Automatic Sentence Simplification in Low Resource Settings for Urdu

Yusra Anees and Sadaf Abdul Rauf
Fatima Jinnah Women University,Pakistan
(yusra.anees96,sadaf.abdulrauf)@gmail.com

Abstract

To build automated simplification systems, corpora of complex sentences and their simplified versions is the first step to understand sentence complexity and enable the development of automatic text simplification systems. We present a lexical and syntactically simplified Urdu simplification corpus with a detailed analysis of the various simplification operations and human evaluation of corpus quality. We further analyze our corpora using text readability measures and present a comparison of the original, lexical simplified and syntactically simplified corpora. In addition, we compare our corpus with other existing simplification corpora by building simplification systems and evaluating these systems using BLEU and SARI scores. Our system achieves the highest BLEU score and comparable SARI score in comparison to other systems. We release our simplification corpora for the benefit of the research community.

1 Introduction

Complex Sentences are always hard to understand for humans as well as automated applications. Complexity of a sentence often hinders proper communication of the intended meaning and hence a bottleneck in the learning pipeline. It has been found that students having problems with language often find it difficult to excel academically (Kyle, 2016). Research in the last decade has largely focused on complexity level identification of texts so that readability of such sentences can be enhanced to facilitate learning for students as per their learning grade.

Simplified sentences have specially been proved useful for producing understandable content for foreign speakers (Paetzold and Specia, 2016), language learners, children and people with lower literacy (Aluísio and Gasperin, 2010; Max, 2006; Petersen and Ostendorf, 2007). They are also recommended for cognitive and reading impairments like dyslexia (Rello et al., 2013), disorder of autism spectrum and aphasia (Carroll et al., 1999). On the other end, they are also valuable for many natural language processing (NLP) applications like semantic role labeling (Vickrey and Koller, 2008), machine translation (Oliveira et al., 2010) and relation extraction (Jonnalagadda et al., 2009) etc.

Generally, literary works have a higher difficulty level as compared to daily life language. This is specifically true for Urdu for which this gap is increasing day by day, literary texts often include complex words and composite sentence structure (Alison and Mushta,2004). Over the time with English taking over as official language and phenomenon of code mixing and switching becoming prevalent, the natives are inclined to use simpler language and often find it difficult to understand the level of Urdu used in traditional literature. Thus, a need arises for simplified versions of the classic works of the language to preserve the gems of literature and also to familiarize younger generation with such works.

Urdu is an international language having large quantity of educational and reading material and many new language learners. But unfortunately, sentence simplification is an unexplored area. Recently, Anees et al. (2020) present a small simplification corpus and Qasmi et al. (2020) present a simplification system using word embedding together with a set of morphological features to generate simplifications without parallel simplification data. It is the need of the day to ad-

Proceedings of the 1st Workshop on NLP for Positive Impact, pages 60–70
Bangkok, Thailand (online), August 5, 2021.

dress this issue and come up with effective complexity reduction and readability enhancement measures.

To be able to study complexity and simplicity parameters for any language, the first step is to have a corpus containing complex sentences and their simplified versions. Such sentence aligned texts are known as simplification parallel corpus and have been prepared for many languages, for example for English there exist simple Wikipedia corpus, PWKP (Zhu et al., 2010), Newsela (Xu et al., 2015), Onestop (Vajjala and Lucic, 2018) and SimPA (Scarton et al., 2018). Sentence simplification corpora for other languages include Ancora (Taulé et al., 2008), ERNESTA (Barbu et al., 2015), CLEAR (Grabar and Cardon, 2018) etc.

To enable research on automatic text simplification systems and text readability for Urdu, development of a simplification corpus providing enough complex sentences and their corresponding simple versions is imperative. We developed one such corpus for the high school students and simplified (lexically and syntactically) short stories from a renowned author. We have considered three-levels in our simplification process: Original, lexical and syntactic simplification. In Lexical Simplification (LS), complex words are replaced with simple and easy words. Whereas, Syntactic Simplification (SS) may result in an entirely new but simpler sentence.

We show the effectiveness of our corpora by human evaluation as well as comparing our corpus with other existing simplification corpora by building simplification systems. For automatic evaluation, we use BLEU (Papineni et al., 2001), an adequacy metric and SARI, simplification metric. Our system achieves the highest BLEU score and comparable SARI score in comparison to other systems. Lexical analysis and metric scores for each corpus, i.e. original, lexically simplified and syntactically simplified show correlation with human evaluations.

