
Subject	**0**	**1**	**agrh**
1	1	1	0.000
2		2	1.000
3		2	1.000
4		2	1.000
5		2	1.000
6		2	1.000
7	2		1.000
8	2		1.000
9	2		1.000
10	2		1.000
11	2		1.000
12	2		1.000
13	2		1.000
14	2		1.000
15	2		1.000
16	2		1.000
17	2		1.000
18	2		1.000
19	2		1.000
20	2		1.000
21	2		1.000
22	1	1	0.000
23	1	1	0.000
24	1	1	0.000
25	1	1	0.000
26	1	1	0.000
27	1	1	0.000
28		2	1.000
29		2	1.000
30		2	1.000
31		2	1.000
32		2	1.000
33	2		1.000
34	2		1.000
35	2		1.000
36	2		1.000
37	2		1.000
38	2		1.000
39	2		1.000
40	2		1.000
Total	53	27	
Pk	0.6625	0.3375	

Table 11 The agreement table of Fleiss' Kappa for Arg2

	Category		
Subject	**0**	**1**	**agrh**
1	2		1.000
2	2		1.000
3	2		1.000
4	2		1.000
5	2		1.000
6	2		1.000
7	2		1.000
8	2		1.000
9	2		1.000
10	2		1.000
11		2	1.000
12		2	1.000
13		2	1.000
14		2	1.000
15		2	1.000
16		2	1.000
17	1	1	0.000
18	1	1	0.000
19	1	1	0.000
20	1	1	0.000
21	1	1	0.000
22	2		1.000
23	2		1.000
24	2		1.000
25	2		1.000
26	2		1.000
27	2		1.000
28	2		1.000
29	2		1.000
30	2		1.000
31	2		1.000
32	2		1.000
33	2		1.000
34	2		1.000
35	2		1.000
36	2		1.000
37		2	1.000
38		2	1.000
39		2	1.000
40	1	1	0.000
Total	56	24	
Pk	0.7	0.3	

(How an agreement table is used to calculate agreement by using Fleiss' Kappa is described in § 2.2.2.)

4.7.3 Krippendorff's Alpha

Unlike Cochran's Q Test and Fleiss' Kappa, no agreement table is used in Krippendorff's Alpha. Yet, the unitized data shall be converted to a binary array where 1s represent annotated units and 0s represent not annotated units.

Arg1 arrays of the annotators for the sample annotation effort are as follows:

Annotator1$_{\text{Arg1}}$ = [1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0]

$\text{Annotator2}_{\text{Arg1}}$ = [0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0]

Arg2 arrays of the annotators for the sample annotation effort are as follows:
$\text{Annotator1}_{\text{Arg2}}$ = [0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0]
$\text{Annotator2}_{\text{Arg2}}$ = [0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1]

(How these presented arrays can be used to calculate agreement by using Krippendorff's Alpha is comprehensively described in § 2.2.3.)

4.8 AGREEMENT RESULTS

Currently, two academic works were presented that discuss inter-annotator agreement among the TDB (Zeyrek et al., 2009; Zeyrek et al., 2010). In this section, these works will be briefly mentioned along with the calculation parameters of the agreement statistics.

In the first work, Zeyrek et al. (2009) presented inter-annotator agreement for three subordinator annotations, namely rağmen 'despite', karşın 'although', and halde 'despite, along with'. The calculations were performed according to the following parameters:

Table 12 Agreement measurement parameters of Zeyrek et al., 2009

Agreement Statistic...:	Cochran's Q Test
Data Handling Method..:	Interval Approach
Coding Unit...........:	Character

In the second work, Zeyrek et al. (2010) presented agreement results for 22 discourse connectives. At the time of the work, there were 60 annotated connectives in the TDB, but in order to increase the validity of measurements they preferred to examine 22 connectives which have 10 or more annotations. The calculations were performed according to the following parameters:

Table 13 Agreement measurement parameters of Zeyrek et al., 2010

Agreement Statistic...:	Fleiss' Kappa
Data Handling Method..:	Boundary Approach
Coding Unit...........:	Character

The main concern of Zeyrek et al. (2010) was the sources of disagreements. Thus, they paid special interest to 8 connectives that have Kappa values less than 0.8.

Both of the above mentioned works are important efforts to mature the inter-annotator agreement measurement approach of the TDB. The questions and the problems that rose during these works contributed much to the agreement statistics, the data handling methods, and the data unitization processes that were presented in this thesis.

CHAPTER V

INTER-ANNOTATOR AGREEMENT MEASUREMENTS ON THE TDB

In this chapter, the inter-annotator agreement measurements on the TDB will be presented by using the set of agreement statistics and the data unitization approaches that are mentioned in Chapter 4. Whether these statistics and approaches are adequate for the TDB will be evaluated. Then, the measurement results will be used in some preliminary analysis. Finally, the features of the agreement measurement computer program developed during this thesis will be presented.

All these discussions will be conducted on 22 connective annotations of the TDB which encompass all types of Turkish discourse connectives (§ 4.1), and which were analyzed by Zeyrek et al. (2010). There are 9144 (3*2*1524) annotation tokens for these 22 connectives as all the annotations were performed by three annotators, and 1232 annotations for each arg1 and arg2 arguments.

5.1 AGREEMENT RESULTS ON THE TDB

As the agreement statistics, Cochran's Q Test (Q), Fleiss' Kappa (K), and Krippendorff's Alpha (A) will be used. For each statistic, the calculations will be performed by taking characters (C) and words (W) as coding units separately. Moreover, boundary (B) and interval (I) approaches will be used separately with each agreement statistic – coding unit pairs. Therefore, 12 agreement results will be discussed in the sub-sections.

The following table shows the list of the agreement result types with their abbreviations. These abbreviations will be used to define the agreement measurement statistics and the data unitization combinations:

Table 14 The TDB's agreement measurement statistics and data unitization combinations

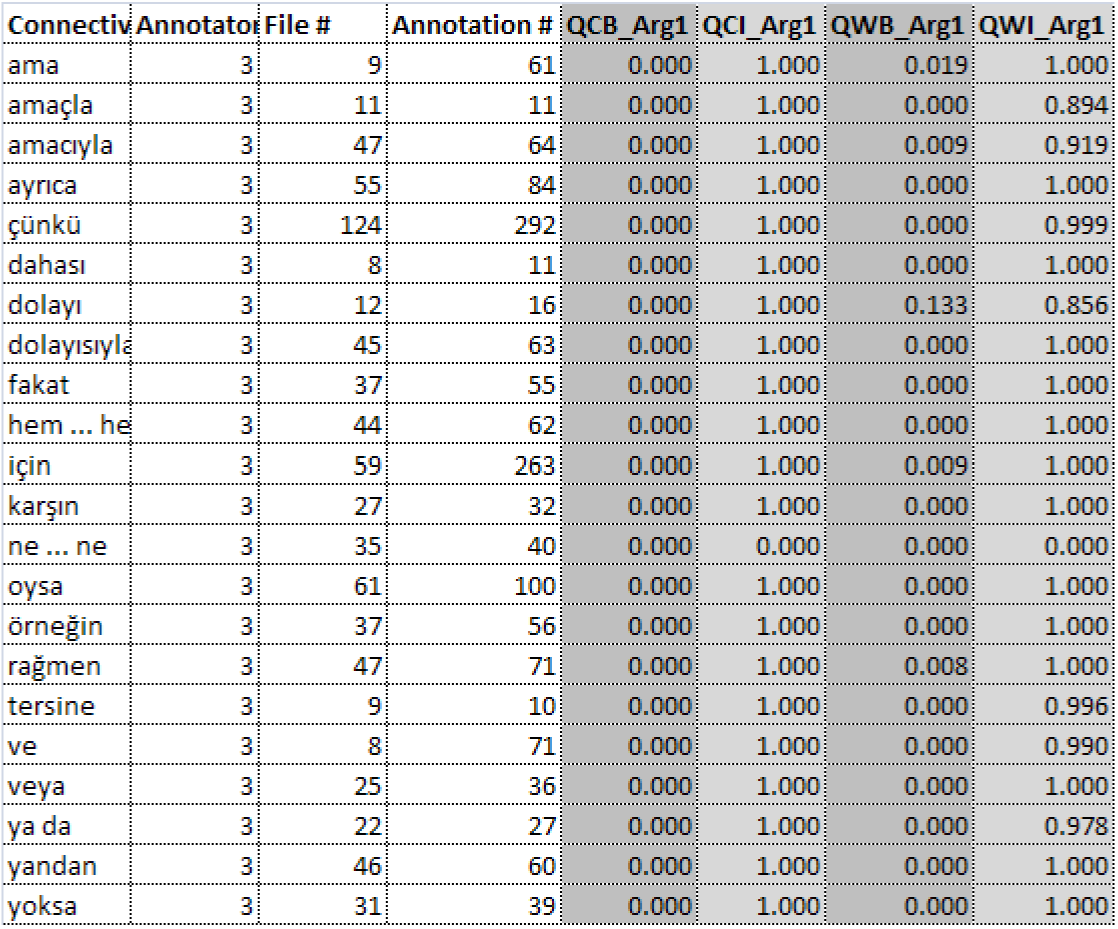
	Agreement Statistics	**Coding Unit Type**	**Context Unit Approach**	**Abbreviation**
1	Cochran's **Q** Test	**C**haracter	**B**oundary	**QCB**
2	Cochran's **Q** Test	**C**haracter	**I**nterval	**QCI**
3	Cochran's **Q** Test	**W**ord	**B**oundary	**QWB**
4	Cochran's **Q** Test	**W**ord	**I**nterval	**QWI**
5	Fleiss' **K**appa	**C**haracter	**B**oundary	**KCB**
6	Fleiss' **K**appa	**C**haracter	**I**nterval	**KCI**
7	Fleiss' **K**appa	**W**ord	**B**oundary	**KWB**
8	Fleiss' **K**appa	**W**ord	**I**nterval	**KWI**
9	Krippendorff's **A**lpha	**C**haracter	**B**oundary	**ACB**
10	Krippendorff's **A**lpha	**C**haracter	**I**nterval	**ACI**
11	Krippendorff's **A**lpha	**W**ord	**B**oundary	**AWB**
12	Krippendorff's **A**lpha	**W**ord	**I**nterval	**AWI**

5.1.1 Cochran's Q Test

In this section, QCB, QCI, QWB, and QWI results will be presented for Arg1 and Arg2 separately. Also, the reasons of the unsuitability of Cochran's Q test for the TDB will be discussed.

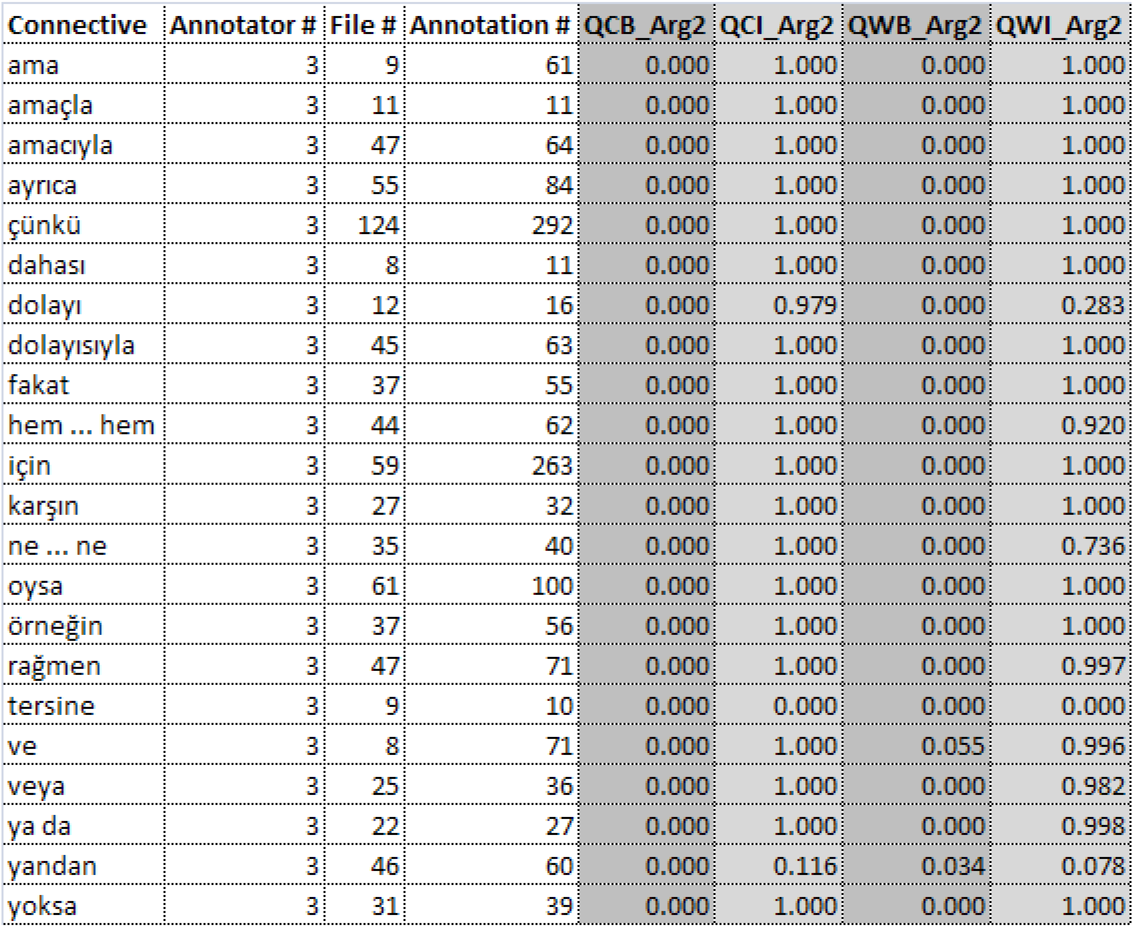
In the following tables, presented agreement results are p-values which indicate acceptable agreement among annotators when their value are greater than 0.05, otherwise they indicate disagreement.

