# Description of Turkish Paraphrase Corpus Structure and Generation Method

Bahar Karaoglan[1], Tarık Kışla[1](✉), and Senem Kumova Metin[2]

[1] Ege University, İzmir, Turkey
{bahar.karaoglan, tarik.kisla}@ege.edu.tr
[2] Izmir University of Economics, İzmir, Turkey
senem.kumova@ieu.edu.tr

**Abstract.** Because developing a corpus requires a long time and lots of human effort, it is desirable to make it as resourceful as possible: rich in coverage, flexible, multipurpose and expandable. Here we describe the steps we took in the development of Turkish paraphrase corpus, the factors we considered, problems we faced and how we dealt with them. Currently our corpus contains nearly 4000 sentences with the ratio of 60% paraphrase and 40% non-paraphrase sentence pairs. The sentence pairs are annotated at 5-scale: paraphrase, encapsulating, encapsulated, non-paraphrase and opposite. The corpus is formulated in a database structure integrated with Turkish dictionary. The sources we used till now are news texts from Bilcon 2005 corpus, a set of professionally translated sentence pairs from MSRP corpus, multiple Turkish translations from different languages that are involved in Tatoeba corpus and user generated paraphrases.

**Keywords:** Turkish · Paraphrase · Corpus generation

## 1 Introduction

Corpora are the fundamental elements in the development and/or testing of the studies in the fields of Natural Language Processing, Information Retrieval and Computational Linguistics. Building a corpus is much more than putting bunch of texts together. Many things are needed to be decided and done. Some of which are: Sources of the texts and the size of the corpus, the structure to store the texts (e.g. plain texts, html, database), the fields to tag, the metrics to assess the quality of the corpus.

In this paper we present our efforts in developing a Turkish paraphrase corpus, PARDER, hoping to serve for studies in machine translation, summarization, language generation, automatic assessment of answers to essay type questions and plagiarism detection. We first give relevant work done in other languages and Turkish then, describe each of the above points within the context of our studies in the following sections.

A. Gelbukh (Ed.): CICLing 2016, LNCS 9623, pp. 208–217, 2018.
https://doi.org/10.1007/978-3-319-75477-2_13

# 2 Relevant Work

The literature of paraphrase studies cover numerous corpora that are constructed based on a variety of methods and are holding various features. This variety complicates the classification and the comparison of the paraphrase corpora.

In this study, we will exemplify the notion of paraphrase corpus and paraphrase corpus construction methods in literature by the use of paraphrase corpora listed in Table 1. Henceforth, the corpora in Table 1 will be mentioned by regarding abbreviations given in the second column of the table.

**Table 1.** Paraphrase Corpora

| Corpus | Abbreviation |
|---|---|
| Microsoft Research Paraphrase [1] | MSRP |
| User Language Paraphrase [2] | ULP |
| Question Paraphrase [3] | QP |
| SIMILAR [4] | SIMILAR |
| Regneri & Wang [5] | R&W |
| WiCoPaCo [6, 7] | WiCoPaCo |
| Question Corpus [8] | QC |
| FAQFinder [9] | FAQ |
| Turkish Paraphrase [10] | TP |

The paraphrase corpora and different corpus construction methods will be examined and compared based on the following features (if available in the previous studies): text sources of the corpus, pre-processing of the source data, identification of candidate pairs, the annotation of paraphrase/non-paraphrase pairs, the corpus size.

## 2.1 Text Sources of the Corpus

The paraphrase corpora may be built using different sources such as comparable texts (e.g. similar news texts from different news papers), parallel texts (e.g. answers to the same question) or text corrections (e.g. revisions in Wikipedia articles). The data collected from different sources are parsed into paraphrase units that may be a sentence, paragraph or a collection of words. Following, the candidates are selected from the source data units.