2 Literature review

Sentence simplification has been an active topic of research since the last decade. Many approaches have been proposed to develop the simplification corpora. Xu et al. (2015) present the first human built English simplification corpus, Newsela. It provides articles, re-written with 4 levels of readability for children of different ages. Brunato et al. (2015) report an Italian simplification corpus made using three-levels: local coherence, global coherence and lexical/syntax. Syntax and lexical simplification is done by reordering, insert, split, merge, transformation and delete. These simplification operations are also followed by (Tonelli et al., 2016; Brunato et al., 2016).

Vajjala and Lucic (2018) provide simplified version of texts taken from websites in three-levels elementary, intermediate and advanced. (Scarton et al., 2018) is a public administration domain corpus produced using syntactically and lexical simplification on around 1000 sentences. Other simplification corpora include (Grabar and Cardon, 2018) for French. Štajner et al. (2019) present an automatic lexical simplifier for Spanish by using synonyms and paraphrases from existing resources. The training corpus is from news and general literature consisting of 764 sentences. These are simplified using the six simplification rules defined in (Mitkov and Štajner, 2014).

For Urdu since no prior work exists on the topic, we follow the simplification schemes defined in the research literature and used the most frequent evaluation metrics to lay the ground work for future research.

3 Corpus Development

We gathered data from Urdu library [1] which has a huge collection of Urdu classic literary works. We chose 69 short, philosophical and thought-provoking stories based on daily life. These stories are written by Ashfaq Ahmad and published in the form of book. It uses complex sentence structure with typical Urdu literature vocabulary which is not very easy. We simplified the sentences using lexical and syntactic methods. Online Urdu Lughat [2] (dictionary) was used to find simpler synonyms.

All the sentences used in our corpus are available online. Initially we consulted language professionals to properly identify complex sentences in literature. Complex sen-

[1] http://www.udb.gov.pk/
[2] http://www.urdulibrary.org/

tences are further processed for removal of irrelevant characters and words to avoid ambiguities in data set. Rules for lexical simplification and syntactic simplification were defined after thorough literature review and discussion with language experts. Simplified corpora are rechecked by language experts to remove any anomalies. Our corpus creation methodology is consistent with the recent works like (Štajner et al., 2019; Scarton et al., 2018; Katsuta and Yamamoto, 2018; Grabar and Cardon, 2018; Brunato et al., 2016, 2015) who also simplified using basic lexical simplification operations and (Yatskar et al., 2010) for syntactic simplification. Since our corpus is composed of short stories, each sentence is linked to the previous and whole theme of stories is based on daily life emotions. The corpus is available at [3].

3.1 Simplification Annotation Scheme

Sentence simplification was performed using two techniques: lexical and syntactic substitution. LS uses lexical operations and SS uses syntactical operations. Most productive simplification operations according to literature including insertion, deletion, splitting, merging, substitution, deletion and reordering are used to produce the simpler sentences. During the corpus development process, we were also able to make a complex: simple word and paraphrase dictionary based on the simplifications applied on our text. Below we explain each of these operations with corresponding examples for a clear understanding of the operations.

3.1.1 Lexical Substitution

Lexical simplification operations include word and phrase replacement. LS replaces complex words in corpus by their simple synonyms or the complex phrase with its suitable analog.

Word level: Word level substitution is the case when a single word or compound word is replaced by the corresponding simple word(s).Rello et al. (2013) reported that dyslexic individuals understand more frequently used words better than their less frequent counterparts. We chose the most frequent synonyms for simplification, for example, for the sentence in example below, with lexical simplification, (e.g.“عہدے”«position» is replaced with ”نوکری“ «Job» and شریک حیات " is replaced with”بیوی“ «wife).

- ***Original.*** کہ شریک حیات کی موت پر آنسو نہ بہانے والا حرکت قلب بند ہو جانے پر اس دار فانی سے کوچ کر گیا ہے۔
- ***English.*** That the spouse who did not shed tears over the death of his spouse has escaped from his ordeal when the heart stopped beating.
- ***Simplified.*** کہ بیوی کی موت پر آنسو نہ بہانے والا دل کی دڑھکن بند ہو جانے پراس دنیا سے چلا گیا ہے۔
- ***English.*** The spouse who did not shed tears over the death of his spouse has died due to heartbeat stoppage.

Phrase level: Is the case when a group of words is replaced by a simple word or two words. Similarly, the complex phrase can also be replaced by the meaning of that phrase. For example in the following the phrase اس کی روح“ پرواز کر گئی” « his soul flew »is replaced by فوت“ ہو گئے”« died ».