Table 15 Cochran's Q test p-value results of the TDB for Arg1 for 22 connectives



Connectiv	Annotator	File #	Annotation #	QCB_Arg1	QCI_Arg1	QWB_Arg1	QWI_Arg1
ama	3	9	61	0.000	1.000	0.019	1.000
amaçla	3	11	11	0.000	1.000	0.000	0.894
amacıyla	3	47	64	0.000	1.000	0.009	0.919
ayrıca	3	55	84	0.000	1.000	0.000	1.000
çünkü	3	124	292	0.000	1.000	0.000	0.999
dahası	3	8	11	0.000	1.000	0.000	1.000
dolayı	3	12	16	0.000	1.000	0.133	0.856
dolayısıyl	3	45	63	0.000	1.000	0.000	1.000
fakat	3	37	55	0.000	1.000	0.000	1.000
hem ... he	3	44	62	0.000	1.000	0.000	1.000
için	3	59	263	0.000	1.000	0.009	1.000
karşın	3	27	32	0.000	1.000	0.000	1.000
ne ... ne	3	35	40	0.000	0.000	0.000	0.000
oysa	3	61	100	0.000	1.000	0.000	1.000
örneğin	3	37	56	0.000	1.000	0.000	1.000
rağmen	3	47	71	0.000	1.000	0.008	1.000
tersine	3	9	10	0.000	1.000	0.000	0.996
ve	3	8	71	0.000	1.000	0.000	0.990
veya	3	25	36	0.000	1.000	0.000	1.000
ya da	3	22	27	0.000	1.000	0.000	0.978
yandan	3	46	60	0.000	1.000	0.000	1.000
yoksa	3	31	39	0.000	1.000	0.000	1.000

Table 16 Cochran's Q test p-value results of the TDB for Arg2 for 22 connectives



Connective	Annotator #	File #	Annotation #	QCB_Arg2	QCI_Arg2	QWB_Arg2	QWI_Arg2
ama	3	9	61	0.000	1.000	0.000	1.000
amaçla	3	11	11	0.000	1.000	0.000	1.000
amacıyla	3	47	64	0.000	1.000	0.000	1.000
ayrıca	3	55	84	0.000	1.000	0.000	1.000
çünkü	3	124	292	0.000	1.000	0.000	1.000
dahası	3	8	11	0.000	1.000	0.000	1.000
dolayı	3	12	16	0.000	0.979	0.000	0.283
dolayısıyla	3	45	63	0.000	1.000	0.000	1.000
fakat	3	37	55	0.000	1.000	0.000	1.000
hem ... hem	3	44	62	0.000	1.000	0.000	0.920
için	3	59	263	0.000	1.000	0.000	1.000
karşın	3	27	32	0.000	1.000	0.000	1.000
ne ... ne	3	35	40	0.000	1.000	0.000	0.736
oysa	3	61	100	0.000	1.000	0.000	1.000
örneğin	3	37	56	0.000	1.000	0.000	1.000
rağmen	3	47	71	0.000	1.000	0.000	0.997
tersine	3	9	10	0.000	0.000	0.000	0.000
ve	3	8	71	0.000	1.000	0.055	0.996
veya	3	25	36	0.000	1.000	0.000	0.982
ya da	3	22	27	0.000	1.000	0.000	0.998
yandan	3	46	60	0.000	0.116	0.034	0.078
yoksa	3	31	39	0.000	1.000	0.000	1.000

When both tables are examined, the most explicit inference is that the results, which are calculated by the boundary approach, are always 0 or very close to 0. As indicated before, p-value that is below 0.05 means disagreement. However, when the results that are calculated via other agreement statistics (see § 5.1.2 and 5.1.3), and even when the annotation data files themselves are analyzed manually, it is seen that there is no definite disagreement for all of the connective annotations. The reason of such disagreement results is easily understood when the formula of Cochran's Q Test is re-considered. In Cochran's Q test, when all of the annotators perform the same number of annotations, in other words, when the summation of 1s in each column of the agreement table of Cochran's Q test is equal, the dividend of the Q formula always becomes 0. Therefore, p-value becomes always 0 regardless of the inter-annotator agreement. In the TDB, in principle, all annotators perform the same number of argument annotations, and thus the same numbers of boundary data are produced. Therefore, all Cochran's Q test results with the boundary approach are always supposed to be 0. However, it is seen that some boundary values are very close to 0 but not 0. In some annotations, the beginning and end of annotations are the same coding units (word or character). For this reason, two boundaries are represented by one value in the agreement table of Cochran's Q test. Thus, the column totals become different when one or two annotators performed an annotation that begins and ends in the same coding unit, and the rest of the annotators performed an annotation that begins and ends in a different coding unit for the same discourse connective. With different column totals, the p values are calculated as bigger than 0 but still very close to 0.

As another striking feature of the tables, the results which are calculated according to the interval approach are mostly 1 or very close to 1, with a few exceptions. In this approach, the results are not always 0 because, as a result of several differences among the annotators, the column totals in Cochran's Q table become different from each other. However, the results of the interval approach are not satisfactory either. The values are always 1 or very close to 1, which does not tell more than that all connective annotations of the TDB are not random annotations. However, with such a method, one cannot observe the progress or recession in the annotations by time, by training, by annotator, and so on. Also, one cannot compare the agreement results of the different connectives.

As a result, Cochran's Q test produces absolute agreement or disagreement results for all observed settings. When the aim is to measure the amount of the inter-annotator agreement, Cochran's Q test cannot be used as the agreement measurement method.

5.1.2 FLEISS' KAPPA

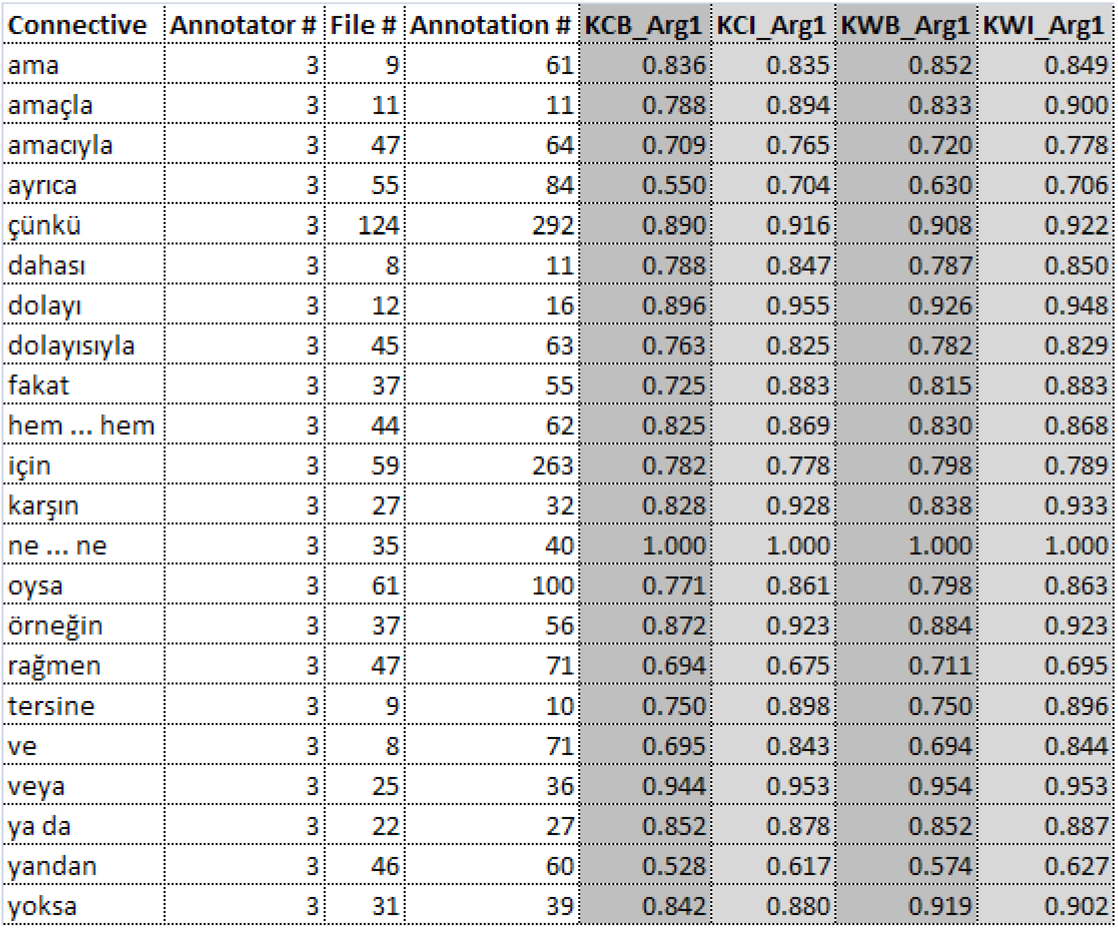
In this section, KCB, KCI, KWB, and KWI results will be presented for Arg1 and Arg2 separately. The result differences that occur with the data unitization preferences (interval-boundary and word-character) will also be discussed.

In the literature, the results of kappa-like agreement measurements are mostly interpreted in six categories (Landis, & Koch, 1971; Artstein, & Poesio, 2008):

1- Measurement > 0.8: Perfect agreement
2- 0.8 > Measurement > 0.6: Substantial agreement
3- 0.6 > Measurement > 0.4: Moderate agreement
4- 0.4 > Measurement > 0.2: Fair agreement
5- 0.2 > Measurement > 0.0: Slight agreement
6- 0.0 > Measurement: Poor agreement

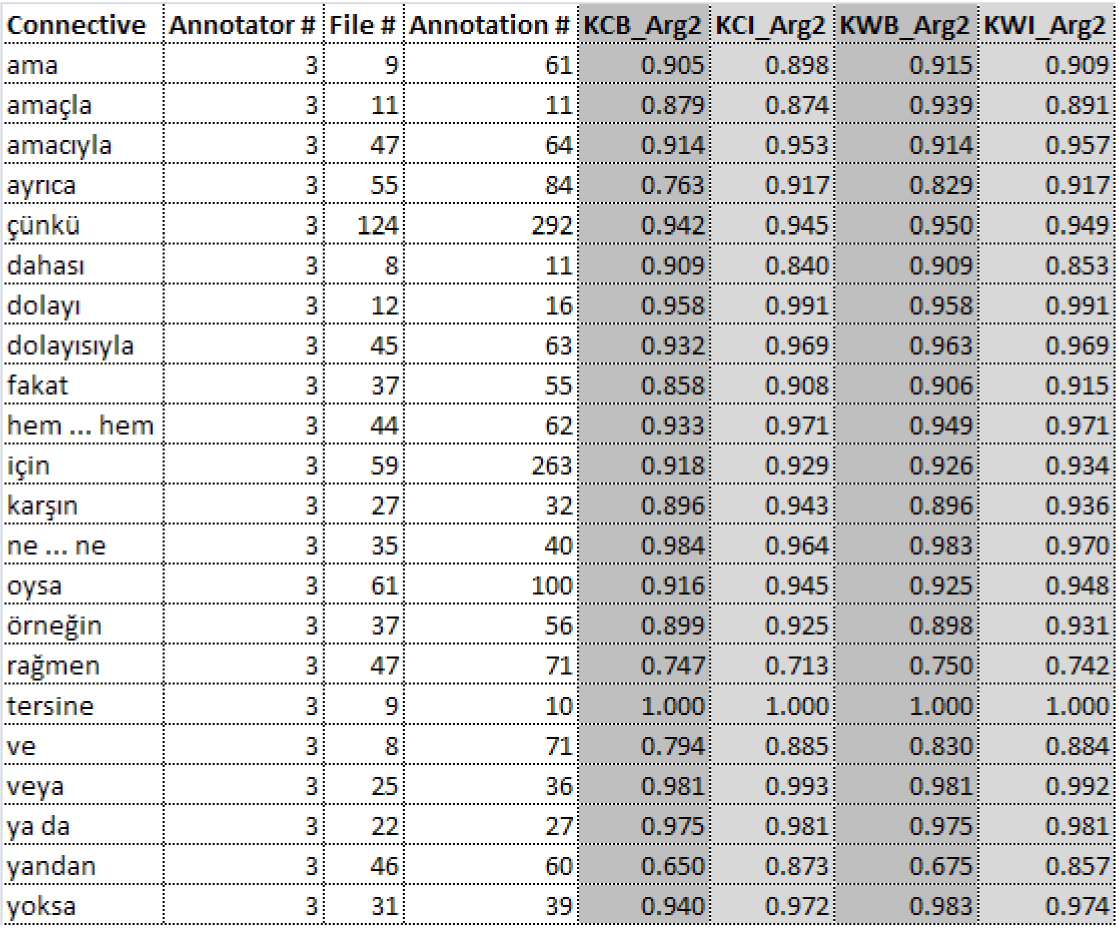
Table 17 and 18 present the results of Fleiss' Kappa for each connective, considering coding units (character - word) and interval approaches (interval - boundary):