MSRP can be considered as the first major public paraphrase corpus. The candidate sentence pairs are obtained from web-sources. This corpus is not only served for numerous studies but also served as a data source for other paraphrase corpora (e.g. SIMILAR). Sentence is the paraphrase unit in ULP as in the MSRP corpus. ULP corpus involves the sentence pairs that are collected by iSTART system. iSTART system is defined to be an educative support system where the students generate paraphrases to a given set of target sentences to improve their linguistic abilities. The sentence pairs in ULP are the student and target sentence pairs. In QP corpus, question-answer sentence pairs from WikiAnswers are used. The paraphrasing pairs are selected from different questions that are directed to the same answer. R&W corpus is compiled from the subtitles of the TV series House MD. The paraphrasing unit is the sentence, and the corpus involves 14735 sentences from 160 documents. WiCoPaCo corpus is built using the revision logs of Wikipedia. In Wikipedia, the users may add a new content (record) to the system and/or may correct an existing record to improve the quality. In the construction of WiCoPaCo corpus, the sentences that are retrieved from different revisions of the same record have been accepted as paraphrasing candidates. Similar to the WiCoPaCo corpus, the data source of the QC corpus is the log-archive of an online encyclopaedia, Encarta, which includes both the queries and the answers. The paraphrasing unit is again sentence in QC corpus.

A web based-question answering system FAQFinder provides the source data of FAQ corpus. In FAQFinder system, the questions of the users are replied by an answer of a previous similar question. The system involves over 600 files of frequently asked questions.

TP corpus is the first developed Turkish paraphrase corpus that is drawn from four different sources: 1. Two different Turkish translations of Ernest Hemingway's "For Whom the Bell Tolls", 2. Two different subtitles of the film: "The silence of the Lambs", 3. Turkish-English sentence pairs used for machine translation, 4. Paraphrased news sentences. The corpus contains only paraphrased sentences annotated with word and phrase alignments.

## 2.2 Pre-processing of the Source Data

In paraphrase corpus construction, the source data is commonly pre-processed. The pre-processing involves tasks such as spelling correction, stop word removal and ignoring some parts of source texts. For example, two types of corrections: removal of multiple spaces between tokens and appending the full stop character when there isn't one at the end of the sentence; are performed on the source data of ULP. In QP corpus, following the removal of stop words, TreeTagger and Porter stemmer are used in determination of the roots and the stems of the words in source data, respectively. R&W corpus involves the pre-processing tasks such as spelling correction, POS tagging, named entity recognition, and co-reference resolution. In TP corpus, a tool that is developed by the researchers entails the source texts.

## 2.3 Identification of Candidate Pairs

The pairs in paraphrase corpora may be determined randomly or using a procedure by researchers. The most comprehensive study on identification of pairs is presented in the construction of MSRP corpus. The identification procedure in MSRP corpus has two steps. Firstly, the source data is filtered by two different criteria sets. Secondly, a support vector machine (SVM) is employed in classification of the filtered pairs. The pairs that are classified as paraphrase by SVM are accepted as candidates of the corpus. This method increases the amount of true paraphrase pairs in MSRP corpus.

In R&W, WiCoPaCo, QC and FAQ corpora, several procedures such as ordering the events in source texts, longest common subsequence filtering, removal of short sentences, are performed. On the other hand, in ULP, QP, SIMILAR and TP corpora, the candidate pairs are selected randomly.

### 2.4 The Annotation of Paraphrase/Non-paraphrase Pairs

The pairs in a corpus may be annotated in binary mode or within a predefined interval by human annotators. In MSRP corpus, sentence pairs are annotated in binary mode as paraphrase or non-paraphrase as it is common in many other studies (e.g. QP corpus). In ULP corpus, the degree of paraphrasing is annotated within 1–6 interval. Moreover, the quality of user-generated paraphrases is described considering 10 dimensions (garbage, frozen expression, irrelevant, elaboration, writing quality, semantic similarity, lexical similarity, entailment, syntactic similarity, paraphrase quality) of paraphrasing. In R&W corpus, the annotation covers four intervals corresponding to paraphrase, containment, backwards containment, unrelated or invalid tags.

