Simplification of Turkish Sentences

Dilara Torunoğlu-Selamet, Tuğba Pamay, Gülşen Eryiğit
Department of Computer Engineering
Istanbul Technical University
Istanbul, 34469, Turkey
[torunoglud, pamay, gulsen.cebiroglu]@itu.edu.tr

***Abstract*—Text Simplification is the process of transforming existing natural language text into a new form aiming to reduce their syntactic or lexical complexities while preserving their meaning. A sentence being long and complicated may pose multiple problems especially for elementary school children. In this paper[1], we focus on Turkish, a morphologically rich language, and examine sentences from an elementary school text book to extract complex structures and propose a sentence simplification system to automatically generate simpler versions of the sentences. Thereby, sentences become easier for children to understand, particularly children with difficulty in reading comprehension. Our system automatically uses simplification operations, namely splitting, dropping, reordering, and substitution.**

***Keywords*—Text Simplification, Sentence Simplification, Turkish**

I. Introduction

Text Simplification is the process of transforming existing natural language text into a new form with aim of reducing their syntactic or lexical complexity while preserving their meaning. Applications of Text Simplification can help people to understand natural text with less effort. The target audience might be people with language disabilities like aphasia, adults learning a foreign language, low-literacy readers [1] and children [2]. Text simplification is also used in areas like Machine Translation (MT) [3] and Text Summarization (TS) [4]. At sentence level, reading difficulties (sentence complexities) lie in the syntactic and lexical levels, so simplification of sentences can be classified into two general categories: Lexical and Syntactical Simplification. Without considering the language level, there are some approaches for lexical and syntactic simplification based on Statistical Machine Translation. The concept of a simple, "easy-to-read" sentence is not universal. Sentence length and syllable count can give a good estimate but it will not be complete since we are taking the preserving of meaning and understandability into account during the simplification process. Also, requirements of "easy-to-read" sentences can vary from audience to audience.

Sentence simplification for highly inflectional or agglutinative languages has significant problems. For example, in Turkish, some words may be omitted from a sentence yet the meaning may remain the same. Elementary school children (preteens) face difficulty in understanding the arguments of the main predicate in the sentence, which may be complicated. Preteens have a tendency to use simple sentence structures in their daily lives, and when they come across complex structured sentences in school text books, they may fall behind in the class. For this reason, in this paper, we focus on Turkish and examine sentences from elementary school textbook to extract complex structures and propose a sentence simplification system to automatically generate simpler versions of the sentences. Thereby, sentences become easier to understand by children, especially ones with difficulty in reading comprehension.

In this paper, we take advantage of inflectional groups in Turkish and investigate certain types of complex structured sentences. We divide these sentences under three main categories as: 1. Coordinate Sentences, 2. Paratactic Sentences, 3. Subordinating Sentences and each main category also has sub-categories. Examples of these categories are explained in Section III in detail. Then, we derive rules corresponding to each category and apply the rules to the sentences which were taken from an elementary school textbook. We prepare a data set which was annotated morphologically and syntactically with the NLP tools [5] to use in the sentence simplification.

The paper is structured as follows: Section II gives brief information about related work, Section III introduces the sentence structures on which we focused and presents our sentence simplification approach and Section IV presents the conclusion and futurework.

II. Related Work

Text simplification has become a highly investigated topic with the increase in the use of NLP systems. These systems suffer lower accuracy results from the complexity of the sentences. One study [6], proposes a sentence simplification model which is based on tree transformation by Statistical Machine Translation (SMT) [7], [8]. This work covers operations like sentence splitting, reordering, deleting (dropping) and phrase/word substitution. The parallel corpora that were used in this work (PWKP) were generated from English Wikipedia and Simple English Wikipedia. Another study [9] presents a data-driven model based on quasi-synchronous grammar. In contrast to state of art solutions [6], operations are not defined explicitly; instead the quasi-synchronous grammar extraction algorithm learns appropriate rules from the training data. In

[1]This work is part of our ongoing research project "A Signing Avatar System for Turkish to Turkish Sign Language Machine Translation" supported by TUBITAK FATIH 1003 (grant no: 114E263).

another study [10] which presents a machine translation based approach similar to [6], differs in that it does not take syntactic information into account and only relies on phrase based machine translation methods to implicitly learn simplifying and paraphrasing of phrases. They claim that they produced a language agnostic solution. However they only worked on lexical operations for sentence simplification. In [11], a lexical approach was followed for sentence simplification for different learning levels and context. Their method has 4 steps: part-of-speech (POS) tagging, synonym probing, context frequency-based lexical replacement and sentence checker. They evaluated their results with human annotators by only asking yes/no questions for testing on meaning and simplicity. They did not use parallel datasets, instead they used context-based books for doing lexical operations. The study [12] focuses on syntactic simplification to make text easier to comprehend for human readers, or process by programs. They formalize the interactions that take place between syntax and discourse during the simplification process and present the results of their system.