Table 17 Fleiss' Kappa results of the TDB for Arg1 for 22 connectives



Connective	Annotator #	File #	Annotation #	KCB_Arg1	KCI_Arg1	KWB_Arg1	KWI_Arg1
ama	3	9	61	0.836	0.835	0.852	0.849
amaçla	3	11	11	0.788	0.894	0.833	0.900
amacıyla	3	47	64	0.709	0.765	0.720	0.778
ayrıca	3	55	84	0.550	0.704	0.630	0.706
çünkü	3	124	292	0.890	0.916	0.908	0.922
dahası	3	8	11	0.788	0.847	0.787	0.850
dolayı	3	12	16	0.896	0.955	0.926	0.948
dolayısıyla	3	45	63	0.763	0.825	0.782	0.829
fakat	3	37	55	0.725	0.883	0.815	0.883
hem ... hem	3	44	62	0.825	0.869	0.830	0.868
için	3	59	263	0.782	0.778	0.798	0.789
karşın	3	27	32	0.828	0.928	0.838	0.933
ne ... ne	3	35	40	1.000	1.000	1.000	1.000
oysa	3	61	100	0.771	0.861	0.798	0.863
örneğin	3	37	56	0.872	0.923	0.884	0.923
rağmen	3	47	71	0.694	0.675	0.711	0.695
tersine	3	9	10	0.750	0.898	0.750	0.896
ve	3	8	71	0.695	0.843	0.694	0.844
veya	3	25	36	0.944	0.953	0.954	0.953
ya da	3	22	27	0.852	0.878	0.852	0.887
yandan	3	46	60	0.528	0.617	0.574	0.627
yoksa	3	31	39	0.842	0.880	0.919	0.902

Table 18 Fleiss' Kappa results of the TDB for Arg2 for 22 connectives



Connective	Annotator #	File #	Annotation #	KCB_Arg2	KCI_Arg2	KWB_Arg2	KWI_Arg2
ama	3	9	61	0.905	0.898	0.915	0.909
amaçla	3	11	11	0.879	0.874	0.939	0.891
amacıyla	3	47	64	0.914	0.953	0.914	0.957
ayrıca	3	55	84	0.763	0.917	0.829	0.917
çünkü	3	124	292	0.942	0.945	0.950	0.949
dahası	3	8	11	0.909	0.840	0.909	0.853
dolayı	3	12	16	0.958	0.991	0.958	0.991
dolayısıyla	3	45	63	0.932	0.969	0.963	0.969
fakat	3	37	55	0.858	0.908	0.906	0.915
hem ... hem	3	44	62	0.933	0.971	0.949	0.971
için	3	59	263	0.918	0.929	0.926	0.934
karşın	3	27	32	0.896	0.943	0.896	0.936
ne ... ne	3	35	40	0.984	0.964	0.983	0.970
oysa	3	61	100	0.916	0.945	0.925	0.948
örneğin	3	37	56	0.899	0.925	0.898	0.931
rağmen	3	47	71	0.747	0.713	0.750	0.742
tersine	3	9	10	1.000	1.000	1.000	1.000
ve	3	8	71	0.794	0.885	0.830	0.884
veya	3	25	36	0.981	0.993	0.981	0.992
ya da	3	22	27	0.975	0.981	0.975	0.981
yandan	3	46	60	0.650	0.873	0.675	0.857
yoksa	3	31	39	0.940	0.972	0.983	0.974

Unlike Cochran's Q test results, there is no systematic problem among Fleiss' Kappa results. The changes among the results are because of data unitization preferences. The following two *box-and-whisker diagrams* visualize the Arg1 and Arg2 measurements and show the data unitization effects. In the four columns (KCB_ArgX, KCI_ArgX, KWB_ArgX, and KWI_ArgX) of the diagrams, the *five-number summaries* and *whiskers* are depicted according to Tukey (1977).

The five-number summaries include the following information:

- First Quartile (Q1): The median of lowest 25% of data

- Second Quartile (Q2): The median
- Third Quartile (Q3): The median of highest 25 of data
- Max Outlier: The largest measurement that falls outside of the *boxes* and *whiskers*
- Min Outlier: The smallest measurement that falls outside of the *boxes* and *whiskers*

The *whiskers* are used to depict the tails of distribution. There are different ways of determining the whiskers, but in this thesis the most common way, which operates through Q3 and Q1 is used:

- The largest end of *whiskers*: Q3+1.5*(Q3-Q1)
- The smallest end of *whiskers*: Q1-1.5*(Q3-Q1)

(In the following diagrams, the results of the annotations that have the exact agreement (1.0) values are not depicted in order to make more explicit the behaviors of each measurement.)

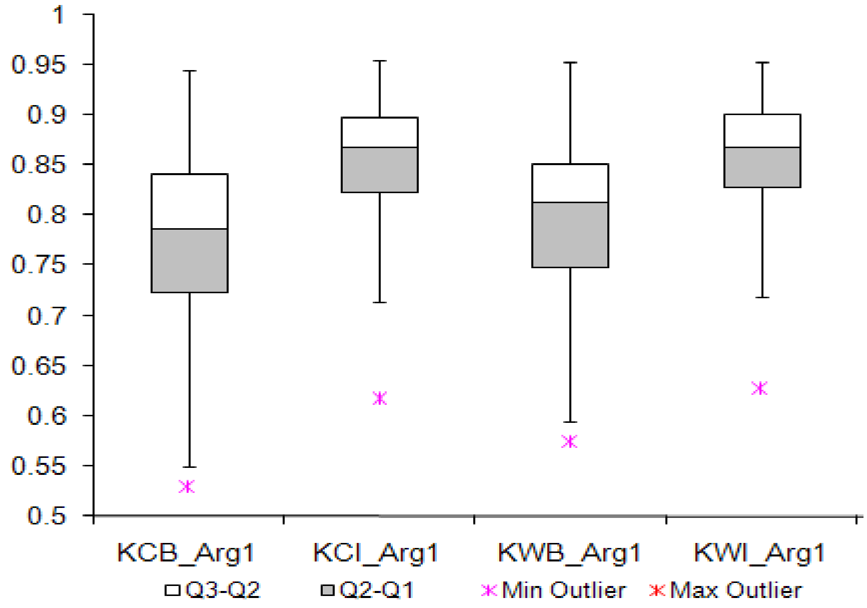


Figure 21 **The box-and-whisker representation of Fleiss' Kappa results of the TDB for Arg1**

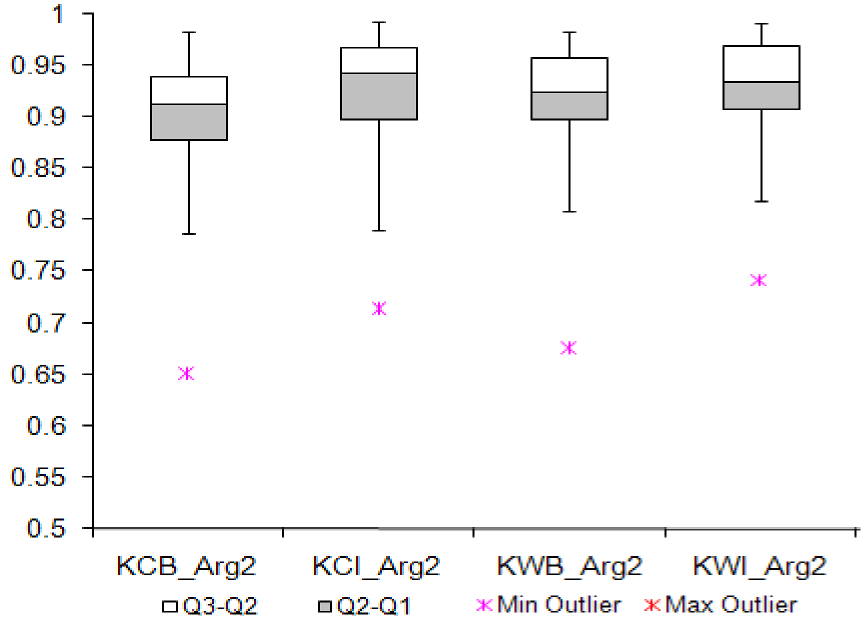


Figure 22 **The box-and-whisker representation of Fleiss' Kappa results of the TDB for Arg2**

As seen from the above diagrams, Arg2 results are higher and closer to each other than Arg1 results, therefore using Arg1 while making inferences is more correct. Figure 21 shows that the KCB and KWB results are closer than the KCI and KWI results for both *boxes* and *whiskers*. Also, it is seen that the interval approach is more prone to produce higher agreement results than the boundary approach. There are two reasons behind these result variations:

(1) The results do not change much by taking characters or words as coding units because the numbers of both units are satisfactory to perform statistically reliable agreement measurements.

(2) The interval approach is more prone to produce higher results because most of the disagreement cases contain partial-overlap among the annotations. (Because of its data producing approach, the boundary approach has lesser capacity to identify the differences between partial-overlap and total disagreement cases.)

Additionally, the presented Arg2 measurements are mostly (for 92% of measurements) greater than Arg1 measurements, which is because Arg2s are syntactically anchoring the discourse connective (Zeyrek et al. 2009).

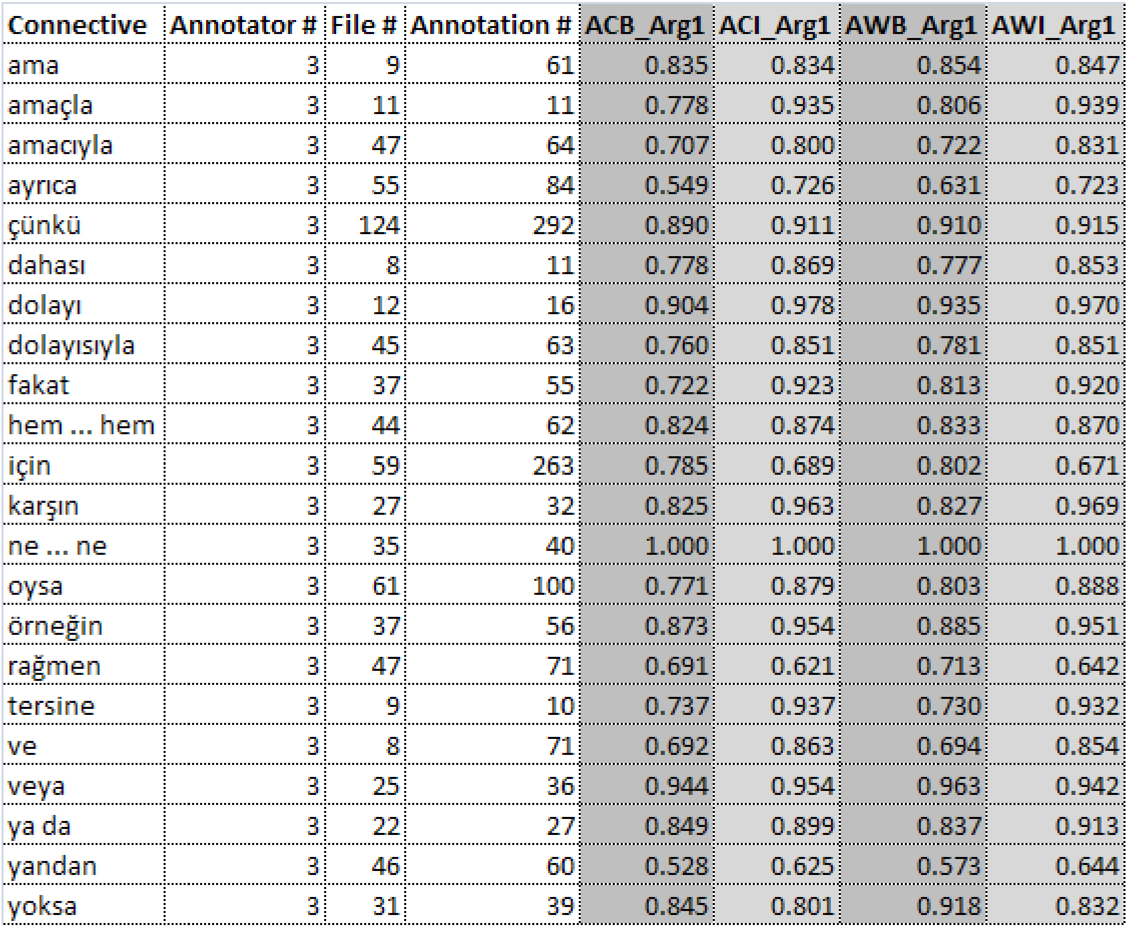
Finally, there is no significant systematic problem and the result variations can be explained reasonably. Fleiss' Kappa can be used to measure the reliability of the TDB with correct research considerations. The evaluation of the methods according to research considerations will be performed in § 5.2.

5.1.3 KRIPPENDORFF'S ALPHA

In this section, ACB, ACI, AWB, and AWI results will be presented for Arg1 and Arg2 separately. The result differences that occur with the data unitization preferences (interval-boundary and word-character) will be discussed.