The other important issues in annotation of pairs are the annotation units, the number of annotators and the method to measure annotator agreement. For example, in MSRP, ULP, QP and TP corpus the annotation unit is sentence. Though the annotation unit is considered as the word in SIMILAR corpus. In construction of most of the corpora (e.g. MSRP, R&W, TP), two annotators are employed in classification of pairs as paraphrase or non-paraphrase and one other annotator is employed to resolve the conflictions. In annotation of WiCoPaCo and SIMILAR corpora, the four researchers of the study, a group of six students are assigned respectively. The annotator agreement is given as 63% in SIMILAR corpus.

The average annotator agreement for WiCoPaCo is calculated based on four different evaluation criteria. The highest average agreement value in presented as 0.65 on the semantic differences criterion and the average kappa value f all criteria is given as 0.62. In R&W and TP corpus, the kappa values are reported as 0.55 and 0.416.

### 2.5 The Size of Corpus

MSRP corpus contains about 5801 sentence pairs where 67% of corpus is annotated as paraphrase. SIMILAR corpus includes 700 pairs from MSRP corpus where the number of paraphrase and non-paraphrase pairs is balanced. The total number of sentence pairs that are annotated in 6 scales is 1998 in ULP corpus.

QP corpus includes 7434 pairs in which there are 1000 question sentences and their corresponding paraphrases. R&W corpus is built from 200 millions of candidate pairs. 1992 of pairs in corpus are annotated as gold standard where 158 pairs are paraphrases, 238 are containments and 194 are tagged as related. 1402 of pairs in R&W corpus is accepted to be unrelated. In WiCoPaCo corpus, 200 pairs are paraphrases and 200 are non-paraphrases. QC corpus contains manually annotated 67379 pairs in which 65750 of pairs are paraphrases and 1629 are non-paraphrases. FAQ and TP corpora involve 679, 1270 paraphrase pairs respectively.

# 3 PARDER Corpus

In this section, the sources for building the PARDER corpus, the structure of the corpus database, the annotation scheme will be introduced respectively.

## 3.1 Corpus Sources

The sources for the sentence pairs to be included in the corpus are: Bilcon2005 [11] corpus, translated sentence pairs from MSRP corpus [12, 13], Tatoeba corpus [14], and human generated paraphrase sentences.

Bilcon2005 Turkish news corpus contains 209.305 news, which are collected from five different Turkish news web sources throughout the year 2005. In our study, news texts from Bilcon2005 are parsed into sentences, normalized, and short sentences with less than 3 words and duplicates are removed. For each topic, we then calculated the distance of each sentence to all other sentences in the same topic with 3 different distance metrics: Chebyshev, correlation and Euclid. For each sentence, we selected two sentences with the least distance calculated by each metric as the paraphrase candidates to be marked by the human annotators via a user interface with five marking options: paraphrase, encapsulated, encapsulating, opposite, not-paraphrase. In the user interface, the target statement (sentence) is shown on top of the screen and three annotators labeled each candidate sentence in the list with a label provided via pull down menu.

MSRP corpus is the other source employed in building PARDER. A set of randomly chosen 2000 sentence pairs, which are 60% paraphrase and 40% not paraphrase, from MSRP corpus are translated by a professional. These translated sentences are re-labeled by human annotators considering the fact that translations may not be in parallel with the original labels.

Tatoeba corpus is referred as a multilingual sentence dictionary consisting of cross language translations of sentences between language pairs. In this study, the multiple translations of English sentences to Turkish are utilized. Most of these sentences were very short and different translations of the same English sentence varied by only one word. After eliminating sentences less than 5 words and multiple translations that vary with one word, we obtained 114 paraphrase sentence pairs for PARDER.

Volunteering people, researchers and Turkish Language Education students are provided with a list of sentences from which they can choose to rephrase. Currently there are 2419 Sentence pairs labeled as paraphrase and 1602 labeled as non-paraphrase by the annotators (Table 2).

