CWITR: A Corpus for Automatic Complex Word Identification in Turkish Texts

BAHAR ILGEN

Language Technology Group, Department of Informatics, Universität Hamburg, Germany

CHRIS BIEMANN

Language Technology Group, Department of Informatics, Universität Hamburg, Germany

baharilgen@gmail.com; christian.biemann@uni-hamburg.de

The *Complex Word Identification* (CWI) task aims to provide support to resolve accessibility barriers for people who experience difficulties with cognitive, language, and learning disabilities. The task is concerned with the detection and identification of complex words that are unusual and difficult to understand by certain target groups. CWI systems have a large impact on the output of Text Simplification (TS) systems. This paper revisits the CWI task by extending available datasets by creating a new CWI corpus. In this study, we collect a new CWI dataset (CWITR) of complex single and multi-token words consisting of different text genres for Turkish and prepare it for investigation of computational methods on discrimination between complex and non-complex words forms.

CCS CONCEPTS • Computing Methodologies • Artificial Intelligence • Natural Language Processing • Language Resources

Additional Keywords and Phrases: complex word identification, lexical complexity, text simplification, crowdsourcing

ACM Reference Format:
Bahar ILGEN, Chris BIEMANN. 2022. CWITR: A Corpus for Automatic Complex Word Identification in Turkish Texts. In Conference Proceeding Series by ACM NLPIR 2022: Symposium on Natural Language Processing and Information Retrieval, December 16–18, 2022, Bangkok. Thailand, 10 pages.

1 INTRODUCTION

The first step of Text Simplification (TS) systems is to predict which words are complex considering a target population before applying any text simplification task. Complex Word Identification (CWI) is the task of determining words and phrases that are considered difficult to understand by the target audience. CWI is a subtask of Lexical Simplification (LS) pipeline and accessibility [1-3]. The definition of a complex word and related parameters have been investigated in the scope of psycho and neurolinguistic research. Previous research identified major parameters related to word complexity consisting of factors such as word frequency, word length, and the position of phonemes within a word [4]. Once complex words and phrases in a text are

determined, these units are replaced by simpler alternatives. Considering all words as complex units is not practical in the LS task. Some LS systems first identify the complex words and replace them with potentially simpler alternatives. In contrast to assuming all words as complex units, selecting too few complex words has the potential of resulting in a bad performance on the task efficiency. On the other hand, identifying too many words might lead to erroneous substitutions and meaning loss [5]. Categories of CWI tasks for the available strategies can be classified into five groups. These groups consist of the following approaches: simplifying everything, threshold-based and lexicon-based approaches, machine learning assisted and implicit CWI strategies [6]. In the scope of implicit approaches, CWI is performed implicitly during other steps of the pipeline instead of an initial step. Regarding the latter approach, the availability of CWI datasets plays an important role in the accuracy of applications such as LS tasks. In this study, we collect a new CWI dataset (CWITR) in different groups of text genres (NEWS, WIKIPEDIA, WIKINEWS, PERIODICALS, BOOK SUMMARIES) for Turkish and prepare it to investigate the performance of different algorithms.

Automatic identification of complex words is linked to several language-related areas of research. Lexically and semantically complex words and phrases may cause difficulties in reading and understanding texts. Text Simplification, Lexical Simplification [7] and Reading Assessment [8] are principal areas that have the potential to benefit from the CWI task. TS task aims to reduce the linguistic complexity of a given text to improve understandability and readability by still maintaining the original meaning [9,10]. The output of the TS task is utilized to improve the comprehension of different groups of people such as individuals with low-literacy levels, children, second language learners, and people with several cognitive impairments. The latter group includes the disorders such as aphasia and dyslexia in which proposed simplification techniques may vary based on the needs of special groups. While second language learners possibly have a limited vocabulary, people with cognitive disorders may have difficulties distinguishing passive/active voice forms which may affect the whole meaning of a sentence drastically. Texts with shorter and more frequent words have been found useful for people with dyslexia since they have difficulties reading and understanding long forms [11]. Apart from its role as a solution in target groups, TS is also a preparatory step to improve the results of other NLP tasks such as automatic text summarization, machine translation, sentence fusion, and semantic role labeling.

