A Case Study Towards Turkish Paraphrase Alignment

Seniz Demir **İlknur Durgar El-Kahlout** **Erdem Unal**
TUBITAK-BILGEM
Gebze, Kocaeli, TURKEY
{seniz.demir,ilknur.durgar,erdem.unal}@tubitak.gov.tr

Abstract

Paraphrasing is expressing the same semantic content using different linguistic means. Although previous work has addressed linguistic variations at different levels of language, paraphrasing in Turkish has not been yet thoroughly studied. This paper presents the first study towards Turkish paraphrase alignment. We perform an analysis of different types of paraphrases on a modest Turkish paraphrase corpus and present preliminary results on that analysis from different standpoints. We also explore the impact of human interpretation of paraphrasing on the alignment of paraphrase sentence pairs.

1 Introduction

Paraphrases are alternative linguistic expressions that convey the same content. Natural languages allow linguistic variations at different levels (e.g., lexical and phrasal) and a change at a level of language may trigger other changes at different levels. Paraphrasing has attracted a growing interest from the research community in a broad range of tasks such as language generation (Power and Scott, 2005), machine translation (Callison-Burch et al., 2006), and question answering (France et al., 2003). Moreover, research on acquisition (Max et al., 2012), generation (Zhao et al., 2010), and recognition (Qiu et al., 2006) of paraphrases has been on the rise for the last decade. Paraphrasing is also an increasingly studied problem by the generation community. One particular text-to-text generation problem being addressed is the generation of sentence-level paraphrases by converting a sentence into a new one with approximately the same meaning (Wubben et al., 2010).

One aspect of paraphrasing is the specification of paraphrase types via a typology. Building paraphrase typologies from different perspectives (e.g., linguistics analysis and discourse analysis) has been an active research area for a number of years now (Vila et al., 2011). In particular, linguistic grounds govern the typologies built by language processing systems (Kozlowski et al., 2003) which are often very generic or system specific.

Research on paraphrase alignment focuses on identifying links between semantically related word strings. Such monolingual alignments can be later used as training data for several natural language processing approaches (e.g., textual entailment and multidocument summarization) (Thadani et al., 2012). Although a wealth amount of research has studied various problems related to Turkish, we here focus on a problem which has not been studied earlier. We present our initial explorations on Turkish paraphrase alignment by considering how alignment is affected by human interpretation of paraphrasing. We conducted a study on a modest corpus from four different sources to investigate answers to the following questions: i) What are the types of paraphrases that can be observed at different levels of Turkish? ii) Do humans agree on the existence of paraphrasing between Turkish paraphrase sentences? iii) How does human interpretation of paraphrasing affect the alignment of paraphrase sentences?

Our study is unique in that it presents a generic typology of paraphrase types found in our Turkish paraphrase corpus and discusses the agreement of human annotators on the identification and classification of observed correspondences between paraphrases. This study also presents our aggregated observations on the relation between interpretation and alignment of paraphrase casts.

Proceedings of the 14th European Workshop on Natural Language Generation, pages 188–192,
Sofia, Bulgaria, August 8-9 2013.

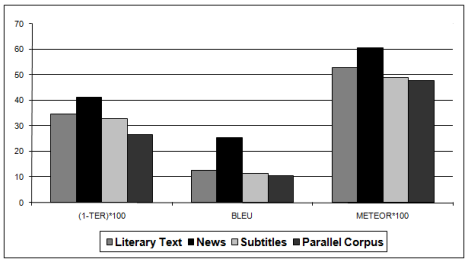


Figure 1: Sentence similarity scores of the corpus.

	Literary Text	News Articles	Subtitles	Parallel Corpus
# Tokens	1879	3379	1632	1581
# Unique Tokens	811	1473	824	609
# Shared Tokens	519	1125	402	354
Lexical Overlap	72.5	82.9	63.2	62.7
Lexical Overlap (lem. cont. words)	68.4	67.2	48.6	45

Table 1: Characteristics of the selected 400 pairs.