- ***Original.*** ابھی وہ مسجد کی سیڑھیاں چڑھ ھی رھا تھا کہ اس کی حرکت قلب بند ھو گئی اور مسجد کے باھر ھی اس کی روح پرواز کر گئی۔
- ***English.*** As he was mounting the stairs of the mosque, his heart stopped and his soul flew, just outside the mosque.
- ***Simplified.*** ابھی وہ مسجد میں داخل ھو رھا تھا کہ اس کی دڑھکن رک گئی اور مسجد کے باھر ھی وہ فوت ھو گئے۔
- ***English.*** As he was entering the mosque his heart stopped and he died outside the mosque.

3.1.2 Syntactical Substitution:

”Syntactic simplification is the process of reducing the grammatical complexity of a text, while retaining its information content and meaning” (Siddharthan, 2006). It involves removal of phrases or words such that main context and meaning of sentence remains same. Syntactic simplification changes the order of words grammatically, and inserts new words

[3] https://github.com/umauh/Urdu-Sentence-Simplfication

to reduce the complexity. Merging and splitting of sentence are also used to reduce the complexity which is frequently used by (Zhu et al., 2010).

Deletion: Deals with removing extra information in a complex sentence to make it short, simple and clear to understand. A simple sentence normally has lesser numbers of words for conveying the important and content information. Often sentences use multiple adjectives which make the text complex and lengthy (Brunato et al., 2016) without conveying any meaningful information. For instance in the following sentence the phrases "وجہ یہ ہے کہ" « The reason is that», "پڑوس میں رہ کر"« living in your neighborhood », "ہاتھوں پل پل" « every moment » are redundant and thus deleted.

- ***Original.*** فوزیہ وجہ یہ ہے کہ میں نے پڑوس میں رہ کر تمہیں بچپن ہی سے اپنی سوتیلی ماں کے ہاتھوں پل پل دکھ جھیلتے اور اذیتیں اٹھاتے ہوئے دیکھا۔ ہے
- ***English.*** Fauzia the reason is that living in your neighborhood, I have seen you suffering every moment from your childhood at the hands of your stepmother.
- ***Simplified.*** فوزیہ میں نے تمہیں بچپن سے اپنی سوتیلی ماں سے دکھ سہتے ہوئے دیکھا ہے۔
- ***English.*** Fauzia I have seen you suffer from your stepmother since childhood.

Insertion: It is interesting that in the syntactically simplification process there is an insertion operation. Such operation is sometimes referred to as an 'elaboration' process, which is not simplification itself but helps improving text understanding. The simplified sentence may also be longer than its original sentence due to the insertion of meaning or some words which make the meaning of the original sentence clearer. Sometimes, it is difficult to predict the meaning of words or the text which requires supportive information for making it easy to understand. We have used the insertion operation for 9.12% sentences to clarify the meaning. For example, in the following sentence inserting "سے ملیں گی" « as well as get »made the meaning of the sentence complete.

- ***Original.*** تب مجھے تنخواہ بھی ملے گی اور ٹپس الگ۔
- ***English.*** Then I will get paid and tips aside.
- ***Simplified.*** تب تنخواہ بھی ملے گی اور ٹپس الگ سے ملیں گی۔
- ***English.*** Then I will get salary and tips will be given separately.

Reordering: This operation is carried out by changing the order of some words or phrases, e.g. changing the order of the clause in the original sentence to form a newer but simpler sentence (Brunato et al., 2016). In Urdu, reordering eliminates the complexity of sentence making it easier to understand, like in the example below changing order of the "میں سب سے بڑا ہوں" « I am the biggest » made the sentence easy to understand.

- ***Original.*** میں بھٹکے ہوئے مسافروں کو راستہ دکھاتا ہوں میں سب سے بڑا ہوں۔
- ***English.*** I guide lost passengers to the right path, I am superior to all.
- ***Simplified.*** میں سب سے بڑا ہوں کیونکہ میں بھٹکے ہوئے مسافروں کو راستہ دکھاتا ہوں۔
- ***English.*** I am the biggest because I guide the stray travelers.

Merging and Splitting: These operations are antithetical to each other in the simplification process. Merging is specifically used to join two or more sentences into one simplified sentence. It is commonly carried out by insertion of one or two suitable words or by placing suitable conjunction between both sentences. On the other hand, splitting is an operation through which one sentence is split into two or more sentences to make a simplified sentence (Gonzalez-Dios et al., 2018). For example in the following sentence merge and split make the sentence quite simple.

Merge.