**Table 2.** Content Summary of The Corpus

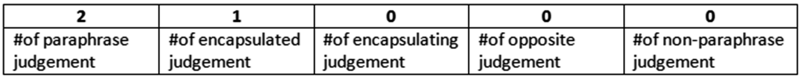
| Source | Paraphrase | Non-paraphrase |
|---|---|---|
| Bilcon | 1.005 | 802 |
| Tatoeba | 114 | – |
| MSRP-translated | 1.200 | 800 |
| User-generated | 100 | – |
| Total | 2.419 | 1.602 |

### 3.2 Corpus Database Structure

The corpus is created on a database structure to make the labeling as flexible and informative as possible. The database consists of 7 tables: *Documents*, *Sentences*, *Similarity*, *Words*, *Dictionary*, *Word-Relation* and *Word-Meaning* table.

*Documents Table* keeps the physical location (path), type (e.g. news, translation, user generated, etc.), author, source of the document in which the sentence appears.

*Similarity Table* stores similarity and type information for each sentence pair in the corpus keeping the overall similarity score of the sentence pairs. Similarity column contains a code that represents the number of similarity values assigned by the annotators. For example, the value 21000 of similarity column, as given in Fig. 1, means that two annotators marked the regarding sentence pair as paraphrase; one has marked them as encapsulated. Type column stores the binary decision that denotes if the sentence pair is paraphrase (1) or not (0).



| 2 | 1 | 0 | 0 | 0 |
|---|---|---|---|---|
| #of paraphrase judgement | #of encapsulated judgement | #of encapsulating judgement | #of opposite judgement | #of non-paraphrase judgement |

**Fig. 1.** The structure of similarity value (in similarity table).

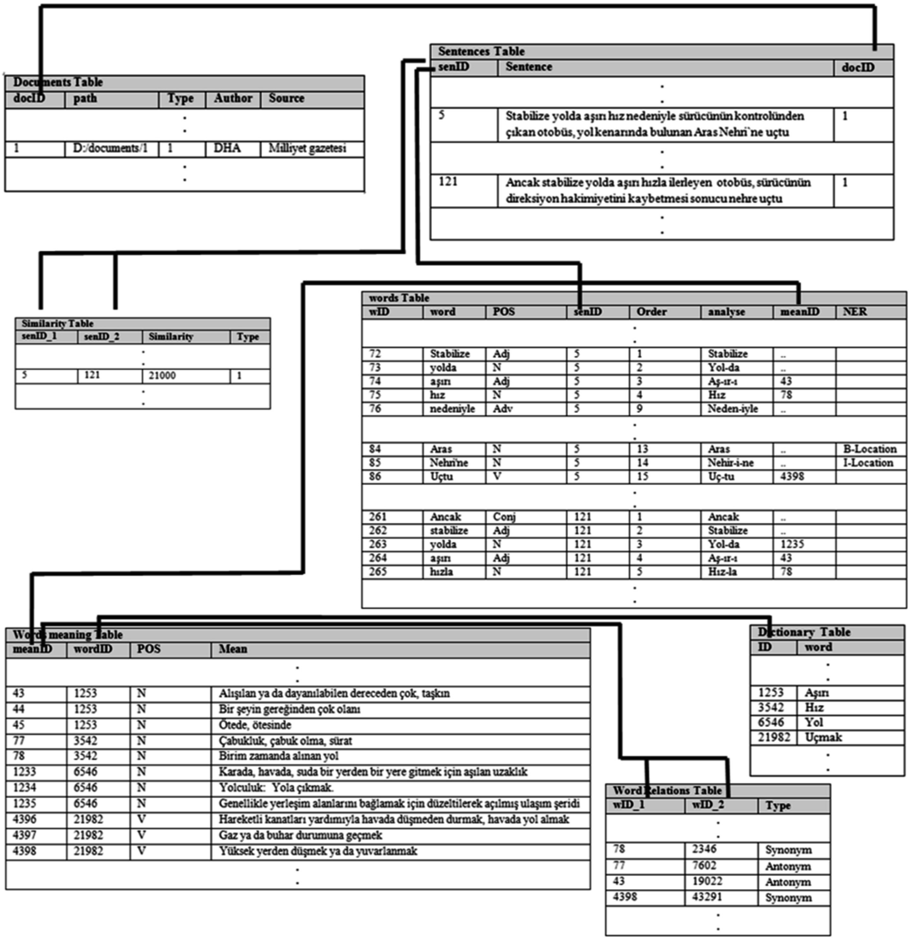
*Words Table* has an entry for each word in each sentence holding part of speech tag (POS), morphological analysis, named entity (NE) tag and its position within the sentence. This table is related to the *Word-Meaning Table* enabling the detection of synonym, antonym and meaning relations between the words.