2 Paraphrase Corpora

The Turkish paraphrase corpus (Demir et al., 2012) comprises 1270 paraphrase pairs from four different sources: i) translations of a literary text, ii) multiple reference translations of English tourism-related sentences, iii) news articles, and iv) subtitles of a movie. We measured sentence similarities of all paraphrase pairs from each domain via three measures typically used in statistical machine translation evaluations: TER (Matthew Snover, 2006), BLEU (Papineni et al., 2002), and METEOR (Lavie and Agarwal, 2007). As shown in Figure 1, the ordering of domains with respect to all metrics are the same where the pairs from the news domain and those from the parallel corpus are the most and the least similar pairs respectively. Since there are divergences across different domains, we randomly drew from each domain an equal number of sentences (i.e., 100 paraphrase pairs[1]). Some characteristic features of the paraphrase pairs selected for this study are shown in Table 1.

3 Paraphrase Typology

To our best knowledge, a Turkish paraphrase typology that we can apply to this study does not yet exist in the literature. On the other hand, building a comprehensive typology is not one of our objectives. There are a number of available typologies built for English (Dras, 1999; Vila et al., 2011). Since our focus in this work is on characterizing paraphrasing at different levels of language, we greatly drew from the linguistically-motivated typology by Vila et al. (2011) while building our generic typology. We examined the selected 400 paraphrase pairs and constructed a typology that covers all paraphrases occurring within these pairs. Our typology covers three levels of language and consists of four classes.

The **lexical class** covers all changes that arise from exchanging words within a phrase with other words and includes four subclasses (i.e., substitution, substitution with opposite polarity, deletion, and pronominalization)[2]:
(1) “Su bize takip edebileceğimiz hiçbir$_1$ iz$_1$ bırakmıyor$_1$.” (Water leaves$_1$ no$_1$ trace$_1$ that we can follow.)
(2) “Su olayın takip edilebilecek bütün$_1$ izlerini$_1$ yok$_1$ ediyor$_1$.” (The water destroys$_1$ all$_1$ traces$_1$ of the event that can be followed.)

The **morphological class** covers inflectional and derivational changes within words and includes two subclasses (i.e., inflectional changes and derivational changes):
(1) “Böyle bir ilaç almaktansa hasta$_1$ kalmak$_1$ iyidir.” (Staying$_1$ sick$_1$ is better than taking such a drug.)
(2) “Hasta$_1$ kalırım$_1$ da yine de bu ilacı içmem.”(I$_1$ stay$_1$ sick$_1$ still I don’t take this drug.)

The **phrasal class** includes changes that arise from exchanging fragments with same meaning:
1) “Bunları biliyorum fakat emri ben$_1$ vermedim$_1$.” (I know all that, but I$_1$ did$_1$ not$_1$ give$_1$ the order.)
(2) “Bunları biliyorum ama, emri veren$_1$ ben$_1$ değilim$_1$.” (I know all that, but I’m$_1$ not$_1$ the$_1$ one$_1$ who$_1$ gave$_1$ the order.)

The **other class** is for all other changes that imply different lexicalizations for the same contextual meaning:
(1) “Savaş çıkınca pek çok çingene eskilerdeki gibi kötü$_1$ kişiler$_1$ oldular$_1$.” (When war broke out, many gypsies became$_1$ just$_1$ as bad$_1$ people$_1$ as those of the past.)
(2) “Savaşta birçok çingene eskiden olduğu gibi yine çok$_1$ kötülük$_1$ yaptılar$_1$.” (Many gypsies did$_1$ much$_1$ evil$_1$ in the war again as in the past.)

[1]The number of paraphrase pairs in the subtitle domain limits the study to 100 pairs from each domain.

[2]Each word in a paraphrase cast receives the same subscript.

Although these classes are language independent, they include several Turkish specific aspects such as morphophonemic processes. For instance, Turkish word changes due to vowel harmony, vowel drops, and consonant drops/changes are all covered by the morphological class.

4 Paraphrase Alignment

While manually aligning the paraphrase sentence pairs, our goal was to jointly identify the paraphrase casts (i.e., the substitutable word strings) and specify the types correspondences between them. We asked three native speakers to align the selected paraphrase sentences by aligning word strings[3] as much as possible and marking the strength of observed correspondences as either "certain" (the correspondences that hold in any context) or "possible" (the correspondences that are context-specific). The annotators were also told to assign each identified correspondence between paraphrase casts to one of the classes in our typology. In cases where the same word strings were aligned, the correspondence was not classified with a class from the typology. Before aligning the corpus, the annotators were trained on a different set of paraphrases using an annotation guideline. Table 2 reports some statistics of the alignment process. The column labelled as "Common" represents the alignments common to all annotators. The rows labelled as "C", "P", and "U" represent the number of certain and possible alignments, and the number of unaligned words respectively[4]. It is noteworthy that the percentage of common certain alignments is significantly higher than the percentage of common possible alignments in all domains.