Most of the recent works focus on the English yet there are some studies on other languages. The study in [13], focuses on Brazilian Portuguese. Another study [14] which is based on dependency parsing of Spanish sentences is capable of lexical simplification, deletion operations and sentence simplification operations. The study [15] aims to develop an approach to syntactic simplification of French sentences.

Another usage of text simplification is to help children understand complex sentences in books. One of the studies conducted for this purpose is [16] which examines children stories and proposes a text simplification system to automatically generate simplified, more comprehensible versions of the stories for children, especially those with difficulty in reading comprehension. Splitting, dropping, reordering and substitution operations can be done with the proposed system. Another study with the same approach is in [2] which chooses children as the target audience of text simplification operations. They perform both syntactic and lexical simplifications. They follow a rule based system for this task. Inspired by these researches in this paper, we focus on simplifying children's textbooks.

III. Discussion and Approach

The morphologically rich nature of Turkish may result in orthographic words to be split into multiple inflectional groups[2]. For the sentence simplification approach, we take advantage of this issue and investigate solutions for the simplification of syntactically complex sentences. We divide them under three main categories as: 1. Coordinate Sentences, 2. Paratactic Sentences, 3. Subordinating Sentences and each main category also has sub-categories. Then, we derive rules corresponding to each category and applied the rules to the sentences which were taken from the elementary school textbook. To apply the rules over the sentences, we benefit from the morphological and syntactic information of tokens in the sentence and also use a morphological generator [5] which is one of the NLP tools to generate surface form of the token from its morphological analysis. Sentence simplification is executed under three steps which are visualized in Figure 1. First step is the analysis operation in which the sentence is analyzed morphologically and parsed syntactically. By this way, we obtain dependency relations between each token. Then, for the transformation stage, each rule is tried on the given sentence, and the first suitable rule is selected to be applied (only one rule could be applied over the sentence. If no suitable rule is found, the sentence will be left in its original form). Insertion of a token is performed in this level if it is considered as necessary. In the insertion step, shared arguments of the original sentence are derived first then each shared argument is inserted to the sub-sentence. Examples of insertion step is given in sections below. The rule in Figure 1 is explained in Section III-C1 in detail. In the generation step, the sentence is divided into sub-sentences corresponding to the information which is obtained from the transformation stage. At this phase, morphological information of the tokens may be updated to fit with the simplified version of the sentence. For this purpose, we use a morphological generator to reconstruct the new form of the token. The morphological generator produces a valid Turkish word by applying all the rules of a morphological analyzer in the reverse order (from lexical form through surface form). For example, the analysis of the participle "görmediğim"(*who I have not seen*) is produced as *gör*+VERB+NEG^DB+ADJ+PASTPART+P1SG by the morphological analyzer. This analysis is converted to *gör*+VERB+NEG+PAST+A1SG in the generation step to construct the predicate of the sub-sentence as "görmedim"(*I have not seen*). These three steps are valid for each rule which are explained in the below sections.

A. Coordinate Sentences

1) Shared Predicate: For this category, we introduce sentences in which the predicate is shared by elements which are interconnected coordination structure. A sample sentence under this category is shown in Figure 2. In the sample sentence, the word "sever" (*like*) is the shared predicate. Turkish allows the non-repetition of some words in the sentences which may cause difficulty in understanding the arguments of the shared predicate for children.

In this category, sentences are split, based on the number of sub-parts in the original sentence. The elements of the sub-parts are decided by the coordinated arguments in the sentence. For example, for the Figure 2, "Ali" and "Mehmet" are coordinated subjects and "basketbolu" (*basketball*) and "futbolu" (*football*) are coordinated objects of the same predicate. After splitting, the sentence is transformed into a new structure which is presented in simplified version of the Figure 2.