*Dictionary Table* keeps the words as they appear in the dictionary and their ids. For the fact that words may have more than one meaning another table, *Word-Meaning Table*, is held to keep the meanings attached to part of speech.

*Word-Relation Table* has an entry for each word in the database to keep synonyms, antonyms and etc. Figure 2 shows snapshot of the database for two sentences drawn from Document #1 of Milliyet newspaper with ids #5 and #121 that are tagged as paraphrases. Sentences #5 and #121 together with their English translations are:

*#5: Stabilize yolda aşırı hız nedeniyle sürücünün kontolünden çıkan otobüs, yol kenarında bulunan Aras Nehri'ne uçtu. (Eng: Due to excessive speed on the stabilized road, getting out of control of the driver, the bus flew into Aras River, which is running on the side of the road.)*

*#121: Ancak, stabilize yolda aşırı hızla ilerleyen otobüs, sürücünün direksiyon hakimiyetin kaybetmesi sonucunda nehre uçtu. (Eng: However, over speeding bus on the stabilized road, as a result of loosing control of the wheel by the driver flew into the river.)*



Documents Table

| docID | path | Type | Author | Source |
|---|---|---|---|---|
| | | ⋮ | | |
| 1 | D:/documents/1 | 1 | DHA | Milliyet gazetesi |
| | | ⋮ | | |

Sentences Table

| senID | Sentence | docID |
|---|---|---|
| | ⋮ | |
| 5 | Stabilize yolda aşırı hız nedeniyle sürücünün kontrolünden çıkan otobüs, yol kenarında bulunan Aras Nehri`ne uçtu | 1 |
| | ⋮ | |
| 121 | Ancak stabilize yolda aşırı hızla ilerleyen otobüs, sürücünün direksiyon hakimiyetini kaybetmesi sonucu nehre uçtu | 1 |
| | ⋮ | |

Similarity Table

| senID_1 | senID_2 | Similarity | Type |
|---|---|---|---|
| | | ⋮ | |
| 5 | 121 | 21000 | 1 |
| | | ⋮ | |

words Table

| wID | word | POS | senID | Order | analyse | meanID | NER |
|---|---|---|---|---|---|---|---|
| | | | | ⋮ | | | |
| 72 | Stabilize | Adj | 5 | 1 | Stabilize | .. | |
| 73 | yolda | N | 5 | 2 | Yol-da | .. | |
| 74 | aşırı | Adj | 5 | 3 | Aş-ır-ı | 43 | |
| 75 | hız | N | 5 | 4 | Hız | 78 | |
| 76 | nedeniyle | Adv | 5 | 9 | Neden-iyle | .. | |
| | | | | ⋮ | | | |
| 84 | Aras | N | 5 | 13 | Aras | .. | B-Location |
| 85 | Nehri'ne | N | 5 | 14 | Nehir-i-ne | .. | I-Location |
| 86 | Uçtu | V | 5 | 15 | Uç-tu | 4398 | |
| | | | | ⋮ | | | |
| 261 | Ancak | Conj | 121 | 1 | Ancak | .. | |
| 262 | stabilize | Adj | 121 | 2 | Stabilize | .. | |
| 263 | yolda | N | 121 | 3 | Yol-da | 1235 | |
| 264 | aşırı | Adj | 121 | 4 | Aş-ır-ı | 43 | |
| 265 | hızla | N | 121 | 5 | Hız-la | 78 | |
| | | | | ⋮ | | | |