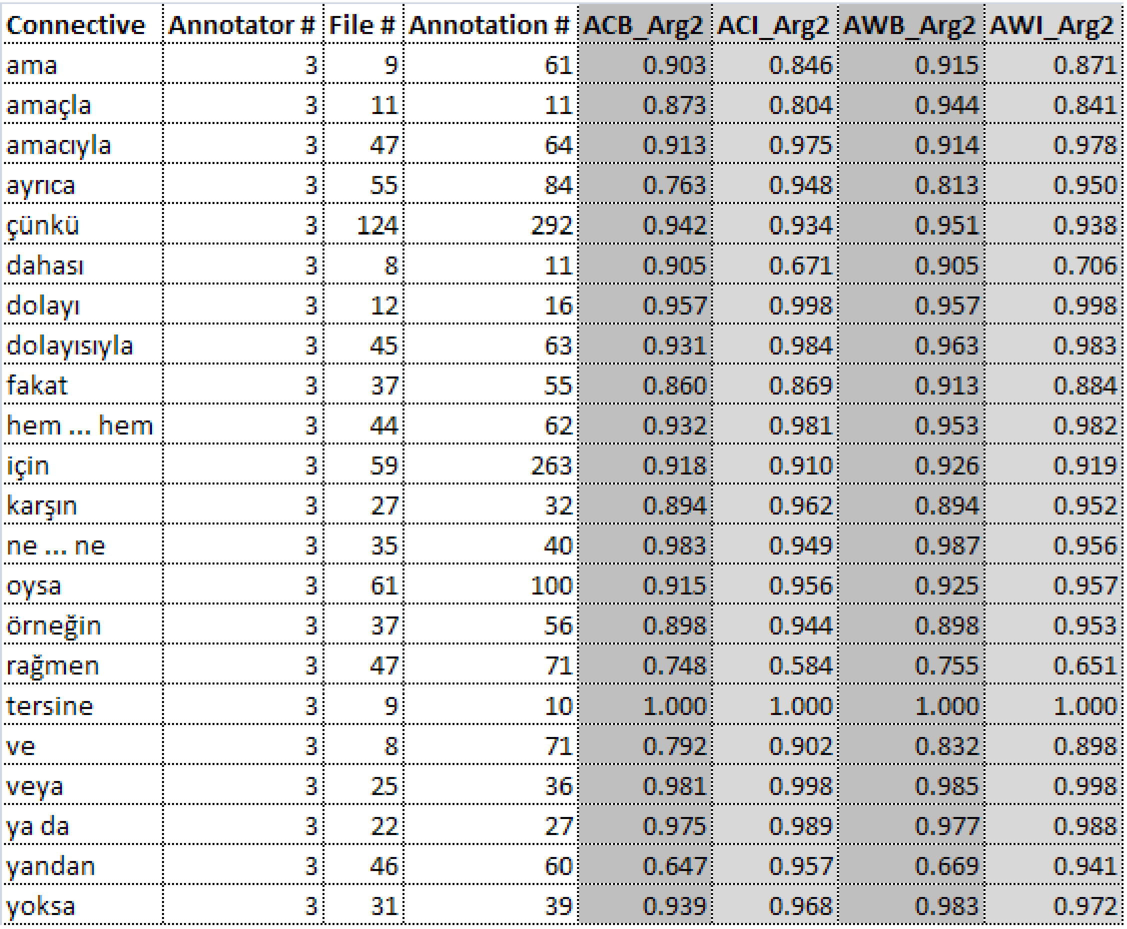
Table 19 and 20 present the results of Krippendorff's Alpha for each connective, considering coding units (character - word) and interval approaches (interval - boundary):

Table 19 Krippendorff's Alpha results of the TDB for Arg1 for 22 connectives



Connective	Annotator #	File #	Annotation #	ACB_Arg1	ACI_Arg1	AWB_Arg1	AWI_Arg1
ama	3	9	61	0.835	0.834	0.854	0.847
amaçla	3	11	11	0.778	0.935	0.806	0.939
amacıyla	3	47	64	0.707	0.800	0.722	0.831
ayrıca	3	55	84	0.549	0.726	0.631	0.723
çünkü	3	124	292	0.890	0.911	0.910	0.915
dahası	3	8	11	0.778	0.869	0.777	0.853
dolayı	3	12	16	0.904	0.978	0.935	0.970
dolayısıyla	3	45	63	0.760	0.851	0.781	0.851
fakat	3	37	55	0.722	0.923	0.813	0.920
hem ... hem	3	44	62	0.824	0.874	0.833	0.870
için	3	59	263	0.785	0.689	0.802	0.671
karşın	3	27	32	0.825	0.963	0.827	0.969
ne ... ne	3	35	40	1.000	1.000	1.000	1.000
oysa	3	61	100	0.771	0.879	0.803	0.888
örneğin	3	37	56	0.873	0.954	0.885	0.951
rağmen	3	47	71	0.691	0.621	0.713	0.642
tersine	3	9	10	0.737	0.937	0.730	0.932
ve	3	8	71	0.692	0.863	0.694	0.854
veya	3	25	36	0.944	0.954	0.963	0.942
ya da	3	22	27	0.849	0.899	0.837	0.913
yandan	3	46	60	0.528	0.625	0.573	0.644
yoksa	3	31	39	0.845	0.801	0.918	0.832

Table 20 Krippendorff's Alpha results of the TDB for Arg2 for 22 connectives



Connective	Annotator #	File #	Annotation #	ACB_Arg2	ACI_Arg2	AWB_Arg2	AWI_Arg2
ama	3	9	61	0.903	0.846	0.915	0.871
amaçla	3	11	11	0.873	0.804	0.944	0.841
amacıyla	3	47	64	0.913	0.975	0.914	0.978
ayrıca	3	55	84	0.763	0.948	0.813	0.950
çünkü	3	124	292	0.942	0.934	0.951	0.938
dahası	3	8	11	0.905	0.671	0.905	0.706
dolayı	3	12	16	0.957	0.998	0.957	0.998
dolayısıyla	3	45	63	0.931	0.984	0.963	0.983
fakat	3	37	55	0.860	0.869	0.913	0.884
hem ... hem	3	44	62	0.932	0.981	0.953	0.982
için	3	59	263	0.918	0.910	0.926	0.919
karşın	3	27	32	0.894	0.962	0.894	0.952
ne ... ne	3	35	40	0.983	0.949	0.987	0.956
oysa	3	61	100	0.915	0.956	0.925	0.957
örneğin	3	37	56	0.898	0.944	0.898	0.953
rağmen	3	47	71	0.748	0.584	0.755	0.651
tersine	3	9	10	1.000	1.000	1.000	1.000
ve	3	8	71	0.792	0.902	0.832	0.898
veya	3	25	36	0.981	0.998	0.985	0.998
ya da	3	22	27	0.975	0.989	0.977	0.988
yandan	3	46	60	0.647	0.957	0.669	0.941
yoksa	3	31	39	0.939	0.968	0.983	0.972

Like Fleiss' Kappa results, there is no systematic problem among the results. The changes in the results can be explained by the data unitization preferences. The following two *box-and-whisker diagrams* visualize the Arg1 and Arg2 measurements and show the data unitization effects. In the four columns (KCB_ArgX, KCI_ArgX, KWB_ArgX, and KWI_ArgX) of the diagrams, the *five-number summaries* and *whiskers* are depicted according to Tukey (1977).

(In the following diagrams, the results of the annotations that have the exact agreement (1.0) values are not depicted in order to make more explicit the behaviors of each measurement.)

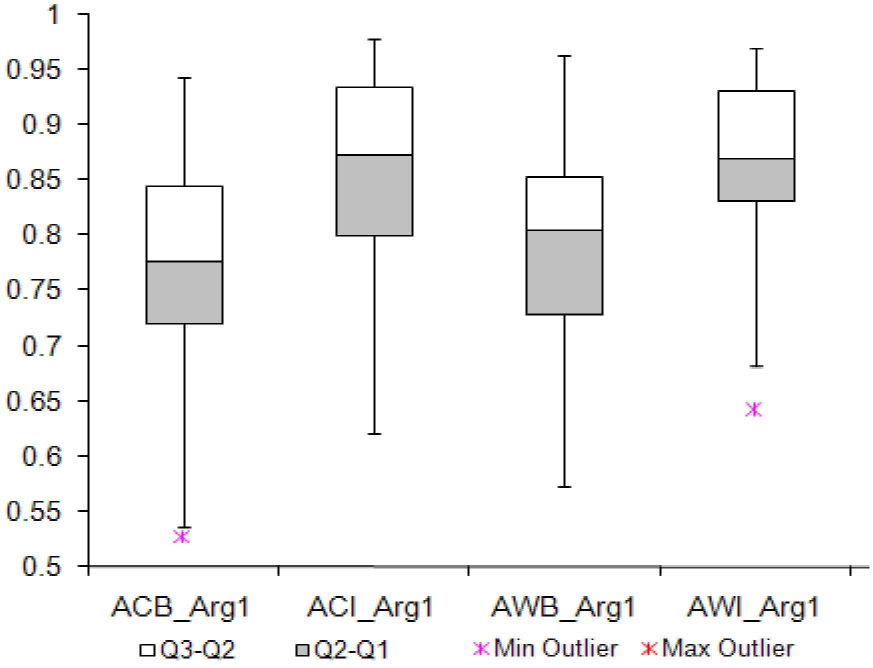


Figure 23 The box-and-whisker representation of Krippendorff's Alpha results of the TDB for Arg1

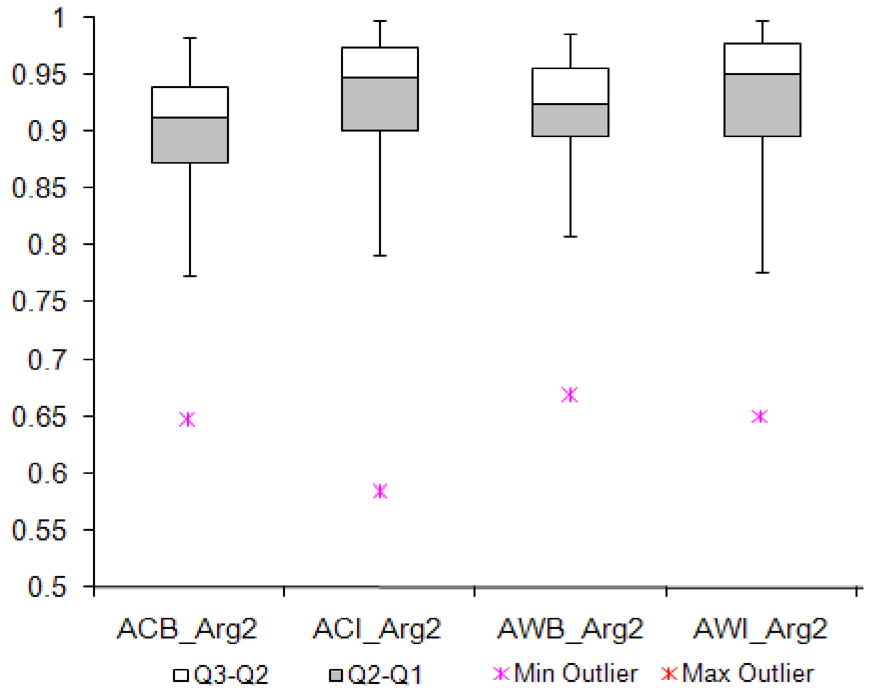


Figure 24 The box-and-whisker representation of Krippendorff's Alpha results of the TDB for Arg2

As seen from the above diagrams, Arg2 results are higher and closer than Arg1 results, therefore using Arg1 while making inferences is more correct. Figure 23 shows that the ACB and AWB results are closer than the ACI and AWI results for both *boxes* and *whiskers*. Also, it is seen that the interval approach is more prone to produce higher agreement results than the boundary approach:

(1) Like Fleiss' Kappa, there is no significant difference between taking characters or words as coding units.

(2) Like Fleiss' Kappa, the interval approach is more prone to produce higher results because most of the disagreement cases contain partial-overlap which the interval approach considers but the boundary approach does not.

Again, like Fleiss' Kappa results, about 83% of Arg2 measurements are greater than Arg1. Arg2 is syntactically anchored to the discourse connective (Zeyrek et al., 2009), which is also supported by Krippendorff's Alpha results.

Finally, Krippendorff's Alpha and Fleiss' Kappa results have similar consistency with different results. By correct research aspects, Krippendorff's Alpha is a promising inter-annotator agreement measurement method for the TDB. After the evaluation of the methods in § 5.2, the role of Krippendorff's Alpha will become clearer.

5.2 EVALUATION OF METHODS

In the previous section (§ 5.1), the inter-annotator agreement measurements, which are calculated by Cochran's Q test, Fleiss' Kappa, and Krippendorff's Alpha, were presented, and some interim conclusions were drawn. In this section, the methods will be evaluated by comparing them to each other. As a result of the evaluations, the appropriate inter-annotator agreement measurement method and the data unitization approaches for the TDB will be determined.

As mentioned in § 5.1.1, Cochran's Q test always produces values around 0 with the boundary approach and always produces values around 1 with the interval approach. The method cannot perform correctly with the boundary approach because of formulation problems. On the other hand, the method works correctly with the interval approach but the results are still not satisfactory for this thesis, which aims to measure the amount of agreement/disagreement among annotators. As a consequence, Cochran's Q test is appropriate for determining whether annotations are random or not random, but it cannot be used to measure the degree of agreement on the TDB annotations. Therefore, Cochran's Q test is not a suitable method for the TDB.

Now that there are two candidate methods left, when elaborating on these methods, the effects of the data unitization preferences shall also be discussed since the results of both Fleiss' Kappa and Krippendorff's Alpha change similarly according to the data unitization preferences. First of all, recall that there are two kinds of data unitization issues: (1) determination of the coding unit type (character or word), and (2) determination of the context unit approach (interval or boundary approach), and these shall be examined separately:

(1) The Coding Unit: § 5.1.2 and § 5.1.3 showed that there is no significant difference between selecting characters or words as coding units. On the other hand, taking words as the coding unit is more plausible because annotators annotate discourse relations by considering words (smallest meaningful syntactic units), not characters (smallest physical units). In other words, when an annotation does not end at the boundaries of a word, it has two meanings either the discourse connective is morphologically adjusted to the word or the annotator has made a mistake. In both cases, extending the annotations to word boundaries does not cause a problem. For the first case, there is no problem because annotations of all annotators are extended, thus the agreement measurement is not affected. For the second case, such an extension will remove the erroneous annotation by considering that the annotator aimed to annotate the whole word. As a result, using words as coding units is more advantageous than using characters as coding units. Therefore, in the rest of the thesis, the result that took words as coding units will be taken into consideration.