- ***Original.*** دیکھتے ہی دیکھتے بچوں کا صفحہ پھٹ گیا۔ اور پھر وہ دونوں لڑ پڑے۔
- ***English.*** As you watch, the children's page tears. And then they both fought

- ***Simplified.*** صفحہ پھٹ گیا اوروہ دونوں لڑ پڑے۔
- ***English.*** The page tore and they both fought.

Split

- ***Original.*** یکن خیال رھے اسٹیج کے سامنے سبھی سوٹ بوٹ والے لوگوں کو بٹھانا اور مزدور پیشہ لوگوں کو نعرے لگانے اور تالیاں بجانے کے لئے پیچھے کھڑے کر دینا۔
- ***English.*** But be careful to place all the well dressed and suited people in front of the stage and make the working class people stand at back to clap and shout slogans.
- ***Simplified.*** لیکن اسٹیج کے سامنے سب امیر لوگوں کو بٹھانا - غریب لوگوں کو نعرے لگانے کے لئے پیچھے کھڑے کر دینا۔
- ***English.*** Seat all the rich people in front of the stage - Make the poor people stand behind to shout slogans.

3.2 Complex:Simple Lexicon

During the course of our simplification, we were able to develop a complex:simple word and phrase lexicon. Our lexicon has 490 dictionary entries, with 270 word-level and 220 phrase-level entries. For example, "تعارف" «introduction» has been translated to "نام" «name», "فکر مند" «concerned» to "پریشان" «upset» and "فصیح و بلیغ" «rhetorical» to "صاف تفصیل سے" «clearly». Similarly, around 220 phrases have been translated into simpler versions. For example,"سوٹ بوٹ والے لوگ" has been converted into "امیر لوگ", "حالت پر قابو پا کر" to"سمبھل کر" and "اس دار فانی سے کوچ کر گیا ہے" to "فوت ہو گیا". Context of a sentence is strictly followed in translation so that meaning of a sentence remains same. List of deleted words and inserted phrases has also been embedded into the corpus.

4 Human Evaluation

For evaluating the quality and simplicity of our corpus, we performed human evaluation which were done by two native Urdu speakers of 35 to 42 year with good grasp on the language.

We evaluated the sentences for adequacy, fluency and simplicity. The annotators were asked to rank the sentence pairs for the three parameters based on the questions given in Table 1. Q1 measures fluency of the sentence, Q2 is based on the adequacy of the sentence which is concerned with meaning preservation, and Q3 measures simplicity. We have the evaluation scheme used by (Sulem et al., 2018).We have made slight modification in Q2 and Q3 w.r.t to our simplification scheme as in Table 1, since our corpus has two levels of simplification in which lexical simplification is carried out by words and phrase transformation so in our case, complexity of words can not be ignored in human evaluation. Possible answers to these questions shown in Table 1 are : 1 is for "no", 2 is for "may-be" and 3 is for "yes". We used 3-scale criteria as (Sulem et al., 2018; Toutanova et al., 2016) prefer 3-scales over 5-scales. We measured inter annotator agreement using Cohen's kappa score which is reported in Table 2.

Human Evaluation Questions	
Fluency	Is the simplified sentence grammatical?
Adequacy	Does the Simplified sentence address the same information, compared to the original?
Simplicity	Is the simplified sentence simpler than the Original.?
Criteria	
1	No
2	May be
3	Yes

Table 1: Human evaluation questions and the criteria

	Fluency	**Adequacy**	**Simplicity**	**Average**
LS	0.76(0.31)	0.91(0.50)	0.8(0.41)	0.82(0.40)
SS	0.85(0.76)	0.78(0.45)	0.9(0.70)	0.84(0.63)

Table 2: Inter-annotator agreement score(Cohen's Kappa score) over human evaluation and Avg Human score carried out on Inter-annotator agreement score.

5 Simplification Statistics

We have produced a corpus of 1220 simplified sentences by simplifying 610 sentences, both lexical and syntactical. These are simplified using two level simplification process. Figure 1 presents the statistics of our simplification procedure. After an in depth analysis of language and content, we have approximately 58.8% sentences which were simplified using

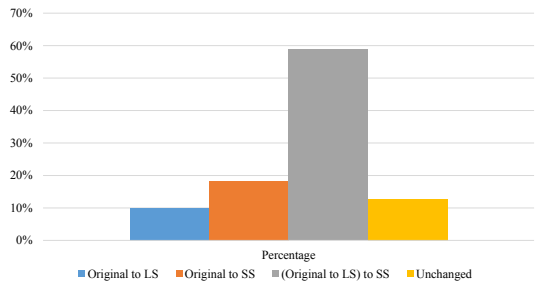

Figure 1: Percentages of each simplification level, LS indicates Lexical simplification and SS indication Syntactic simplification