Initial attempts on the TS task include the approaches using hand-crafted syntactic rules, generating shorter sentences, and active/passive voice transformations [12-14]. In a more data-driven attempt, Narayan and Gardent [15] utilized the English Wikipedia (EWKP) and the Simple English Wikipedia (SWKP) to form a parallel corpus for the simplification task. Most of the LS systems rely on the usage of parallel corpora, sentence alignments and news articles. The CWIG32 dataset [16] was annotated by both native and non-native English speakers. The CWIG32 covers three text genres and provides an extension to the Wikipedia genre which is basically addressed in most of the previous studies [2, 17, 18]. Additional categories of professionally written news articles, amateurishly written articles, and Wikipedia articles are the new genres of this dataset. In addition, both native and non-native annotators take part in the annotation process.

In the scope of this study, we collect a Turkish Complex Word Identification dataset using Amazon Mechanical Turk (MTurk) crowdsourcing platform for annotations following similar settings that are made for the CWIG32 dataset. As in the CWIG32 dataset, users have been displayed paragraph contexts to let them annotate both complex words and word phrases. Annotators are expected to provide native/non-native information with their additional language-level information. We use Wikipedia and WikiNews genres with additional professionally written texts on several subjects. A sample HIT (Human Intelligence Task) with its

sample complex word/phrase selections is shown in Figure 1. These annotations are supposed to be utilized for the automatic prediction of complex words and phrases and be investigated in terms of their success and impact on different genres.

2 RELATED WORK

As CWI systems have gained more attention in recent years, several competitions were organized such as CWI2016, CWI2018, and CWI2021. The first shared CWI task was organized under the International Workshop on Semantic Evaluation (SemEval-2016). Users are asked to label complex and non-complex words to perform binary classification. Participants were selected from the pool of non-native English speakers. In the scope of the competition, 21 teams took place with the submission of 42 systems. Several features such as syntactic, semantic, morphological, word and character n-grams, word embeddings, psycholinguistic features, and Zipfian distribution were utilized by the participants. The second edition of the competition was held in the scope of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA) in 2018. The second organization brought a new perspective to the research area by including languages and datasets other than English such as French, German, and Spanish [19].

The complexity of a given word can be explained by several parameters. The following parameters have been pointed out after analyzing the systems and datasets participated in CWI-2016 and CWI-2018. The word might be an archaic word or an atypical one because it was borrowed from some other language. It might be one of the uncommon or infrequent words. It might relate to a very specific concept. Although it is a common word, it may have very uncommon usage in the given context as a polysemous word. The complex 2.0 dataset has been prepared and annotated for complexity levels. During the collection of the dataset three different sources have been used to provide sufficient complexity levels. These consist of Bible, Europarl, and Biomedical sources. Since these resources are sufficiently diverse, it is possible to cover different complexity levels (e.g., Bible usually does not have archaic words or very specialized types of usages can be found in the biomedical domain) [19].

The Lexical Complexity Prediction (LCP) task was organized at SemEval-2021. During the Semeval-2021 LCP task, participants were provided with the augmented version of the Complex Corpus [20]. This is a multi-domain corpus with words and multi-word expressions (MWEs), which are annotated using a five-point Likert scale (i.e., very easy, easy, neutral, difficult, very difficult). The task also featured focusing on two subtasks namely, words and MWEs. The participated systems are mainly categorized into three types. These consist of feature-based systems, deep learning systems, and a final group of systems that utilizes a hybrid approach of the other two categories [21]. Although the results have shown that deep learning-based system results are superior to the others, the results of feature-based systems have been found successful and not far behind this group. Word embeddings from resources such as GLOVE and Word2Vec with other lexical complexity features are the popular and widely used ones together with regression systems such as Gradient Boosted Regression and Random Forest Regression [19]. Pre-trained language models and fine-tuning using transfer learning is followed by the groups which are opted to follow deep learning approaches. In this context, BERT and RoBERTa were widely used in the scope of Task-1. ALBERT and ERNIE were also utilized by the participants [22].

Earlier studies on complex word identification handle the problem by attempting to simplify all the words or to use frequency threshold approaches [7, 23]. During more recent competitions, probabilistic classification was

also performed on the given tasks. This information could be gathered by considering the total number of annotators for a complex word.