(2) The Context Unit Approach: In § 4.6.1.2, it is mentioned that discourse relation annotation itself is a context unit determination process. On the other hand, there are two ways to determine the context units: interval and boundary approaches. Table 21, 22, 23 and 24 show that in both Fleiss' Kappa and Krippendorff's Alpha results, the differences between the boundary and interval approaches are

considerably large. The followings are the examination of these two approaches to identify the appropriate research aspects for each:

a. The Boundary Approach: This approach only considers if the beginnings and ends of the argument annotations agree or not. Therefore, one would use this method when one wants to measure agreement by considering the partial-overlap disagreements as total disagreements. Krippendorff's Alpha seems as a weak candidate for that kind of aim. However, Krippendorff's Alpha gets closer to Fleiss' Kappa when unit lengths get shorter (Krippendorff, 2004c Chapter5). Therefore, these two agreement statistics can be used interchangeably with the boundary approach.

b. The Interval Approach: Contrary to the boundary approach, the interval approach considers annotations as a whole. Therefore, one would use this method when one wants to measure agreement by considering partial-overlap disagreements separately from total disagreements. The best way to perform such an analysis is rating partial-overlaps according to the overlapping amount of text spans. As Krippendorff's Alpha has such a weighing approach, and Fleiss' Kappa does not, the appropriate method is Krippendorff's Alpha. The difference between Krippendorff's Alpha and Fleiss' Kappa can be monitored via the following *box-and-whisker diagram* (Figure 25). The diagram shows that, when the other settings except the agreement statistic are kept same, Krippendorff's Alpha results are more prone to oscillate than Fleiss' Kappa results. This shows that Krippendorff's Alpha is more selective on the partial-overlap disagreements:

(In the following diagram, the results of the annotations that have the exact agreement (1.0) values are not depicted in order to make more explicit the behaviors of each measurement.)

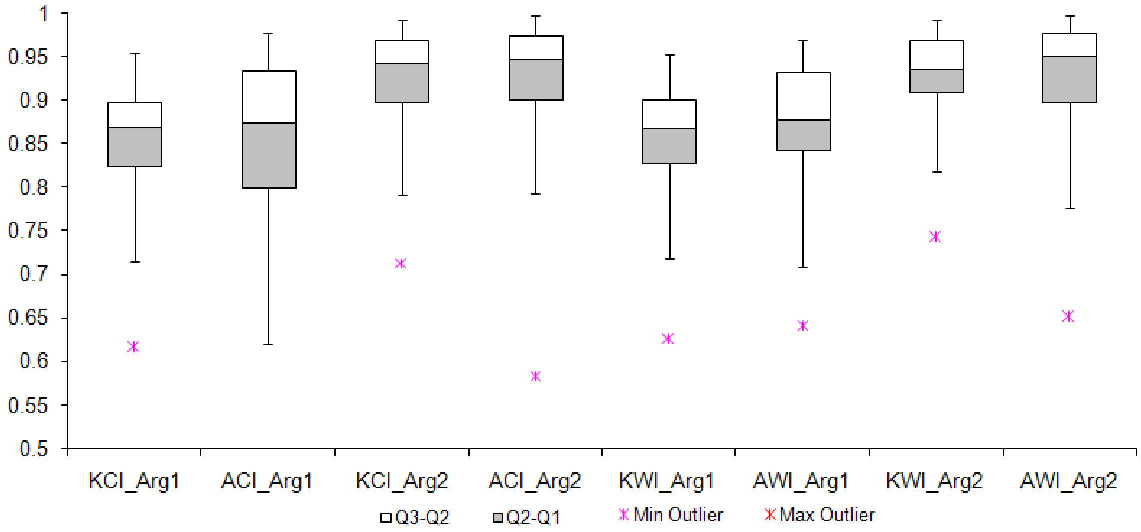


Figure 25 The box-and-whisker comparison of Interval Approach Results of Fleiss' Kappa and Krippendorff's Alpha

According to the mathematical summary above, the difference between the two methods is obvious. Also, with the support of theoretical discussions that are presented in Chapter 4, it is concluded that Krippendorff's Alpha is the most appropriate method to be used with the interval approach.

In brief, in the agreement measurements of the TDB, words should be used as coding units, and by considering research aspects, both the interval and the boundary approach can be used as context unit approaches. When the boundary approach is preferred, both Fleiss' Kappa and Krippendorff's Alpha can be used to measure agreement. However, when the interval approach is preferred, only Krippendorff's Alpha can be used to measure the agreement. In order to decrease the number of methods used, one should use Krippendorff's Alpha for both the boundary and interval approaches. With these statements, the twelve combinations of agreement statistics and their data unitization preferences (see Table 13 in § 5.1) are decreased to two (AWB and AWI).

5.3 Future Discourse Analysis

So far, the only published TDB paper has been Zeyrek et al (2009), which evaluates annotations quantitatively. In this paper, Zeyrek et al. (2009) elaborate on two discontinuous Turkish discourse connectives (hem … hem 'neither … nor', ne … ne 'both … and'). They state that the inter-annotator agreements on these two discontinuous connectives are higher than the anaphoric connectives. In the discontinuous connective annotations, the boundary of arguments are determined by the connective itself and according to Zeyrek et al. (2009) this is the sole reason of the higher agreement.

Surely, with this thesis, the number of quantitative annotation evaluations will increase. Because this thesis reveals research opportunities on qualitative annotation evaluation by both clarifying the usage of the statistical measurements on the TDB and presenting a handy software tool for these statistical measurements (see § 5.4). In this section, possible research directions will be grouped and discussed in two classes in order to reveal some future research directions: (1) directly using the measurements in evaluations, and (2) comparing the measurements that are collected with different data unitization approaches on the same discourse connective annotations.

The following is a brief discussion of some examples with respect to the mentioned research directions:

1- In addition to rating the reliability of connective annotations, inter-annotator agreement measurements can be used to evaluate linguistic aspects of different connectives. As Zeyrek et al. (2009) conducted, connectives or connective types can be compared with each other, and the variation can be used to support or contradict a linguistic proposal. For example, current measurements can be used to support the following proposal: *Annotating the Arg1 of anaphoric connectives is cognitively harder than annotating the Arg1 of other types of connectives* *because Arg1 of anaphoric connectives can be away from its connective a few sentences or even paragraphs.* According to the current (AWI) measurements, Arg1 measurements exhibit the perfect agreement (0.8) limit: yandan 'on the one hand' (0.638), rağmen 'despite' (0.642), için 'because' (0.671), ayrıca 'in addition' (0.719). Even with a superficial look at the Arg1 measurements, it is seen that the lowest two agreement values are produced by the annotations of discourse connectives which seem like subordinators, but also contain anaphoric forms (e.g., buna rağmen 'despite this', bunu için 'because of this', etc.). Additionally, by performing the measurements at different stages of the annotation effort, the change in the inter-annotator agreement by time can be observed to determine the effect of gaining annotation experience.

2- As discussed in the preceding section, there are two proposed measurement ways (AWB and AWI), which are sound and suitable for the TDB. Unlike the AWB method, the AWI method makes a distinction between total disagreement and partial disagreement by weighing the overlapping annotation spans. One can make various inferences by using these two methods with the same connective annotations. When such research is conducted, there will occur three results to interpret:

   a. Higher AWB result: In general, the overlapping parts of the annotations are shorter than the non-overlapping parts.

   b. Higher AWI result: In general, the overlapping parts of the annotations are longer than the non-overlapping parts.

   c. Equal results: Total agreement or disagreement.

The two research approaches above can be diversified into aim-specific subgroups with more fine-tuned research questions.

5.4 THE RATER AGREEMENT TOOL (RAT)

During the study of this thesis, an original computer program has been developed to perform inter-annotator agreement measurements on the TDB. The program is called Rater Agreement Tool (RAT). As the name indicates, the program is designed to perform inter-rater agreement measurement not only for the TDB but also for any other rating-based efforts.

The RAT works with an input configuration file which is supposed to feed the program for the aimed measurement (See Appendix F for a sample configuration file). In the configuration file, the following input is expected from the user:

- The path of the annotation files, and original texts (workspace).
- The list of connectives for which the inter-annotator agreement will be measured.
- The list of annotators among which the inter-annotator agreement will be measured.
- The path of the output files.
- The setting of the measurements: (1) the agreement statistic, (2) the coding unit (character or word), and (3) the context unit approach (interval or boundary). These setting options totally cover the combinations those are presented in this thesis (see Table 13 in § 5.1).

The work flow of the RAT is described in the following flow charts. In the first part of the flow chart, the data internalization and data handling mechanisms of the RAT are described:

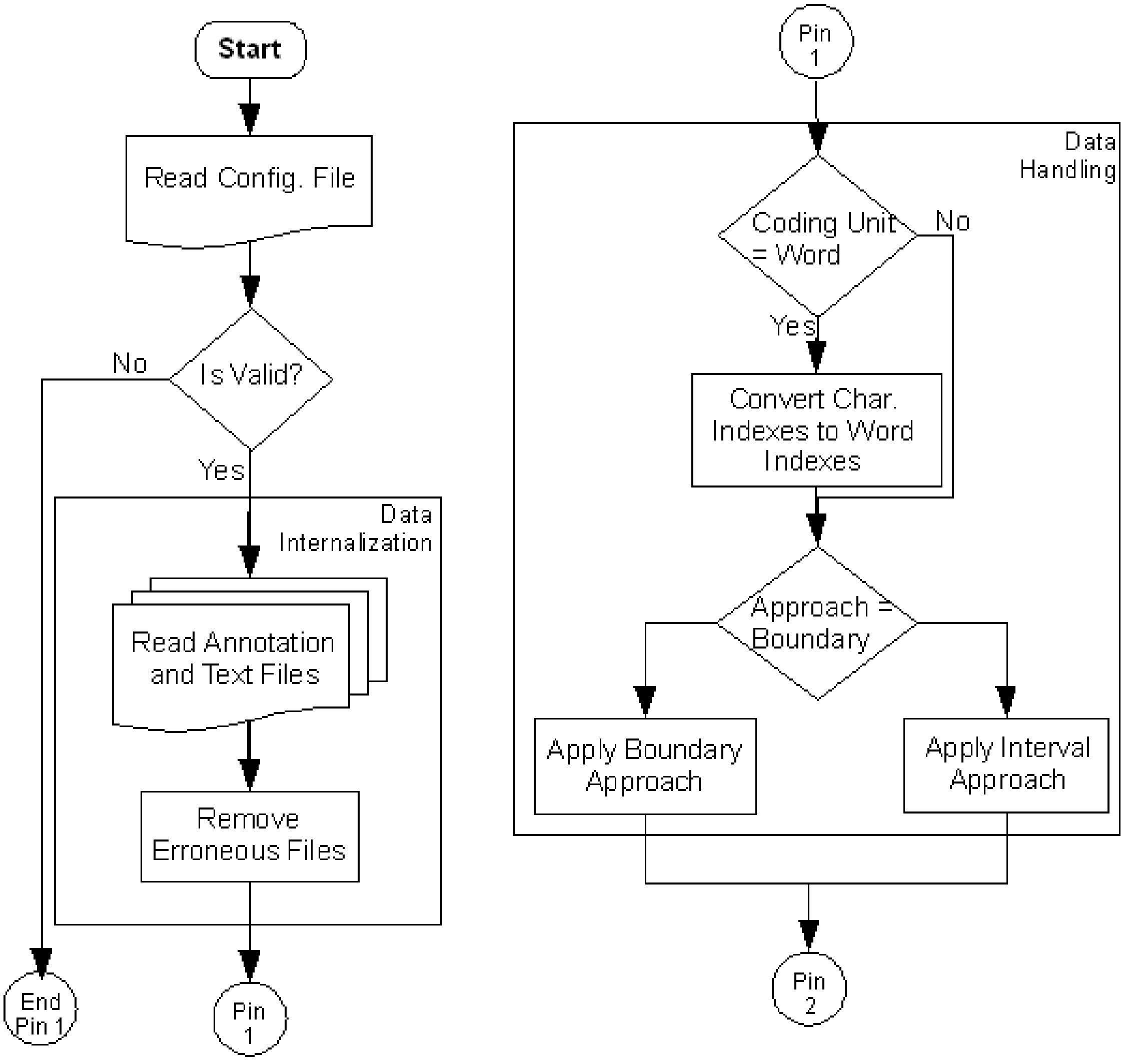


Figure 26 The flow chart of RAT - Part I

In the second part of the flow chart, agreement measurement and output mechanisms are described:

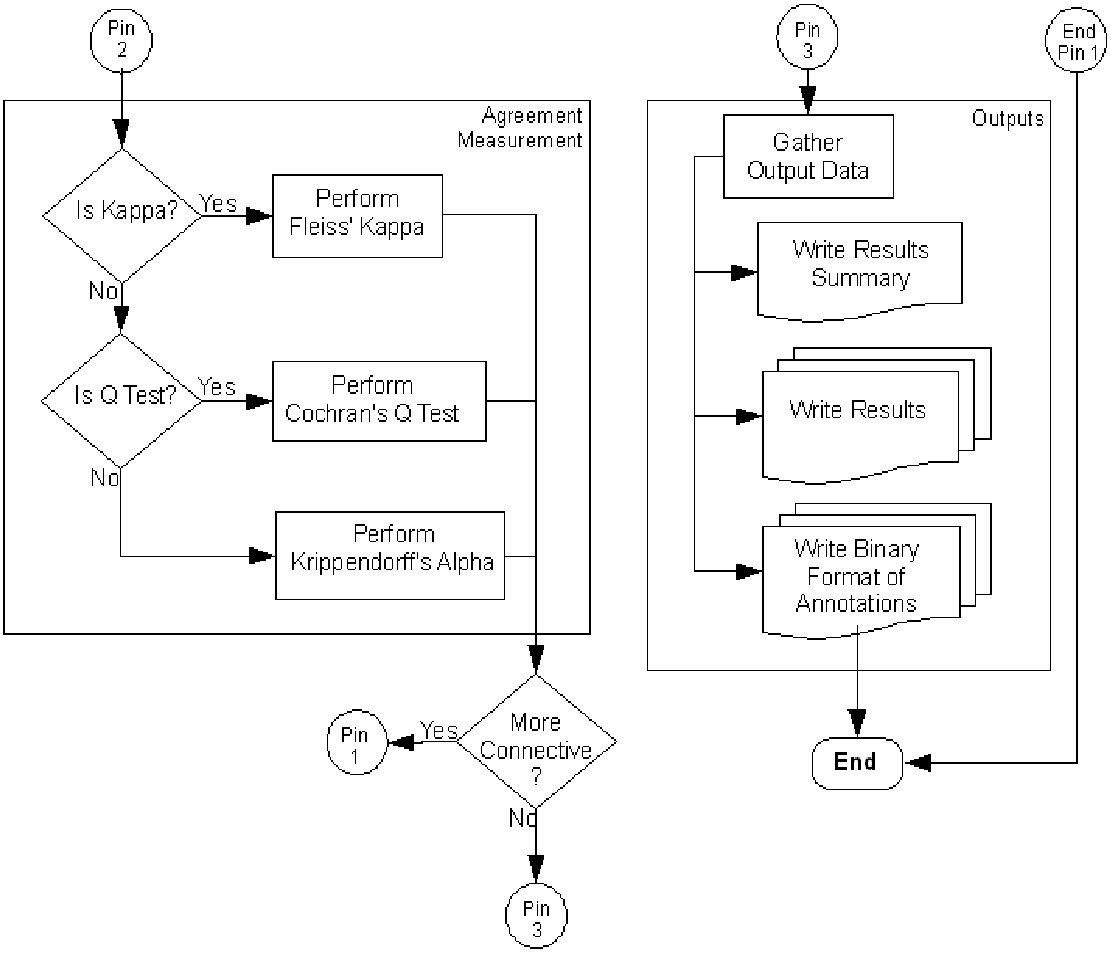


Figure 27 The flow chart of RAT - Part II

The RAT needs a valid configuration file in order to measure the inter-annotator agreement. After validating the configuration file, the program reads all annotation data files. The reading process is followed by a validation step. In the validation step, it is checked whether:

(1) The annotation files have their raw texts in the workspace. This step is performed to be sure that the annotation file is deliberately in the workspace (when the raw texts are not available this step can be canceled).