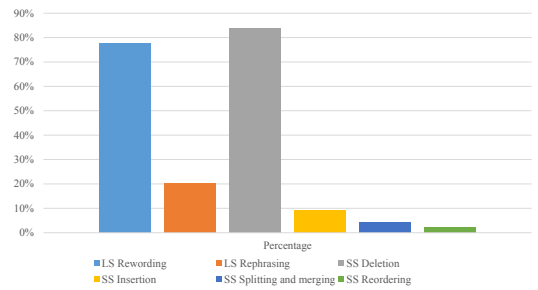

Figure 2: the percentage of each operation applied. LS indicates Lexical simplification and SS indication Syntactic simplification

three-level simplification, original to LS then Lexical simplification to syntactic simplification. On the other hand, 10% sentences were not very complex and only LS was sufficient to produce the final simplified version, whereas 18.3% sentences could only be simplified by SS. Around 12.7% sentences were simple enough not to require simplification of any form.

Figure 2 also summarizes the percentage of each of the simplification operation. Rewording is the most significant operation followed in Lexical Simplification through which 77.61% of simplification was accomplished. Same trend was observed by (Coster and Kauchak, 2011) where they report 65% rewording operations for English. In case of Syntactic Simplification, deletion was found to be the most frequent operation accounting to 84% of cases, this is also in line with results from previous researches (Coster and Kauchak, 2011; Brunato et al., 2016; Gonzalez-Dios et al., 2018). Insertion, split and merge and reordering follow with 9.12%, 4.24% and 2.14% usage respectively.

Figure 3 shows the data statistics found in the corpus, which depicts the average characters and words per sentence in the form of graph. Total numbers of words are lesser in lexical and syntactically simplified sentences as compared to original sentences. Average words per sentence in original sentence, Lexical simplified sentence and syntactically simplified sentence are 13.87, 13.51 and 10.33 respectively. This corpus can be specifically useful for developing automatic Sentence simplification as well as for improving many NLP tasks like text summarizing (Siddharthan, 2014), machine translation (MT) (Štajner and Popovic, 2016) and generation of questions (Heilman and Smith, 2010).

6 Text Simplification model

We used phrase based MT to build Automatic Text simplification models as has been commonly done in the literature. We divided our corpora into three parallel groups: (1) original to simplified lexical corpus, with 641 pairs of sentences with 31 sentences from the news domain, (2) lexical to syntactic simplified corpus with 661 sentences pair with 51 sentences added from kids stories and (3) (Original-Lexical-Syntactic) the concatenation of the both first and second group with the 1,302 sentences pair in which original appears two times as source data and lexical and syntactic level corpora as the target data. Each group is divided into 3 parts to build the PB-SMT models on random selection as 55% of sentences for training, 25% of sentences pairs for tuning and 20% of sentences pairs for testing.

Moses toolkit (Koehn et al., 2007) was used to train the simplification models separately.The models were evaluated using the EASSE toolkit and obtained the BLEU score 66.41 for first group of data, 40.18 for second group and 54.28 for concatenation of both data as shown in the Table 3. Our corpus may not be sufficient to build powerful models for simplifying sentences, but it is useful to test the generalization of the model for simplification of sentences.

Table 4 shows the simplified sentences from the model; first row sentences are simplified by Original to the lexical simplification system; second row is by Lexical to syntactic sim-

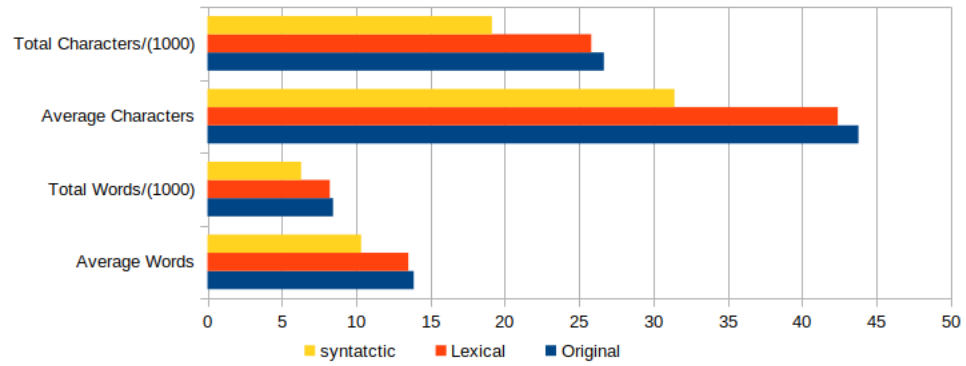

Figure 3: Average words and character per level in the corpus

Corpus	sentences pair	BLEU	SARI
Original to Lexical	641	70.437	37.382
Lexical to Syntactic	661	44.387	21.862
Original-lexical-Syntactic	1,302	53.615	28.333