There are several techniques such as feature-based and deep learning approaches for identifying complex words. The set of features that are utilized in this scope usually consists of; morphological features such as frequency counts, term frequency and several statistics, syntactic and lexical features, psycholinguistic and lexical features, word embedding features, and classical ML learning methods. Aroyehun et al. (2018), [24] compared the results of experiments with feature engineering approaches and Deep Learning approaches using Convolutional Neural Networks (CNN). Sheang (2019), [25] utilized word embeddings and engineered features with an approach to CWI based on CNN trained on pre-trained word embeddings with morphological and linguistic features. Hartmann and Dos Santos (2018) [26] developed approaches using feature engineering, a shallow neural network method using only word embeddings, and a *Long Short-Term Memory* (LSTM) language model that is pre-trained on a large text corpus.

3 COLLECTING CWI TURKISH DATASET

We collected complex word and phrase annotations (sequences of words, up to a maximum of 50 characters), using the Amazon Mechanical Turk crowdsourcing platform, from native and non-native Turkish speakers. We asked participants whether they are native or non-native Turkish speakers or not and collected their proficiency levels for non-native speakers. Because Turkish is not widely used as a second language, all participating annotators were native speakers in our experiments. We also prepared a language proficiency exam that is required to be taken before the annotation starts. The proficiency test consists of 9 questions with a total of 100 points. Within the scope of the test, questions about Turkish spelling mistakes, semantic integrity, and grammatical structures were asked to participants. The exam requires a browser login so any user can only take it once. Only the annotations of workers who have been successful (i.e., participants with a score of 65 and above) in this exam were accepted for the tasks.

3.1 Data Selection

Collected texts consist of Wikipedia news, Wikipedia articles, news, novel summaries, and periodicals (i.e., newspaper columns on different domains including history, technology, science, society, and others). These are paragraph-length texts that can vary between specified number of sentences. Figure 1 shows a sample HIT highlighted with complex word annotations. In Figure 2 and Figure 3, selections of annotators, and instructions for the annotation process are displayed respectively. Figure 3 displays the rules of the process that are given to annotators. It is expressed that the difficulty level in Turkish written texts will be considered and evaluated in terms of non-native language users, language learners, children, and people with cognitive disorders. The information given for annotators includes the minimum and the maximum number of words that should be highlighted as well as illegal selection examples (e.g., selecting a whole sentence, or selecting part of a word). It is also noted that proper nouns and several surface forms of the same words should be avoided for annotation.

Figure 1: Sample HIT for identifying complex words

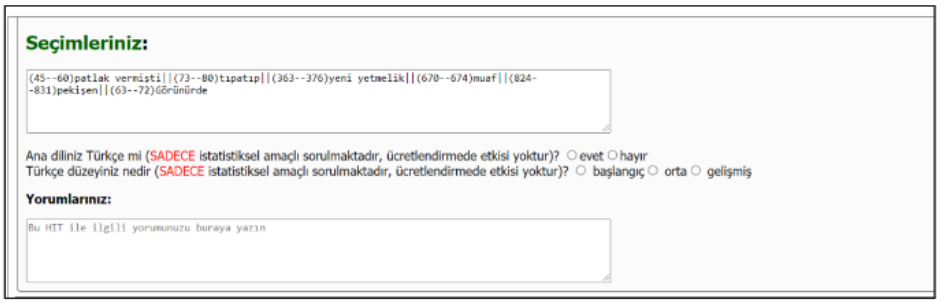

Figure 2: Sample HIT screen of annotated CPs with comment box and user's language level information

3.2 Annotation Procedure

Paragraph level texts have been displayed to the workers on the Amazon Mechanical Turk crowdsourcing platform. These consist of texts from 5 to 10 sentences. Workers are supposed to highlight at least 3 complex words and/or complex phrases (CP). Otherwise, they are informed that they should leave a comment in the text box that is shown in Figure 2. It is prohibited to annotate an arbitrarily large number of selections. The selection cannot exceed 10 complex words and/or complex phrases. In addition, a worker cannot select a whole sentence, a part of words or phrases, etc. Annotators are also notified not to select proper nouns and surface forms of the same word or phrases. There are also two additional questions that workers provide, whether they are native Turkish language speakers, and their knowledge level in Turkish (beginner, intermediate, advanced). Although the system is designed to collect annotations of both native and non-native language speakers of participants, all MTurk annotators were native speakers in our annotation tasks.