Words meaning Table

| meanID | wordID | POS | Mean |
|---|---|---|---|
| | | | ⋮ |
| 43 | 1253 | N | Alışılan ya da dayanılabilen dereceden çok, taşkın |
| 44 | 1253 | N | Bir şeyin gereğinden çok olanı |
| 45 | 1253 | N | Ötede, ötesinde |
| 77 | 3542 | N | Çabukluk, çabuk olma, sürat |
| 78 | 3542 | N | Birim zamanda alınan yol |
| 1233 | 6546 | N | Karada, havada, suda bir yerden bir yere gitmek için aşılan uzaklık |
| 1234 | 6546 | N | Yolculuk: Yola çıkmak. |
| 1235 | 6546 | N | Genellikle yerleşim alanlarını bağlamak için düzeltilerek açılmış ulaşım şeridi |
| 4396 | 21982 | V | Hareketli kanatları yardımıyla havada düşmeden durmak, havada yol almak |
| 4397 | 21982 | V | Gaz ya da buhar durumuna geçmek |
| 4398 | 21982 | V | Yüksek yerden düşmek ya da yuvarlanmak |
| | | | ⋮ |

Dictionary Table

| ID | word |
|---|---|
| | ⋮ |
| 1253 | Aşırı |
| 3542 | Hız |
| 6546 | Yol |
| 21982 | Uçmak |
| | ⋮ |

Word Relations Table

| wID_1 | wID_2 | Type |
|---|---|---|
| | ⋮ | |
| 78 | 2346 | Synonym |
| 77 | 7602 | Antonym |
| 43 | 19022 | Antonym |
| 4398 | 43291 | Synonym |
| | ⋮ | |

**Fig. 2.** Sentence pair structure in corpus database.

### 3.3 Annotation Scheme

The sentence pairs drawn from different sources are all normalized and transferred to the database of the annotation software developed by the researchers. Three human annotators tagged the sentence pairs on pentad scale (paraphrase, encapsulating, encapsulated, non-paraphrase and opposite). This analysis is also done on binary-scaled (paraphrase/non-paraphrase) judgment. For binary scaled judgment, those sentence pairs with similarity score in the similarity table, greater than a predefined threshold score 12 are considered as paraphrase and the rest as non-paraphrase. Similarity score is calculated using similarity value. For example, if we assume that similarity value is 21000, similarity score will be 14 ($\mathbf{2} \times 5 + \mathbf{1} \times 4 + \mathbf{0} \times 3 + \mathbf{0} \times 2 + \mathbf{0} \times 1$). The reliability of the agreement between the annotators is assessed by Fleiss Kappa values [15] that are given in Table 3.

**Table 3.** The results of the Fleiss Kappa Analysis

| Scale | kappa | $SE_{fleiss}$* | z | $CI_{lower}$ | $CI_{upper}$ | p |
|---|---|---|---|---|---|---|
| Binary Scaled | 0.634 | 0.004 | 148.11 | 0.626 | 0.642 | 0.00 |
| Pentad Scaled | 0.671 | 0.003 | 228.30 | 0.665 | 0.667 | 0.00 |

*Standard error (SE) values are calculated using formula given in [16].

## 4 The Analysis of PARDER

PARDER corpus is analyzed considering the averaged values of sentence-pair based attributes: number of the words in sentence (*SL*), the ratio of common words (*MW*), the ratio of sentence lengths (*LS*), the ratio of common consequent sets (*MB*) [17] and the ratio of sequencing (*OW*) [17]. Table 4 gives the overall analysis results of PARDER corpus (P: Paraphrase NP: non-Paraphrase). In the analysis, the attribute values are obtained individually for each set of sentence pairs that are extracted from three different sources.

**Table 4.** The analysis results of PARDER corpus using syntactic features

| Attributes | Sources | | | | | |
|---|---|---|---|---|---|---|
| | Bilcon2005 | | MSRP | | Tatoeba | User-generated |
| | P | NP | P | NP | P | P |
| *SL* | 18.8 | 17.5 | 17 | 15.5 | 7 | 13.9 |
| *LS* | 0.72 | 0.62 | 0.85 | 0.79 | 0,88 | 0,83 |
| *MW* | 0.34 | 0.15 | 0.55 | 0.33 | 0.36 | 0,31 |
| *MB* | 0.15 | 0.03 | 0.31 | 0.14 | 0.13 | 0.24 |
| *OW* | 0.29 | 0.11 | 0.50 | 0.29 | 0.32 | 0.37 |

The semantic attributes that we consider are the difference in the polarity (*DP*) (positive/negative) and tenses between the sentences (*DT*). Table 5 shows the statistics related to the mentioned attributes.