Table 3: BLEU and SARI score of Urdu Systems

plification system; and the last row is simplified by Original-lexical-syntactic. The system has successfully replaced the complex word with a common word in the target sentences. The first system replaced نگاهیں <look> with نظر<look>. The second system performed a deletion operation in simplification, but the output sentence still needs to be corrected in comparison to the reference sentence. Increasing corpus size can improve the system output. In the last sentence, The system has simplified the sentence via both lexical and syntactically. The system has replaced the complex word حقارت<contempt> with کمتر <Inferior> and deleted کی the same as in reference sentence.

System	Sentence
1	
Input	اُسے چاروں طرف نگاهیں دوڑائی
Output	اُسے چاروں طرف نظر دوڑائی
Reference	اُسے چاروں طرف نظر دوڑائی (She looked around)
2	
Input	اس نے اپنے بچوں کی پرورش اور دیکھ بھال میں کوئی کمی نہیں آنے دیٔ ۔
Output	اس نے اپنے بچوں کی پرورش اور دیکھ بھال میں کمی نہیں آنے دیٔ ۔
Reference	اس نے اپنے بچوں کی پرورش میں کوئی کمی نہیں آنے دیٔ۔ (She did not allow her children's upbringing to diminish.)
3	
Input	وہ ہمیں حقارت کی نظروں سے دیکھتے ہیں۔
Output	وہ ہمیں کمتر نظروں سے دیکھتے ہیں۔
Reference	وہ ہمیں کمتر نظروں سے دیکھتے ہیں۔ (They look down on us.)

Table 4: Output example of each Urdu system mentioned in Table 3

7 Comparison of Systems

In order to establish a comparison of our prepared corpus with corpora of other languages, we built systems using various corpora including OneStopEnglish (Vajjala and Lucic, 2018), SimPA (Scarton et al., 2018), and PWKP (Hwang et al., 2015) corpus which are in English language, and Simpliki (Tonelli et al., 2016) and PaCCSS-It (Brunato et al., 2016) which are Italian simplification corpora. We translated these corpora to Urdu using google translate randomly selected 1,302 pairs of sentences. Automatic translations of the Turk corpus (Xu et al., 2015) were used as test set. Tables 5 and 6show the metrics scores and output of these systems. The system build on simUR (our created corpora) showed an excellent BLEU score. The SARI score of all systems is between 24 and 29. SARI score of SimUr is 26.036. Paccss-it obtained the best SARI score which is 29.441 but obtained the lowest BLEU score. The simplification of Paccss-it corpus is based on few additional operations such as verbal features, sentence type and this corpus original level is also more complex than our corpus. The lowest SARI score is obtained by the simPA-ls that are 24.738 where this corpus is based solely on lexical simplification but has obtained a higher BLUE score.

8 Discussion and Analysis

Table 5 shows that simUr got a better BLEU score as compared to other systems. However, the SARI score for the system is average. If scores of simUr are compared with other systems, it shows that the corpus level of this

system is intermediate because the values of this system are nearer to OneStopEnglish (Ele-Adv) corpus level. We can therefore conclude from the result that if the small corpus is to be built, then it should be complex on an advanced level as the paCCSS-It corpus. The paCCSS-It system achieved the highest SARI score, since it includes complex sentences in comparison with our corpus. The more complicated the corpus, the more vocabulary it will cover.

Because of the short dataset the PBSMT works well on lexical substitution. As Table 6 shows some simple sentences by all systems built on different corpora. simUr is the only system that has substituted the complex word حیثیت سے <As> with simple word طور پر <As> in the first sentence. simUr has simplified the sentence with a lexical operation, but the reference sentence is simplified with a syntactic operation. In second sentence, simUR has replaced the word لمبے لمبے <long long> with the بڑے بڑے <Big big> , PaCCSS-IT system has replaced ظاھر <appears> with شاید <Probably> and simPA-ls replaced طرف <side> with چاروں <All four>. In the reference sentence, لمبے لمبے شے <long long things> is replaced with پھیلی ہوئی چیز <Spread out>. In all these changes, the replacement of the simUr system is more similar in the sense of the original word based on the context of the Urdu language.