5 Corpus Study Findings

In this study, we aim to explore whether humans agree on the existence (i.e., identifying two word strings as paraphrases) and type of paraphrasing between Turkish paraphrase sentences. We are also interested in how the alignment of paraphrase casts is affected from human interpretation of paraphrasing between Turkish paraphrase sentences. Please note that the alignment of paraphrase casts consequently affects the sentence alignment of paraphrase sentence pairs.

[3]A word string consists of one or more words which may not be contiguous. Two word strings are aligned when one or more words in one string are paired with one or more words in the other string.

[4]Please note that these scores represent all alignments including the alignments of the same word strings.

Domain		Ant.1	Ant. 2	Ant. 3	Common
Literary Text	C	647	639	578	376 (58%)
	P	88	121	178	10 (5.62%)
	U	165	140	144	101 (61.2%)
News Articles	C	1384	1330	1259	988 (71.4%)
	P	53	186	214	3 (1.4%)
	U	203	124	167	102 (50.2%)
Subtitles	C	578	546	530	306 (52.9%)
	P	101	112	119	13 (10.9%)
	U	104	122	131	71 (54.2%)
Parallel Corpus	C	565	531	542	313 (55.4%)
	P	109	126	70	6 (4.8%)
	U	112	129	174	80 (45.9%)

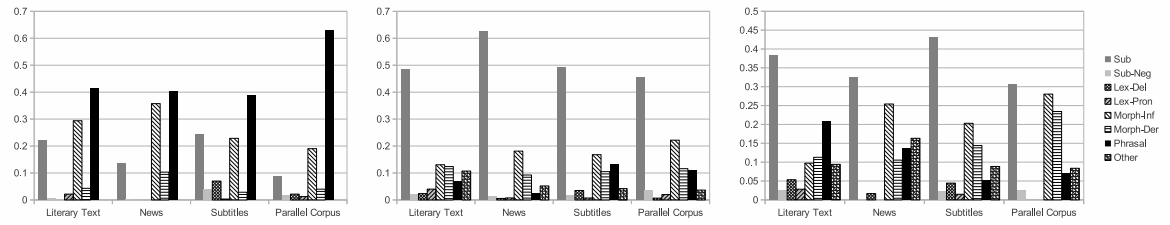
Table 2: Alignment statistics of paraphrase pairs.

Our analysis started with examining how often our annotators agreed on identifying paraphrasing between two word strings. The agreement scores in Table 3 show that the annotators (pairwise) had a reasonable level of agreement in all domains. In majority of these cases, the annotators also agreed on the strength of the correspondence (i.e., both annotators either classified the correspondence as "Certain-Certain" or "Possible-Possible").

Domain	Ant.1&2	Ant. 1&3	Ant. 2&3
Literary Text	0.78%	0.73%	0.77%
News Articles	0.81%	0.86%	0.81%
Subtitles	0.68%	0.72%	0.86%
Parallel Corpus	0.59%	0.61%	0.78%

Table 3: Agreement on paraphrase identification.

The agreement scores in Table 3 show the agreement of annotators on the fact that two word strings are paraphrases and thus should be aligned. But it does not mean that the reason behind similar identifications is the same. We thus explored whether the annotators similarly classified the word strings that they identified as paraphrases. In all domains, the agreement scores between the annotators (given in Table 4) are dramatically lower than the scores in Table 3. It is particularly noteworthy that the smallest drop is observed in the parallel corpus domain (the domain that contains the least similar sentences). In cases where the annotators (pairwise) classified the same word strings with the same paraphrase class, they had a high agreement (between 78% and 91%) on the strength of the correspondence in all domains. We also computed the inter-annotator agreement via Kappa (Cohen, 1960). Kappa scores (shown bold in Table 4) represent fair to good agreement be-



(a) Annotator 1 (b) Annotator 2 (c) Annotator 3

Figure 2: Distribution of paraphrase classes across domains.

tween the annotators.