(2) The number of annotations in an annotation file that are annotated by an annotator is the same with the rest of the annotators. In the scope of this thesis, the connectives are not accepted as annotation elements. Therefore, it is expected that all annotators have the same number of connective annotations for the same source file.

When one of the above validation steps fails, the erroneous annotation file is excluded from the inter-annotator agreement measurements.

After internalizing the annotation data, the data handling stage begins. In this stage, the RAT digitizes the annotations according to the configuration file preference by following the methodology that is described in the sections 4.6 and 4.7. Afterwards, the digitized data is used to calculate the inter-annotator agreement results by Cochran's Q test, Fleiss' Kappa, or Krippendorff's Alpha according to the settings in the configuration file. In the cases where more than one connective or more than one setting is given as input, the RAT repeats data handling and agreement measurement methods. When the agreement results for all the given connectives and settings are calculated, the RAT produces three output files:

1- Results Summary: One results summary file that contains the following information is produced:
    a. Names of connectives
    b. Annotator number for each connective annotations
    c. Raw text file number for each connective annotation
    d. Annotated discourse relation number for each connective
    e. Arg1 and Arg2 results for each setting for each connective annotation

    (See Table 14, 15, 16, 17, 18, and 19 for result summary)
2- Results: For each connective, one result file is produced. Each result file contains the following information for each setting:
    a. Context unit approach (boundary - interval)

b. Coding unit (character - word)

c. Annotator number

d. Name of the agreement statistic

e. Raw text file number

f. Annotated discourse structure number

g. Arg1 and Arg2 results with some intermediate results

h. Paths of the annotation files that is used in calculations

(See Appendix G for a sample result file)

3- Binary Format of Annotations: One result file is produced for each setting of each connective. Each file contains binary formatted and space delimited, digitized annotations. The format of these files enables the usage of third party agreement statistics tools. As a result, the user can verify the results that are found by the RAT via the third party tools.

The RAT is currently only available to the TDB project group and it is not open for the community usage. Despite the generic design of its agreement measurement part, the other parts of the RAT are still TDB-specific. As future works, the RAT shall be revised and documented, prior to its presentation to the community usage. In addition to these two future works, the RAT shall also be updated to handle discontinuous argument annotations (§ 4.5). Currently, the RAT discards discontinuous annotations.

The RAT is an original, steady and stand-alone computer program that can perform and report inter-annotator agreement measurements on the TDB. It is developed completely during the study of this thesis in Java programming language. The sources of algorithms and code parts that are used in RAT are as follows:

- The Java class that calculates Cochran’s Q test is developed according to Cochran (1950).

- The code part which is used to calculate p-value of Cochran's Q test is taken from http://www.alglib.net/specialfunctions/distributions/chisquare.php (Bochkanov & Bystritsky, 2010).
- The Java class that calculates Krippendorff's Alpha is developed according to Krippendorff (1995 & 2004b).
- The code part that calculates Fleiss' Kappa is taken from http://en.wikibooks.org/wiki/Algorithm_implementation/Statistics/Fleiss%27_kappa (Anonymous, 2010)

Instead of using a third party agreement statistics tool, the RAT is developed to handle the agreement statistics that are discussed in this thesis, and to enable a comprehensive inter-rater agreement measurement tool to the community.

CHAPTER VI

SUMMARY AND CONCLUSION

In this thesis, firstly Cochran’s Q test (1950) and a selection of agreement coefficients were discussed in detail. The agreement coefficients were elaborated under two main headings: the Kappa family, and Krippendorff’s Alpha (1995). In the Kappa family, Benett's Sigma (1954), Scott's Pi (1955), Cohen's Kappa (1960), and Fleiss’ Kappa (1971) were compared. According to the theoretical proposals and the statistical capabilities of the methods, Cochran’s Q test, Fleiss’ Kappa and Krippendorff’s Alpha were selected as the candidate agreement statistics for the reliability measurements of the TDB.

Afterwards, the PDTB, the PDTB-like annotation efforts (the HDRB, the LADTB, and the CDTB), and an annotation effort which is based on the RST framework were examined. During the examinations, the annotation effort itself and the statistical methods that are used to measure the inter-annotator reliability were analyzed. In order to improve the reliability approach of the TDB, some lessons are drawn from the reported statistical methods of these efforts.

After the statistic method discussions, and the examination of the annotation efforts, the TDB was presented in detail by elaborating on the following issues:

- The Turkish connectives that are defined in the TDB, the annotation cycle of the TDB,
- The dependency analysis of the Turkish discourse structures that are encountered in the TDB,
- The annotation representation/storage of the TDB.

Subsequently, the data handling challenges of the TDB were defined and discussed. There were two main challenges: the unitization of annotations and the removal of the computational complexities. For the unitization of annotations, two unitization approaches were suggested:

1- Using characters or words as coding unit
2- Using the boundary or the interval approach as context unit approach

Then, solutions were suggested to remove the computational complexity in the annotations, i.e., to find computationally simpler equivalents of the annotations. All the suggestions and solutions for the challenges are original proposals of this thesis.

According to the discussed data handling challenges, inter-annotator agreement measurements of the TDB was performed by using Cochran's Q test, Fleiss' Kappa, and Krippendorff's Alpha with the combinations of character and word as coding units, and the boundary and the interval approach as context unit approaches. The measurements were performed by a computer program (RAT), which is originally developed as an end-product of this thesis.

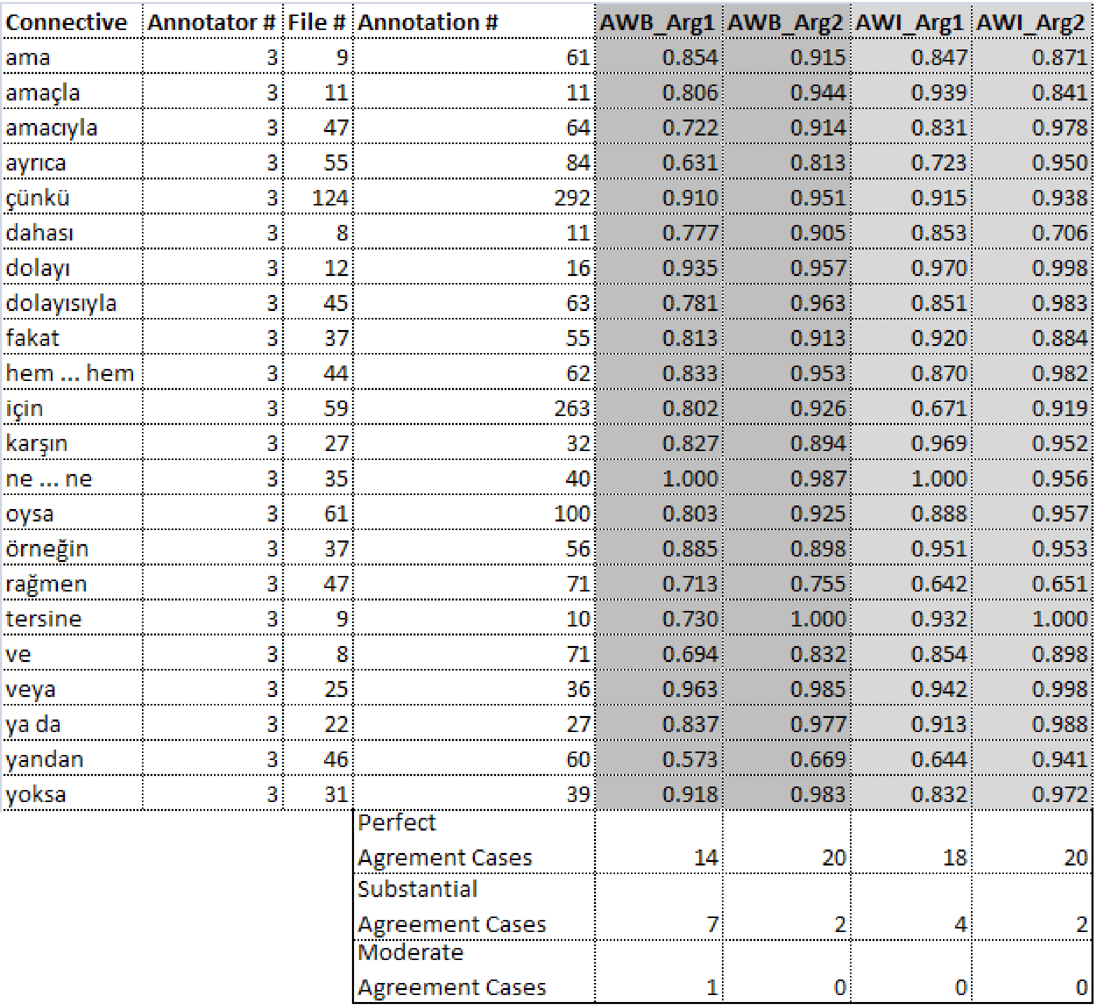
As a result, two statistical methods were defined as valid and sound for the TDB inter-annotator agreement measurements. A summary is presented in Table 21.

Table 21 The valid statistical methods for the TDB

Useful for	Statistical Method	Coding Unit	Context Unit Approach
Weighting the partial-overlap disagreements	Krippendorff's Alpha	Word	Interval
Treating all disagreement sources the same	Krippendorff's Alpha	Word	Boundary

The agreement measurements of the TDB for Arg1 and Arg2 for 22 connectives for AWB and AWI are presented in Table 22. The table also includes the number of perfect, substantial and moderate agreement cases determined by each statistical method for each argument.

Table 22 AWB and AWI results of the TDB for Arg1 and Arg2 for 22 connectives



Connective	Annotator #	File #	Annotation #	AWB_Arg1	AWB_Arg2	AWI_Arg1	AWI_Arg2
ama	3	9	61	0.854	0.915	0.847	0.871
amaçla	3	11	11	0.806	0.944	0.939	0.841
amacıyla	3	47	64	0.722	0.914	0.831	0.978
ayrıca	3	55	84	0.631	0.813	0.723	0.950
çünkü	3	124	292	0.910	0.951	0.915	0.938
dahası	3	8	11	0.777	0.905	0.853	0.706
dolayı	3	12	16	0.935	0.957	0.970	0.998
dolayısıyla	3	45	63	0.781	0.963	0.851	0.983
fakat	3	37	55	0.813	0.913	0.920	0.884
hem ... hem	3	44	62	0.833	0.953	0.870	0.982
için	3	59	263	0.802	0.926	0.671	0.919
karşın	3	27	32	0.827	0.894	0.969	0.952
ne ... ne	3	35	40	1.000	0.987	1.000	0.956
oysa	3	61	100	0.803	0.925	0.888	0.957
örneğin	3	37	56	0.885	0.898	0.951	0.953
rağmen	3	47	71	0.713	0.755	0.642	0.651
tersine	3	9	10	0.730	1.000	0.932	1.000
ve	3	8	71	0.694	0.832	0.854	0.898
veya	3	25	36	0.963	0.985	0.942	0.998
ya da	3	22	27	0.837	0.977	0.913	0.988
yandan	3	46	60	0.573	0.669	0.644	0.941
yoksa	3	31	39	0.918	0.983	0.832	0.972
			Perfect Agrement Cases	14	20	18	20
			Substantial Agreement Cases	7	2	4	2
			Moderate Agreement Cases	1	0	0	0

According to Table 22, the average agreements of all the measurement methods indicate that there is perfect agreement for both Arg1 and Arg2 annotations of 22 connectives of the TDB:

- AWB_Arg1: 0.809
- AWB_Arg2: 0.912
- AWI_Arg1: 0.862
- AWI_Arg2: 0.923

In this thesis, two future research directions were also briefly discussed. First, it was mentioned that, the inter-annotator agreement measurements are not only used to determine the reliability of the TDB but they can also be used to compare the cognitive complexity of the discourse connectives. Secondly, it was proposed that the disagreements on the discourse connective types can be characterized by using the results of the above presented two measurement combinations (AWB and AWI) together.

The thesis also emphasized the need for cross-linguistic comparison of annotation efforts. Needless to say, the prerequisites of a cross-linguistic comparison are performing the annotations by following similar principles and annotation cycles. Afterwards, the comparability of the annotations of the different project groups can be sustained by the assets of this thesis (the statistical methods, the data handling mechanisms, data unitization approaches, and the developed computer program).