9 Summary

We have done experimentation through the supervised method using the Moses toolkit (Koehn et al., 2007) on our corpus. Three systems are constructed using the corpus; the first is based on a simplified lexical corpus, the second system is based on the syntactically simplified corpus, and the third system is based on the lexical and syntactic corpus. The lexical system achieved an excellent score of BLEU and SARI as compared to the other two systems.

Corpus	Sub-levels	BLEU	SARI
SimUR		50.736	26.036
Wiki		49.653	27.244
PaCCSS-IT		46.287	29.441
SimPA	*SS-sim*	46.19	28.854
	SS-LS-sim	49.276	27.351
	LS-sim	52.066	24.738
OneStopEnglish	*Ele-adv*	49.822	27.018
	Adv-ini	49.741	27.678
	Adv-ele	48.612	27.943
Simpitiki		50.523	25.835

Table 5: BLEU and SARI score of all systems

No.	System	Sentence
1	Input	یہ جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی حیثیت سے جاری ہے۔
simUR	Output	یہ جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی طور پر جاری ہے۔
paCCSS-IT	Output	جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی حیثیت سے جاری ہے۔
Wiki	Output	ہ جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی حیثیت سے جاری ہے۔
SimPA-ss	Output	یہ جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی حیثیت سے جاری ہے۔
SimPA-ls	Output	یہ جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی حیثیت سے جاری ہے۔
SimPA-ss-ls	Output	یہ جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی حیثیت سے جاری ہے۔
Adv-Ele1	Output	یہ جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی حیثیت سے جاری ہے۔
Adv-ini	Output	یہ جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی حیثیت سے حرارت چھوڑ ہے۔
ele-Adv	Output	یہ جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کی حیثیت سے آگے ہے۔
	Reference	اسے اب بھی جمہوریہ چیک میں بوھیمیا سوئٹزرلینڈ کہا جاتا ہے۔ (It is still called Bohemia Switzerland in the Czech Republic.)
2	Output	وائجر 2 امیجوں میں اوفیلیا لمبے لمبے شے کے طور پر ظاھر ہوتا ہے ، اھم محور یورینس کی طرف اشارہ کرتا ہے۔
simUR	Output	وائجر 2 امیجوں میں اوفیلیا بڑے بڑے شے کے طور پر ظاھر ہوتا ہے ، اھم محور یورینس کی طرف اشارہ کرتا ہے۔
PaCCSS-IT	Output	وائجر 1 امیجوں میں اوفیلیا لمبے لمبے شے کے طور پر شاید ہوتا ہے ، اھم محور یورینس کی طرف اشارہ کرتا ہے۔
Wiki	Output	وائجر 2 امیجوں میں اوفیلیا لمبے لمبے شے کے طور پر ظاھر ہوتا ہے ، اھم محور یورینس کی طرف اشارہ کرتا ہے۔
simPA-ss	Output	وائجر 2 امیجوں میں اوفیلیا لمبے لمبے شے کے طور پر ظاھر ہوتا ، ، اھم محور یورینس کی طرف اشارہ کرتا ہے۔
simPA-ls	Output	وائجر 2 امیجوں میں اوفیلیا لمبے لمبے شے کے طور پر ظاھر ہوتا ہے ، اھم محور یورینس کی چاروں اشارہ کرتا ہے۔
simPA-ss-ls	Output	وائجر 2 امیجوں میں اوفیلیا لمبے لمبے شے کے طور پر ظاھر ہوتا ہے ، اھم محور یورینس کی طرف اشارہ کرتا ہے۔
Adv-Ele	Output	وائجر 2 امیجوں میں اوفیلیا لمبے لمبے شے کے طور پر ظاھر ہوتا ہے ، کرم اھم محور یورینس کی طرف اشارہ کرتی ہے۔
Adv-ini	Output	وائجر 2 امیجوں میں اوفیلیا لمبے لمبے شے کے طور پر ظاھر ہوتا ہے ، اھم محور یورینس کی طرف اشارہ کرتا ہے۔
Ele-Adv	Output	وائجر 2 امیجوں میں اوفیلیا لمبے لمبے شے کے طور پر ظاھر ہوتا ہے ، اھم محور یورینس کی طرف اشارہ کرتا ہے۔
	Reference	وائجر 2 تصویروں میں اوفیلیا ایک پھیلی ہوئی چیز کے طور پر ظاھر ہوتا ہے۔ ایک پھیلی ہوئی چیز بڑی محور تھی۔ یہ یورینس کی طرف اشارہ کرتا ہے۔ (Ophelia appears in the Voyager 2 images as a spreading object. A spreading object was a big axis. It refers to Uranus.)