2) Shared Object: The sentence structure of this category is similar to the sentences in the Section III-A1. However, in this case an object is shared by the elements of the coordinated structure. An example under this category is given in Figure 3.

[2]In Turkish NLP, words are generally split into sub-word units from their derivational boundaries, each resulting unit having a potentially different part of speech tag and dependency relation.

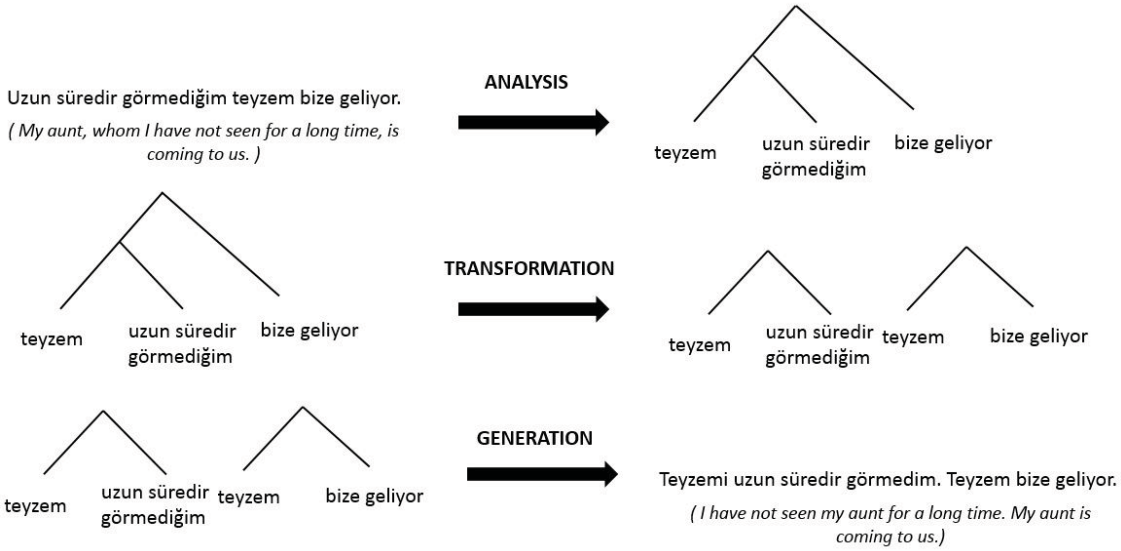
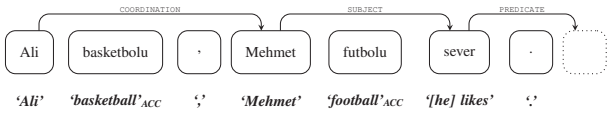


Fig. 1: Sentence Simplification steps

Dependency Graph



Original Version

'Ali basketbolu, Mehmet futbolu sever.'
(*'Ali [likes] basketball, Mehmet likes football.'*)

Simplified Version

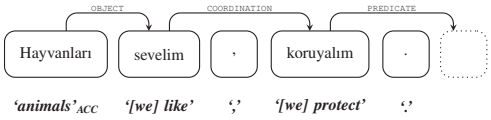
'Ali basketbolu sever. Mehmet futbolu sever.'
(*'Ali likes basketball. Mehmet likes football.'*)

Fig. 2: Example for Shared Predicate Category

The word "hayvanları" (*animals*) is the shared object by the two predicates "sevelim" (*like*) and "koruyalım" (*protect*). Instead of non-repetition of the shared argument, this may be a good way to use the same argument twice in the sentence. By this way, the meaning of the sentence may be given more clearly to preteens.

In this category, sentences are split based on the number of sub-parts in the original sentence. The elements of the sub-parts are decided by the coordinated predicates in the sentence. For example, for the sentence in Figure 3, "sevelim" (*like*) and "koruyalım" (*protect*) are coordinated predicate by the same object. In the splitting operation, the sentence is split into new sentences corresponding to the coordinated predicates and the shared arguments are put into all sub-sentences. After simplification, the sentence is divided into a number of parts, two in this case, and the split sentence is given in the simplified version of the Figure 3. This way, the sentence gives the same meaning, but with a syntactically simpler structure.