REFERENCES

Aktaş, B., Bozşahin, C., & Zeyrek, D. (2010). *Discourse relation configurations in Turkish and an annotation environment*. Paper presented at the Fourth Linguistic Annotation Workshop (LAW IV) ACL 2010.

Al-Saif, A., & Markert, K. (2010). *The Leeds arabic discourse treebank: annotating discourse connectives for arabic.* Paper presented at the Proceedings of the conference on Language Resources and Evaluation.

Anonymous. Algorithm Implementation/Statistics/Fleiss' kappa Retrieved 29 August 2010, from http://en.wikibooks.org/wiki/Algorithm_implementation/Statistics/Fleiss%27_kappa

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555-596.

Begum, R., Husain, S., Dhwaj, A., Sharma, D. M., Bai, L., & Sangal, R. (2008). *Dependency annotation scheme for indian languages.* Paper presented at the Proceedings of IJCNLP-2008.

Benett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited questioning. *Public Opinion Quarterly, 18*(3), 303-308.

Bochkanov, S., & Bystritsky, V. Chi-square distribution - ALGLIB Retrieved 29 August 2010, from http://www.alglib.net/specialfunctions/distributions/chisquare.php

Burnage, G., & Baguley, G. (1996). *The British National Corpus*: South Bank University, Library Information Technology Centre.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics, 22*(2), 249-254.

Cochran, W. G. (1950). The Comparison of Percentages in Matched Samples. *Biometrika, 37*, 256-266.

Cohen, J. (1960). A Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101.

Demirşahin, I. n., Yalçınkaya, I., & Zeyrek, D. (2010). *Pair annotation: adaption of pair programming to corpus*. Cognitive Science Department. Middle East Technical University. Ankara.

Di Eugenio, B., & Glass, M. (2004). The Kappa statistic: a second look. *Computational Linguistics, 30*(1), 95-101.

Dillon, W. R., & Mulani, N. (1984). A Probabilistic Latent Class Model for Assessing Inter-Judge Reliability. *Multivariate Behavioral Research, 19*(4), 438-458.

Dinesh, N., Alan, L., Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2005). *Attribution and the (non)-alignment of syntactic and discourse arguments of connectives.* Paper presented at the Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Francis, W. N., Kuc\030Cera, H., & Mackie, A. W. (1982). *Frequency analysis of English usage : lexicon and grammar*. Boston: Houghton Mifflin.

Gwet, K. (2001). *Handbook of inter-rater reliability*. Gaithersburg: STATAXIS Publishing Company.

Hartmann, D. (1977). Considerations in the choice of interobserver reliability estimates. *J Appl Behav Anal, 10*(1), 103–116.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.

Hsu, L. M., & Field, R. (2003). Interrater agreement measures: comments on kappan, cohen's kappa, scott's π, and aickin's agr. *Understanding Statistics, 2*(3), 205-219.

Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology, 25*, 47-76.

Krippendorff, K. (2004a). *Content analysis : an introduction to its methodology* (2nd ed.). Thousand Oaks, Calif.: Sage.

Krippendorff, K. (2004b). Measuring the Reliability of Qualitative Text Analysis Data. *Humanities, Social Sciences and Law, 38*(6), 787-800.

Krippendorff, K. (2004c). Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411-433.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *International Biometric Society, 33*(1), 159-174.

Lee, A., Prasad, R., Joshi, A., Dinesh, N., & Webber, B. (2006). *Complexity of dependencies in discourse.* Paper presented at the Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories.

Lynn, C., Marcu, D., & Okurowski, M. E. (2003). *Building a discourse-tagged corpus in the framework of rhetorical structure theory*. Paper presented at the Current and New Directions in Discourse and Dialogue, Kluwer, Dordrecht.

Maamouri, M., & Bies, A. (2004). *Developing an arabic treebank: methods, guidelines, procedures, and tools.* Paper presented at the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (COLING).

Mann, W. C., & Thompson, S. A. (1987). *Rhetorical Structure Theory: A Theory of TextOrganization*. No. ISI/RS-87–190. Information Sciences Institute. Marina del Rey, CA.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory:Toward a functional theory of text organization. *Text, 8*(3), 243-281.

Marcu, D., Amorrortu, E., & Romera, M. (1999). *Experiments in constructing a corpus of discourse trees.* Paper presented at the Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging, College Park, MD.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpusof english: the penn treebank. *Computational Linguistics, 19*(2), 313-330.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics : an introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

McNemar, Q. (1949). *Psychological Statistics*. New York: John Wiley and Sons.

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2004). *The Penn Discourse TreeBank.* Paper presented at the Language Resources and Evaluation Conference, Lisbon, Portugal.

Oza, U., Prasad, R., Kolachina, S., Meena, S., Sharma, D. M., & Joshi, A. (2009a). *Experiments with annotating discourse relations in the hindi discourse relation bank.* Paper presented at the Proceedings of the Seventh International Conference on Natural Language Processing.

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., & Joshi, A. (2009b). *The Hindi discourse relation bank.* Paper presented at the Proceedings of the Third Linguistic Annotation Workshop, Suntec, Singapore.

Palmer, M., Guildea, D., & Kingsbury, P. (2005). The Proposition bank: an annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71–106.

PDTB-Group. (2006). *The Penn Discourse TreeBank 1.0 Annotation Manual*. Technical Report IRCS-06-01. Institute for Research in Cognitive Science. University of Pennsylvania.

PDTB-Group. (2008). *The Penn Discourse TreeBank 2.0 Annotation Manual*. Technical Report IRCS-08-01. Institute for Research in Cognitive Science. University of Pennsylvania.

Prasad, R., Dinesh, N., Lee, A., Joshi, A., & Webber, B. (2006). *Annotating Attribution in the Penn Discourse TreeBank.* Paper presented at the Proceedings of the COLING/ACL Workshop on Sentiment and Subjectivity in Text, Sydney, Australia.

Prasad, R., Dinesh, N., Lee, A., Joshi, A., & Webber, B. (2007). Attribution and its annotation in the Penn Discourse TreeBank. *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse, 47*(2).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., et al. (2008a). *The Penn Discourse TreeBank 2.0.*

Prasad, R., Husain, S., Sharma, D. M., & Joshi, A. (2008b). *Towards an annotated corpus of discourse relations in hindi.* Paper presented at the Proceedings of the IJCNLP-08 Workshop on Asian Language Resources.

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

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly, 19*(3), 321-325.

Siegel, S. (1957). Nonparametric statistics. *The American Statistician, 11*(3), 13-19.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Taboada, M., & Mann, W. C. (2006). Rhetorical structure theory: looking back and moving ahead. *Discourse Studies, 8*(3), 423-459.

TDB-Group. (2010). *ODTÜ MEDİD İşaretleme İlkeleri*. Cognitive Science Department. Middle East Technical University. Ankara.

Truex, D. (2006). *Text-based analysis: a brief introduction*. Georgia State University.

Tukey, J. W. (1977). *Exploratory Data Analysis*: Addison-Wesley Publishing Company.

Webber, B., Joshi, A., Miltsakaki, E., Prasad, R., Dinesh, N., Lee, A., et al. (2006). *A Short introduction to the penn discourse treebank.* Paper presented at the Copenhagen Working Papers in Language and Speech Processing.

Xue, N. (2005). *Annotating discourse connectives in the chinese treebank.* Paper presented at the Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky.

Xue, N., Xia, F., Chiou, F.-d., & Palmer, M. (2005). The Penn chinese treebank: phrase structure annotation of a large corpus. *Natural Language Engineering, 11*(2), 207-238.

Yöndem-Turhan, M. (2001). *Identifying the Interactions of Multi-Criteria In Turkish Discourse Segmentation*. Department of Computer Engineering. Middle East Technical University.

Zeyrek, D., Demirşahin, I. n., Sevdik-Çallı, A., Ögel-Balaban, H., Yalçınkaya, I., & Turan, Ü. D. (2010). *The Annotation scheme of the turkish discourse bank and an evaluation of inconsistent annotations*. Paper presented at the Fourth Linguistic Annotation Workshop (LAW IV).

Zeyrek, D., Turan, Ü. D., Bozşahin, C., Çakıcı, R., Sevdik-Çallı, A., Yalçınkaya, I., et al. (2009). *Annotating subordinators in the turkish discourse bank.* Paper presented at the Proceedings of the Third Linguistic Annotation Workshop.

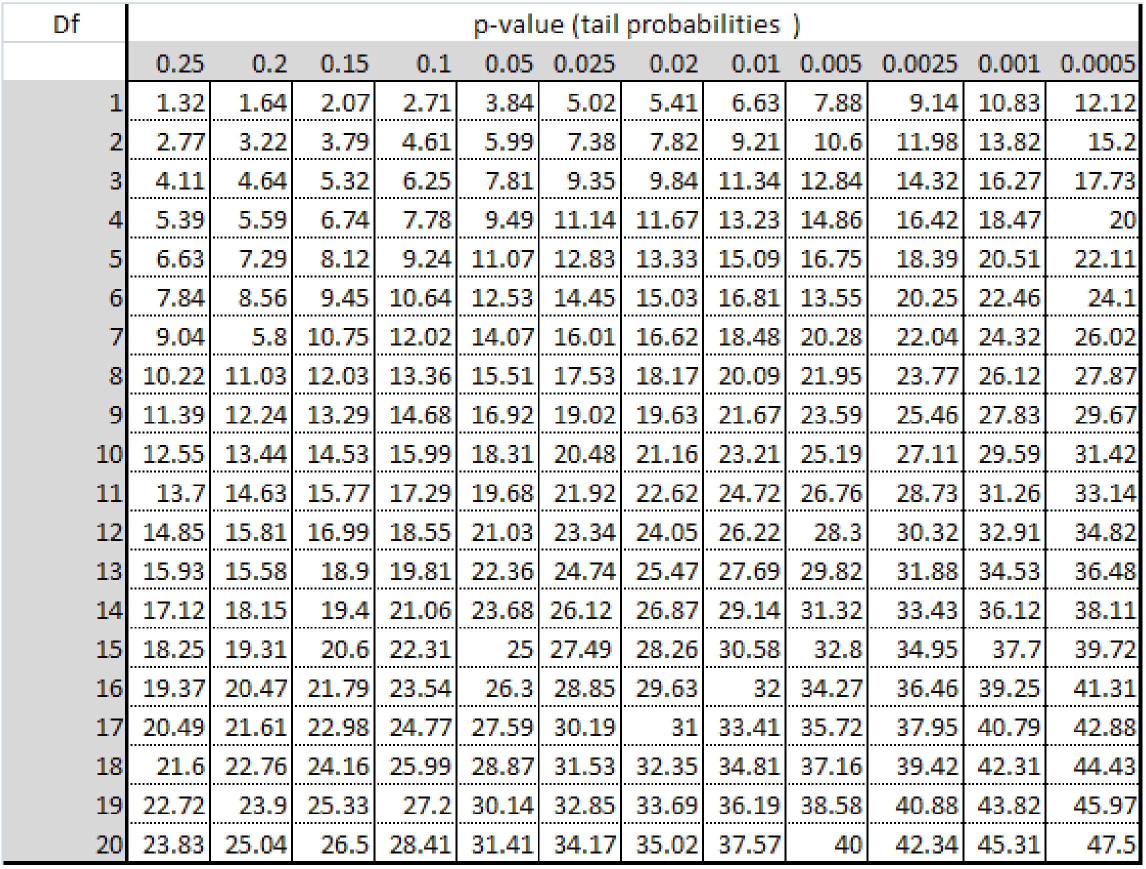
Zeyrek, D., & Webber, B. (2008). *A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus.* Paper presented at the In Proceedings of IJCNLP-2008., Hyderabad, India.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*(3), 374-378.

APPENDICES

APPENDIX A – CHI SQUARE TABLE

Table 23 Chi Square Table



Df	p-value (tail probabilities)											
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.6	11.98	13.82	15.2
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.59	6.74	7.78	9.49	11.14	11.67	13.23	14.86	16.42	18.47	20
5	6.63	7.29	8.12	9.24	11.07	12.83	13.33	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.53	14.45	15.03	16.81	13.55	20.25	22.46	24.1
7	9.04	5.8	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.63	21.67	23.59	25.46	27.83	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.7	14.63	15.77	17.29	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.3	30.32	32.91	34.82
13	15.93	15.58	18.9	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.4	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.6	22.31	25	27.49	28.26	30.58	32.8	34.95	37.7	39.72
16	19.37	20.47	21.79	23.54	26.3	28.85	29.63	32	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31	33.41	35.72	37.95	40.79	42.88
18	21.6	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.9	25.33	27.2	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.5	28.41	31.41	34.17	35.02	37.57	40	42.34	45.31	47.5

APPENDIX B – THE JAVA IMPLEMENTATION OF COCHRAN'S Q TEST

```
public static double CalculateQ(Vector<Integer>[] Data, int rows, int
columns)
{
    double Q=0;
    int i,j;
    double[] RT = new double[rows];
    double[] RT2 = new double[rows];
    double sumOfRT2s = 0;
    double[] CT = new double[columns];
    double[] CT2 = new double[columns];
    double sumOfCTs =0;
    double sumOfCT2s =0;

    for(i=0; i<rows;i++)
    {
        RT[i]=0;
        for(j=0; j<columns; j++ )
        {
            RT[i]+=Data[j].elementAt(i);
        }
        RT2[i] = RT[i]*RT[i];
        sumOfRT2s += RT2[i];
    }

    for(j=0; j<columns; j++)
    {
        CT[j]=0;
        for(i=0; i<rows; i++)
        {
            CT[j]+=Data[j].elementAt(i);
        }
        CT2[j]=CT[j]*CT[j];
        sumOfCTs +=CT[j];
        sumOfCT2s += CT2[j];
    }

    Q= ((columns)*(columns-1)*sumOfCT2s)-((columns-
1)*sumOfCTs*sumOfCTs);
    if((columns*sumOfCTs-sumOfRT2s)==0)
    {
        Q = 0;
    }
    else
    {
        Q = Q / (columns*sumOfCTs-sumOfRT2s);
    }
    return Q;
}
```