Table 6: comparison of output of all systems

We compared our corpus with other available simplification corpus by construction system through PBMT. All systems are trained on the same size of the corpus. These systems are test on the Turk corpus (Xu et al., 2015). For preparing other corpora, we translated all corpora to Urdu language via google translate.

Although the best score of BLEU is achieved by the system build on our simUr corpora as shown in table 5; however, the simUr system got a comparable SARI score. PaCCSS-It achieved the best SARI score. The simplification of this corpus is based on few additional operations such as verbal features and sentence type of PaCCSS-It, so it is also complex than simUr corpora. A PBMT system on Wiki corpus achieves almost similar levels of BLEU as SimUR. This shows that good simplification systems can be built for Urdu even with such small amounts of parallel corpus for lexical simplification.

10 Conclusion

We have introduced the first monolingual parallel Urdu corpus for sentence simplification using text from a famous writer's book. The corpus is the basic requirement for developing an automatic simplification system and has a multitude of applications in NLP. Our corpus contains 1220 simple sentences based on 610 complex sentences along with their simpler versions lexical and syntactic. This sim-

plification is carried out by using simplification operations including substitution, deletion, insertion and reordering of words and phrases. We also built simplification systems using our corpus and have taken an initiative towards Urdu simplification systems.

References

Sandra Maria Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53. Association for Computational Linguistics.

Yusra Anees, Sadaf Abdul Rauf, Nauman Iqbal, and Abdul Basit Siddiqi. 2020. Developing a monolingual sentence simplification corpus for Urdu. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 92–95, Seattle, USA. Association for Computational Linguistics.

Eduard Barbu, M Teresa Martín-Valdivia, Eugenio Martínez-Cámara, and L Alfonso Ureña-López. 2015. Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42(12):5076–5086.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

William Coster and David Kauchak. 2011. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2018. The corpus of basque simplified texts (cbst). *Language Resources and Evaluation*, 52(1):217–247.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Michael Heilman and Noah A Smith. 2010. Extracting simplified statements for factual question generation. In *Proceedings of QG2010: The Third Workshop on Question Generation*, volume 11.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado. Association for Computational Linguistics.

Siddhartha Jonnalagadda, Luis Tari, Jörg Hakenberg, Chitta Baral, and Graciela Gonzalez. 2009. Towards effective sentence simplification for automatic processing of biomedical text. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 177–180. Association for Computational Linguistics.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Confer-*

ence on Language Resources and Evaluation (LREC-2018).

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Kristopher Kyle. 2016. Measuring syntactic development in l2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication.

Aurélien Max. 2006. Writing for language-impaired readers. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 567–570. Springer.

Ruslan Mitkov and Sanja Štajner. 2014. The fewer, the better? a contrastive study about ways to simplify. In *Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 30–40.

Francisco Oliveira, Fai Wong, and Iok-Sai Hong. 2010. Systematic processing of long sentences in rule based portuguese-chinese machine translation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 417–426. Springer.

Gustavo Paetzold and Lucia Specia. 2016. Understanding the lexical simplification needs of non-native speakers of english. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 717–727.

K Papineni, S Roukos, T Ward, and W Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, 2001. *URL http://www1. cs. columbia. edu/nlp/sgd/bleu. pdf*.

Sarah Elizabeth Petersen and Mari Ostendorf. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. Citeseer.

Namoos Hayat Qasmi, Haris Bin Zia, Awais Athar, and Agha Ali Raza. 2020. Simplifyur: Unsupervised lexical text simplification for urdu. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3484–3489.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219. Springer.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Simpa: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Sanja Štajner, Horacio Saggion, and Simone Paolo Ponzetto. 2019. Improving lexical coverage of text simplification systems for spanish. *Expert Systems with Applications*, 118:80–91.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Semantic structural evaluation for text simplification. *arXiv preprint arXiv:1810.05022*.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*.

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpitiki: a simplification corpus for italian. *Proc. of CLiC-it*.

Kristina Toutanova, Chris Brockett, Ke M Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs.

Sowmya Vajjala and Ivana Lucic. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification.

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*, pages 344–352.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association of Computational Linguistics*, 3(1):283–297.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1353–1361. Association for Computational Linguistics.