Domain	Ant.1&2	Ant. 1&3	Ant. 2&3
Literary Text	0.37% **(0.34)**	0.34% **(0.33)**	0.56% **(0.63)**
News Articles	0.40% **(0.38)**	0.51% **(0.48)**	0.54% **(0.53)**
Subtitles	0.37% **(0.41)**	0.37% **(0.38)**	0.70% **(0.72)**
Parallel Corpus	0.33% **(0.46)**	0.35% **(0.46)**	0.64% **(0.75)**

Table 4: Agreement on paraphrase classes.

Figure 2 presents the distribution of paraphrase classes identified by each annotator across different domains. Notably, the identified paraphrase classes between word strings appear to diverge in several respects. We are currently exploring the reason behind this poor annotator agreement on paraphrase classes. One possible reason might be different understanding of the typology.

As a second step, we explored the impact of different interpretations of paraphrasing between sentence pairs on the alignment of these sentences. We analyzed the alignment differences of sentences and classified them into four classes:

- **Different Classification:** Although both annotators identify the same correspondence between two word strings, they classify that correspondence differently.
- **Missing Alignment:** One annotator identifies an alignment between two word strings but the other annotator does not identify a correspondence between these word strings.
- **Missing Word:** The annotators identify a correspondence of the same paraphrase class between two word strings which differ only in one word.
- **Different Grouping:** Two word strings are identified as having a single correspondence by one annotator whereas a number of disjoint correspondences between these word strings are identified by the other annotator.

All these differences except those classified as "different classification" result in different alignments between word strings. Such different alignments of paraphrase casts then change the alignment of paraphrase sentences.

6 Conclusion and Future Work

In this paper, we present our initial explorations on Turkish paraphrase alignment by exploiting a modest corpus. We built a generic and linguistically grounded Turkish paraphrase typology that covers the types of paraphrases observed in the corpus. In the study, the paraphrases identified by human annotators were aligned and annotated with paraphrase classes from the typology. The agreement of the annotators with respect to the existence and alignment of paraphrases as well as the associated paraphrase classes were reported. The study showed that the way how humans interpret paraphrasing between Turkish paraphrase sentences has an impact on how they align these sentences.

We have two main directions for future research: i) conducting a larger corpus study for drawing generalizations about Turkish paraphrasing and enhancing the typology if necessary, and ii) building Turkish paraphrase applications (e.g., automatic paraphrase acquisition) in correlation with the collected insights. We believe that the current findings for Turkish paraphrase alignment and our corpus enriched with paraphrase types enable future research on paraphrase phenomena in different fields such as language generation, textual entailment, summarization, and machine translation to be empirically assessed.

References

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 17–24.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Seniz Demir, Ilknur Durgar El-Kahlout, Erdem Unal, and Hamza Kaya. 2012. Turkish paraphrase corpus. In *Language Resources and Evaluation Conference - LREC*.

Mark Dras. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University.

Florence Duclaye France, Franois Yvon, and Olivier Collin. 2003. Learning paraphrases to improve a question-answering system. In *EACL Workshop NLP for Question-Answering*.

Raymond Kozlowski, Kathleen F. McCoy, and K. Vijay-Shanker. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *Proceedings of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 1–8.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231.

Richard Schwartz. Matthew Snover, Bonnie J. Dorr. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*.

Aurélien Max, Houda Bouamor, and Anne Vilnat. 2012. Generalizing sub-sentential paraphrase acquisition across original signal type of text pairs. In *EMNLP-CoNLL*, pages 721–731.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Richard Power and Donia Scott. 2005. Automatic generation of large-scale paraphrases. In *IWP*.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua. 2006. Paraphrase recognition via dissimilarity significance classification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 18–26.

Kapil Thadani, Scott Martin, and Michael White. 2012. A joint phrasal and dependency model for paraphrase alignment. In *COLING*, pages 1229–1238.

Marta Vila, M. Antonia Marti, and Horacio Rodriguez. 2011. Paraphrase concept and typology: A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2010. Paraphrase generation as monolingual translation: data and evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG '10, pages 203–207.

Shiqi Zhao, Haifeng Wang, Xiang Lan, and Ting Liu. 2010. Leveraging multiple mt engines for paraphrase generation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1326–1334.