Dependency Graph



Original Version

'Hayvanları sevelim, koruyalım.'
(*'Let's like animals, protect (them).'*)

Simplified Version

'Hayvanları sevelim. Hayvanları koruyalım.'
(*'Let's like animals . Let's protect animals.'*)

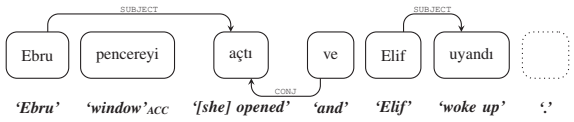
Fig. 3: Example for Shared Object Category

B. Paratactic Sentence

For this category we focused on sentences which do not have any shared argument or predicate. These consist of independent clauses separated by conjunctions or punctuation. As the predicates share no arguments, each sub-sentence has its own elements. An example sentence under this category is shown in Figure 4. As seen from the sample, there are two coordinated predicates: "açtı" (*open*) and "uyandı" (*woke up*). These predicates have their own arguments. For example, "Ebru" is the subject of "açtı" and "Elif" is the subject of "uyandı". In this category, sentences are split at

the conjunctions or punctuation marks which separate the independent clauses, resulting in a number of sub-sentences. Since these predicates have their own arguments, insertion of any argument is not performed in this process. The example is given in Figure 4.

Dependency Graph



Original Version

'Ebru pencereyi açtı ve Elif uyandı.'
(*'Ebru opened the window and Elif woke up.'*)

Simplified Version

'Ebru pencereyi açtı. Elif uyandı.'
(*'Ebru opened the window. Elif woke up.'*)

Fig. 4: Example for Paratactic Sentence Category

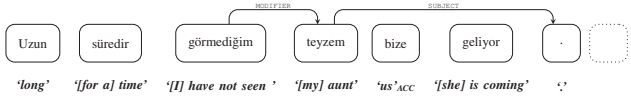
C. Subordinating Sentences

Subordinating sentence is a sentence which contains subclause. Subordinating sentences are not complete sentences by themselves, however they make additional information to complete the meaning of the whole sentence. These subclauses are formed by subordinate conjunctions (i.e. when, until and so on) and relative pronouns (i.e. who, which and so on). In this study, for Turkish, we also focused on these categories under two topics: 1. Participle Subclauses, 2. Converbial Subclauses.

1) Participle Subclauses: For this category, we introduce sentences containing subclauses the heads of which are participles. Participles are adjectives derived from a verb. An example under this category is given in Figure 5. When the English translation of the sentence is considered, the part which starts with the relative pronoun, "who" forms a subclause which modifies the word "aunt". In the Turkish sentence, the part "uzun süredir görmediğim" (*whom I have not seen for a long time* forms a subclause. This is a participle subclause because the head of this part is used as an adjective which modifies the word "teyzem" (*my aunt*).

In this category we benefit from the inflectional groups of the word in the sentence. In the example, when the sentence is semantically analyzed, the person whom I have not seen and the person who is coming are the same person. Using this property, the sentence is split into two parts. The first part covers the subclause arguments and the second one the main sentence arguments. In this category, there is an important issue. The token which is modified by the participle subclause is inserted to the first split part with the proper dependency relation. The word "teyzem" (*my aunt*) is in nominative case. Thus, when this token is inserted to the subclause part in the simplification process, the morphological analysis is changed to accusative case before using the morphological generator. This way, we ensure that the simplified sentences are gramatically correct.

Dependency Graph



Original Version

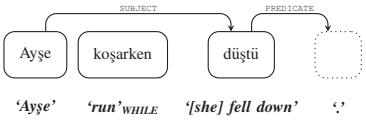
"Uzun süredir görmediğim teyzem bize geliyor."
(*"My aunt whom I have not seen for a long time, is coming to us"*)

Simplified Version

'Teyzemi uzun süredir görmedim.
Teyzem bize geliyor.'
(*'I have not seen my aunt for a long time.
My aunt is coming to us.'*)

Fig. 5: Example for Participle Subclause Category

Dependency Graph



Original Version

"Ayşe koşarken düştü."
(*"Ayşe fell down while [she was] running."*)

Simplified Version

"Ayşe koştu. Ayşe düştü."
(*"Ayşe ran. Ayşe fell down."*)

Fig. 6: Example for Converbial Subclause Category

2) Converbial Subclauses: The sentence structure of this category is similar to the sentences in the Section III-C1. For this category, we introduce the sentences containing subclause whose head clause is a converb. Converbs are adverbs derived from a verb inflectional group. An example under this category is given in Figure 6.