**Table 5.** The analysis results of PARDER corpus using semantic features

| Attributes | Sources | | | | | |
|---|---|---|---|---|---|---|
| | Bilcon2005 | | MSRP | | Tatoeba | User-generated |
| | P | NP | P | NP | P | P |
| *DP* | 34% | 80% | 20% | 60% | 21% | 12% |
| *DT* | 7% | 14% | 6% | 11% | 5% | 8% |

## 5 Conclusion

In this paper, we have presented Turkish paraphrase corpus, PARDER, and described the corpus construction steps in our on-going project. The data sources, candidate selection procedure, data structure and annotation scheme of the PARDER corpus are introduced. The corpus currently contains nearly 4000 sentences with the ratio of 60% paraphrase and 40% non-paraphrase sentence pairs.

PARDER corpus is built to serve for many purposes in the field of language processing. We aim to increase the corpus size to 6000 pairs and make it accessible on web for researchers in future.

**Acknowledgement.** This work is carried under the grant of TÜBİTAK – The Scientific and Technological Research Council of Turkey to Project No: 114E126, Using Certainty Factor Approach and Creating Paraphrase Corpus for Measuring Similarity of Short Turkish Texts and Ege University Scientific Research Council Project No 2015/BİL/034, Developing a Paraphrase Corpus for Turkish Short Text Similarity Studies.

## References

1. Dolan, B., Quirk C., and Brockett C.: Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In: Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics (2004)
2. McCarthy, P.M., McNamara, D.: The user-language paraphrase challenge. In: Special ANLP Topic of the 22nd International Florida Artificial Intelligence Research Society Conference, Florida (2008)
3. Bernhard, D., Gurevych, I.: Answering learners' questions by retrieving question paraphrases from social Q&A sites. In Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications, pp. 44–52. Association for Computational Linguistics, Stroudsburg (2009)
4. Rus, V., Lintean, M., Moldovan, C., Baggett, W., Niraula, N., Morgan, B.: SIMILAR Corpus: a resource to foster the qualitative understanding of semantic similarity of texts. In: LREC, pp. 50–59 (2012)
5. Regneri, M., Wang, R.: Using discourse information for paraphrase extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 916–927 (2012)
6. Max, A., Wisniewski, G.: Mining naturally-occurring corrections and paraphrases from Wikipedia's Revision History. In: LREC (2010)
7. Dutrey, C., Bouamor, H., Bernhard, D., Max, A.: Local modifications and paraphrases in Wikipedia's revision history. Procesamiento del Lenguaje Natural **46**, 51–58 (2010)
8. Zhao, S., Zhou, M., Liu, T.: Learning question paraphrases for QA from encarta logs. In: IJCAI (2007)
9. Lytinen, S., Tomuro, N.: The use of question types to match questions in FAQFinder. AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases (2002)
10. Demir, S., El-Kahlout, I.D., Unal, E., Kaya, H.: Turkish paraphrase corpus. In: LREC, pp. 4087–4091 (2012)

11. Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H.C., Uyar, E.: New event detection and topic tracking in Turkish. J. Am. Soc. Inform. Sci. Technol. **61**(4), 802–819 (2010)
12. Dolan, W., Brockett, C.: Automatically Constructing a Corpus of Sentential Paraphrases. In Third International Workshop on Paraphrasing (2005)
13. Brockett, C., Dolan, W.: Support vector machines for paraphrase identification and corpus construction. In: Third International Workshop on Paraphrasing (IWP2005) (2005)
14. Tiedemann J.: Parallel data, tools and interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC) (2012)
15. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bull. **76**(5), 378–382 (1971)
16. Fleiss, J.L., Nee, J.C., Landis, J.R.: Large sample variance of kappa in the case of different sets of raters. Psychological Bull. **86**(5), 974–977 (1979)
17. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Discov. Data (TKDD), **2**(2), Article 10, 25 pages (2008)