APPENDIX C – THE JAVA IMPLEMENTATION OF FLEISS’ KAPPA

```
//Source:
http://en.wikibooks.org/wiki/Algorithm_implementation/Statistics/Fleiss
%27_kappa
public static float computeKappa(short[][] mat)
{
    final int n = checkEachLineCount(mat) ;
    final int N = mat.length ;
    final int k = mat[0].length ;
    // Computing p[]
    float[] p = new float[k] ;
    for(int j=0 ; j<k ; j++)
    {
        p[j] = 0 ;
        for(int i=0 ; i<N ; i++)
            p[j] += mat[i][j] ;
        p[j] /= N*n ;
    }
    // Computing P[]
    float[] P = new float[N] ;
    for(int i=0 ; i<N ; i++)
    {
        P[i] = 0 ;
        for(int j=0 ; j<k ; j++)
            P[i] += mat[i][j] * mat[i][j] ;
        P[i] = (P[i] - n) / (n * (n - 1)) ;
    }
    // Computing Pbar
    float Pbar = 0 ;
    for(float Pi : P)
```

```
        Pbar += Pi ;
    Pbar /= N ;
    // Computing PbarE
    float PbarE = 0 ;
    for(float pj : p)
        PbarE += pj * pj ;
    final float kappa = (Pbar - PbarE)/(1 - PbarE) ;
    return kappa ;
}

private static int checkEachLineCount(short[][] mat)
{
    int n = 0 ;
    boolean firstLine = true ;
    for(short[] line : mat)
    {
        int count = 0 ;
        for(short cell : line)
            count += cell ;
        if(firstLine)
        {
            n = count ;
            firstLine = false ;
        }
        if(n != count)
            throw new IllegalArgumentException("Line count != "+n+" (n
value).") ;
    }
    return n ;
}
```

APPENDIX D – THE JAVA IMPLEMENTATION OF KRIPPENDORFF'S ALPHA

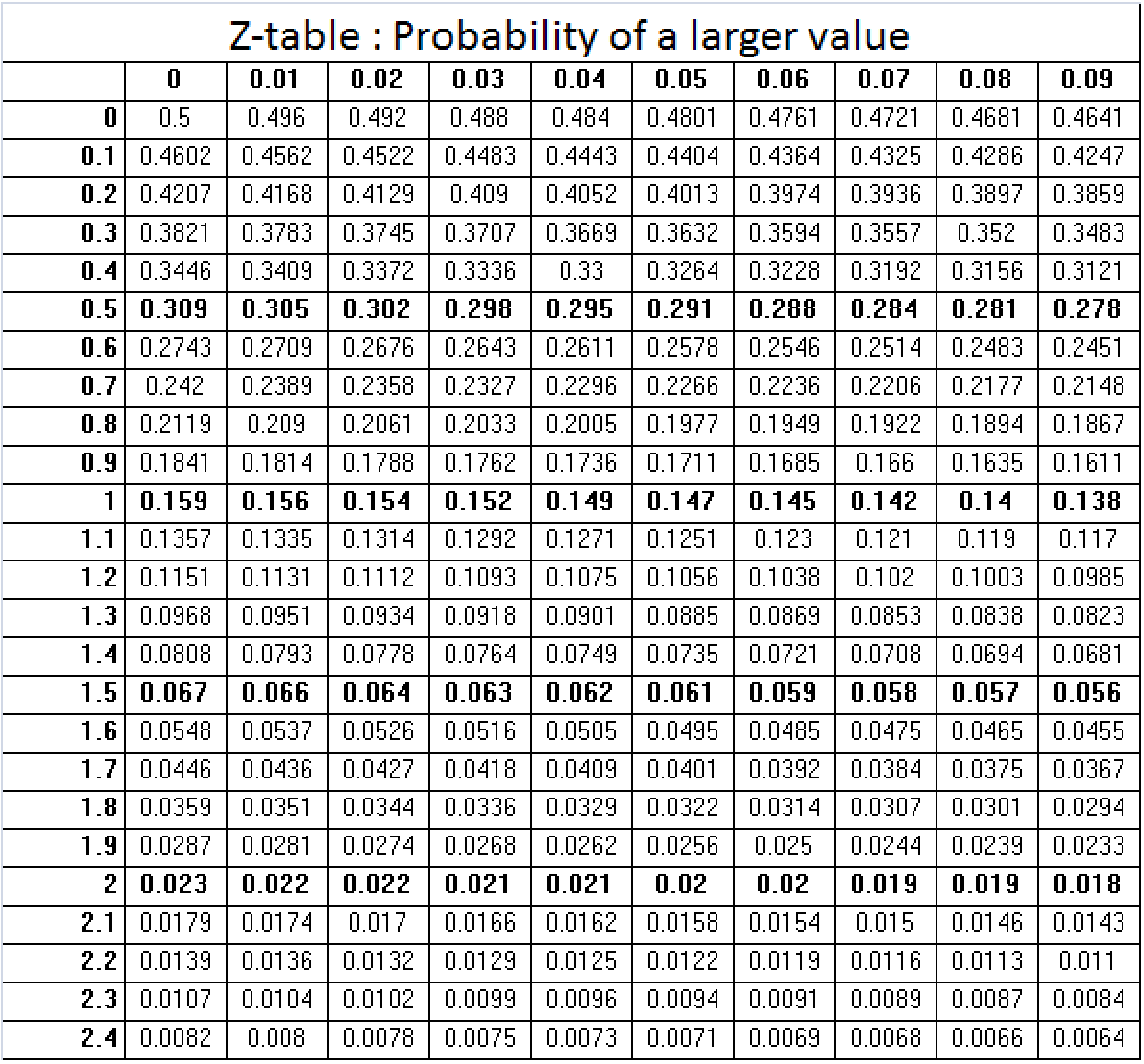
```
    public double calculateAlpha(int categoryIndex)
    {
        //A=1 perfect relaibility, A=0 absence of relaibilty
        double A;
        if(categoryIndex == 0 || categoryIndex == 1)
        {
            double Doc =calcDoc(categoryIndex);
            double Dec =calcDec(categoryIndex);
            A = 1.0-(Doc/Dec);
        }
        else
        {
             A = 1.0-((calcDoc(0)+calcDoc(1))/(calcDec(0)+calcDec(1)));
        }
        return A;
    }

    private double distance(Segment segA, Segment segB)
    {
        double dis=0.0;
        if(segA.v==1 && segB.v==1 && ((-1*segA.l)<(segA.b-segB.b)) &&
((segA.b-segB.b)<segB.l))
        {
            dis = (segA.b-segB.b)*(segA.b-segB.b)+(segA.b+segA.l-
segB.b-segB.l)*(segA.b+segA.l-segB.b-segB.l);
        }
        else if(segA.v==1 && segB.v==0 && ((segB.l-segA.l)>=(segA.b-
segB.b)) && ((segA.b-segB.b)>=0))
        {
            dis = segA.l*segA.l;
        }
        else if(segA.v==0 && segB.v==1 && ((segB.l-segA.l)<=(segA.b-
segB.b)) && ((segA.b-segB.b)<=0))
        {
            dis = segB.l*segB.l;
        }
        else
        {
            dis=0.0;
        }
        return dis;
    }

    public double calcDoc(int categoryIndex)
    {
        double Doc = 0.0;
        if(categoryIndex>=_NumCategory)
        {
            return -999999.9;
        }
        for(int i=0; i<_NumAnnotator; i++)
        {
            for(int g=0; g<_Annotations[categoryIndex][i].size();g++)
            {
                for(int j=0; j<_NumAnnotator; j++)
                {
                    if(i!=j)
                    {
                        for(int h=0;
h<_Annotations[categoryIndex][j].size();h++)
                        {

Doc+=distance(_Annotations[categoryIndex][i].elementAt(g),
_Annotations[categoryIndex][j].elementAt(h));
                        }
                    }
                }
            }
        }
        Doc /= _NumAnnotator*(_NumAnnotator-1)*_Length*_Length;
        return Doc;
    }

    double calcDec(int categoryIndex)
    {
        double Dec = 0.0;
        double delimDec = 0.0;
        double Nc = calcNc(categoryIndex);
        Segment mySeg,segCIG, segCJH;
        double tmp1;

        for(int i=0; i<_NumAnnotator; i++)
        {
            for(int g=0; g<_Annotations[categoryIndex][i].size(); g++)
            {
                segCIG = _Annotations[categoryIndex][i].elementAt(g);
                tmp1=0;
                if(segCIG.v!=0)
                {
                    for(int j=0; j<_NumAnnotator; j++)
                    {
                        for(int h=0;
h<_Annotations[categoryIndex][j].size(); h++)
                        {
                            segCJH =
_Annotations[categoryIndex][j].elementAt(h);
```

```
                            if(segCJH.l>=segCIG.l)
                            {
                                tmp1 += (1-segCJH.v)*(segCJH.l-
segCIG.l+1);
                            }
                        }
                    }
                    tmp1*=segCIG.l*segCIG.l;
                    Dec+=segCIG.v*((((Nc-
1)/3.0)*(2*segCIG.l*segCIG.l*segCIG.l-
3*segCIG.l*segCIG.l+segCIG.l))+tmp1);
                }
            }
        }

        Dec *= (2.0/(double)_Length);

        for(int i=0; i<_NumAnnotator; i++)
        {
            for(int g=0; g<_Annotations[categoryIndex][i].size(); g++)
            {
                mySeg = _Annotations[categoryIndex][i].elementAt(g);
                delimDec+= mySeg.v*mySeg.l*(mySeg.l-1);
            }
        }
        delimDec = _NumAnnotator*_Length*(_NumAnnotator*_Length-1)-
delimDec;

        Dec /= delimDec;

        return Dec;
    }

    double calcNc(int categoryIndex)
```

```
{
    double Nc = 0.0;
    for(int i=0; i<_NumAnnotator; i++)
    {
        for(int g=0; g<_Annotations[categoryIndex][i].size();g++)
        {
            Nc+=_Annotations[categoryIndex][i].elementAt(g).v;
        }
    }
    return Nc;
}
```

APPENDIX E – Z TABLE

Table 24 Z-table: Probability of a larger value



Z-table : Probability of a larger value

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.496	0.492	0.488	0.484	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.409	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.352	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.33	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	**0.309**	**0.305**	**0.302**	**0.298**	**0.295**	**0.291**	**0.288**	**0.284**	**0.281**	**0.278**
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.242	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.209	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.166	0.1635	0.1611
1	**0.159**	**0.156**	**0.154**	**0.152**	**0.149**	**0.147**	**0.145**	**0.142**	**0.14**	**0.138**
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.123	0.121	0.119	0.117
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.102	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	**0.067**	**0.066**	**0.064**	**0.063**	**0.062**	**0.061**	**0.059**	**0.058**	**0.057**	**0.056**
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.025	0.0244	0.0239	0.0233
2	**0.023**	**0.022**	**0.022**	**0.021**	**0.021**	**0.02**	**0.02**	**0.019**	**0.019**	**0.018**
2.1	0.0179	0.0174	0.017	0.0166	0.0162	0.0158	0.0154	0.015	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.011
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.008	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064

APPENDIX F – THE RAT SAMPLE CONFIGURATION FILE

```
[Work Path]
C:/ThesisMeasurements/ortam1_240710/
[Connectives]
oysa
[Annotators]
Annotator1
Annotator2
Annotator3
[Out Directory]
C:/ThesisMeasurements/folder_out/262/
[Data Handle Method] //{1=word, 0=char}-{1=boundary, 0=interval}-
{0=QTest, 1=FleissKappa, 2=Krippendorff}
0-0-1
```

APPENDIX G – THE RAT SAMPLE RESULT FILE

```
**********
Interval method is used.
Annotator number is 3.
Feliss' Kappa is applied.
Connective.........: oysa
Test is applied to..: 61 files.
Test is applied to..: 100 annotations.
  Fleiss Value of Arg1............: 0.86060727
      Matrix path..:
C:/ThesisMeasurements/folder_out/262/5_asis_char_interval_annotatorNum_
fleiss_Arg1.dat
  Fleiss Value of Arg2............: 0.9450765
      Matrix path..:
C:/ThesisMeasurements/folder_out/262/5_asis_char_interval_annotatorNum_
fleiss_Arg2.dat
  Fleiss Value of All............: 0.89587396
      Matrix path..:
C:/ThesisMeasurements/folder_out/262/5_asis_char_interval_annotatorNum_
fleiss_BothArg1Arg2.dat
----------


Processed Files..
+++++++++++++++++++++
C:/ThesisMeasurements/ortam1_240710/oysa\00003121_ASC_oysa.xml
C:/ThesisMeasurements/ortam1_240710/oysa\00003221_ASC_oysa.xml
C:/ThesisMeasurements/ortam1_240710/oysa\00005121_ASC_oysa.xml
C:/ThesisMeasurements/ortam1_240710/oysa\00005221_ASC_oysa.xml
**********
C:/ThesisMeasurements/ortam1_240710/oysa\00003121_deniz_oysa.xml
C:/ThesisMeasurements/ortam1_240710/oysa\00003221_deniz_oysa.xml
C:/ThesisMeasurements/ortam1_240710/oysa\00005121_deniz_oysa.xml
C:/ThesisMeasurements/ortam1_240710/oysa\00005221_deniz_oysa.xml
**********
C:/ThesisMeasurements/ortam1_240710/oysa\00003121_idemirsahin_oysa.xml
C:/ThesisMeasurements/ortam1_240710/oysa\00003221_idemirsahin_oysa.xml
C:/ThesisMeasurements/ortam1_240710/oysa\00005121_idemirsahin_oysa.xml
C:/ThesisMeasurements/ortam1_240710/oysa\00005221_idemirsahin_oysa.xml
**********
+++++++++++++++++++++
```