When the English translation of the sentence is considered, the part which starts with the subordinating conjunction, "while" forms a subclause which modifies the predicate of the main sentence, "fell down". In the Turkish sentence, the part "koşarken" (*while [she was] running*) forms a subclause.

This is a converbial subclause because the head clause of this sub-part is a converb which modifies the main predicate.

In the example, when the sentence is semantically analyzed, the person who fell down and the person who was running are the same person, "Ayşe". This word is only assigned as the subject of the main predicate in syntactic analysis. As a result, in the simplification process, this token is also inserted into the sub-sentence which is formed by the subclause. Also, the head of the converbial subclause is used as the verb of the first part sub-sentence. Therefore, this converb token is converted to the verb form using the morphological generator tool.

IV. Conclusion and Futurework

A sentence being long and complicated can pose multiple problems in daily life. For example, in Turkish, some words may be omitted from a sentence yet the meaning may remain the same. However, elementary school children (preteens) may face difficulty in understanding the arguments of the main predicate in a complicated sentence. For this reason, in this paper, we focus on solving this problem by simplifying the given sentences.

In this paper, we take advantage of inflectional groups in Turkish and investigate certain types of complex structured sentences. We divide them under three main categories as: 1. Coordinate Sentences, 2. Paratactic Sentences, 3. Subordinating Sentences. Then, we derive rules corresponding to each category and apply the rules to the sentences taken from an elementary school textbook. We present an automatic sentence simplifier for these categories and propose an approach to divide sentences to help children understand better.

Thus, as a future work we plan to verify the effectiveness of our simplification and preservation of meaning by testing our results on child readers. For validating our rules, we intend to use a human-focused evaluation based system with elemantary-school children as a testing audience.

V. Acknowledgements

This work is part of our ongoing research project "A Signing Avatar System for Turkish to Turkish Sign Language Machine Translation" supported by TUBITAK FATIH 1003 (grant no: 114E263). The authors want to thank Umut Sulubacak and Memduh Gökırmak for their valuable discussions and helps.

References

[1] W. M. Watanabe, A. C. Junior, V. R. Uzêda, R. P. d. M. Fortes, T. A. S. Pardo, and S. M. Aluísio, *Facilita: reading assistance for low-literacy readers*, ACM Std., 2009.

[2] J. De Belder and M.-F. Moens, *Text simplification for children*, ACM Std., 2010.

[3] S. Tyagi, D. Chopra, I. Mathur, and N. Joshi, *Classifier based text simplification for improved machine translation*, IEEE Std., 2015.

[4] A. Siddharthan, A. Nenkova, and K. McKeown, *Syntactic simplification for improving content selection in multi-document summarization*, Association for Computational Linguistics Std., 2004.

[5] G. Eryiğit, *ITU Turkish NLP Web Service*, Std., April 2014.

[6] Z. Zhu, D. Bernhard, and I. Gurevych, *A monolingual tree-based translation model for sentence simplification*, Association for Computational Linguistics Std., 2010.

[7] *A syntax-based statistical translation model*, Association for Computational Linguistics, 2001.

[8] K. Yamada and K. Knight, *A decoder for syntax-based statistical MT*, Association for Computational Linguistics Std., 2002.

[9] K. Woodsend and M. Lapata, *Learning to simplify sentences with quasi-synchronous grammar and integer programming*, Association for Computational Linguistics Std., 2011.

[10] S. Wubben, A. Van Den Bosch, and E. Krahmer, *Sentence simplification by monolingual machine translation*, Association for Computational Linguistics Std., 2012.

[11] B. P. Nunes, R. Kawase, P. Siehndel, M. A. Casanova, and S. Dietze, *As simple as it gets-a sentence simplifier for different learning levels and contexts*, IEEE Std., 2013.

[12] *Syntactic simplification and text cohesion*, vol. 4, no. 1, 2006.

[13] *Natural language processing for social inclusion: a text simplification architecture for different literacy levels*, 2009.

[14] S. Bott, L. Rello, B. Drndarevic, and H. Saggion, *Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish.*, Std., 2012.

[15] *Simplification syntaxique de phrases pour le français*, 2012.

[16] T. T. Vu, G. B. Tran, and S. B. Pham, "Learning to simplify children stories with limited data," in *Intelligent Information and Database Systems*. Springer, 2014, pp. 31–41.