3.3 Preprocessing

Datasets from batches were preprocessed to gather approved annotations and remove irrelevant information from those such as HITTypeId, title, keywords, several timestamps, and worker information. The final format of the dataset has the following information: HITId, text (the sentence in which a complex phrase occurs), offset for the complex phrase where it starts and ends in the sentence, number of native language speakers, number of non-native speakers, and the total number of annotators. Some of the annotations that are not Turkish words were removed from the dataset. Multi-word phrases have also been excluded when they are not in a proper

form (e.g., annotation of plain comma-separated words instead of compound word phrases). In Figure 4., the dataset format with sample complex phrase annotations is shown. The dataset contains information about complex phrases annotated with the following statistics. Each line in the dataset represents the information of a sentence with one complex phrase annotation and relevant information (HITId, text, start and end offsets, complex word, number of native and non-native speaker annotators, and the total number of annotators respectively), each separated by a TAB character. The first sentence in Figure 4 is “Görüntüler düşük çözünürlüklü kamera ile çekildi.” (“Images were taken with a low-resolution camera.”) and the word “çözünürlüklü” (“resolution”) is annotated as a complex word.

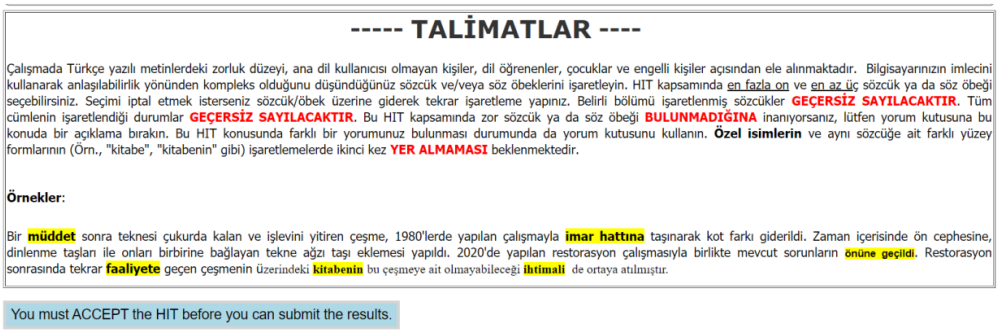

Figure 3: HIT Instructions

```
33CLA8O0MIBF9S1KXTPP1DA7XNTRFP  Görüntüler düşük çözünürlüklü bir kamera ile çekildi    17  29  çözünürlüklü    2   0   2
33CLA8O0MIBF9S1KXTPP1DA7XNTRFP  Merkür manyetik alanı ile güneş rüzgarı ile etkileşimi incelenecek  44  54  etkileşimi  1   0   1
33CLA8O0MIBF9S1KXTPP1DA7XNTRFP  Merkür manyetik alanı ile güneş rüzgarı ile etkileşimi incelenecek  7   21  manyetik alanı  1   0   1
33CLA8O0MIBF9S1KXTPP1DA7XNTRFP  Merkür'ün keşfi için görevlendirilen bir misyon , gezegenden alınan ilk görüntüleri Dünya'ya gönderdi   41  47  misyon  4   0   4
33CLA8O0MIBF9S1KXTPP1DA7XNTRFP  Yavaş hareket eden aracın 2025 yılı sonunda gezegen etrafında istikrarlı bir yörüngeye oturması bekleniyor  62  72  istikrarlı  4   0   4
378G7J1SJLYSBAIA1CFP6UXOD1TEWX  Avustralya'da elektrikçi lisansınız yoksa ampulünüzü değiştiremezsiniz  25  35  lisansınız  3   0   3
378G7J1SJLYSBAIA1CFP6UXOD1TEWX  Hastane çalışanları ve cenazeye katılanlar ise bu yasaktan muaftır  59  66  muaftır 4   0   4
378G7J1SJLYSBAIA1CFP6UXOD1TEWX  Milano'da çok eskiden beri yürürlükte olan yasaya göre kamusal alanda somurtmak yasaklanmıştır  70  79  somurtmak   4   0   4
378G7J1SJLYSBAIA1CFP6UXOD1TEWX  Örneğin Florida'da bekar kadınlar Pazar günleri paraşüt atlayışı yapamıyor  48  55  paraşüt 2   0   2
```

Figure 4: Sample lines from the dataset that represent sentences with one complex phrase (CP) annotation and relevant information, each separated by a TAB character.

4 ANALYSIS OF COLLECTED ANNOTATIONS

A total of 25 native speakers of Turkish participated in the annotation task and a total of 21,436 complex phrase annotations have been collected from all genres, out of which 13,837 unique CPs. These were deemed as complex by at least one annotator. Among these, there are 9,229 single and 4,608 multi-token complex phrases. In the following sections, we discuss details of the data collected. In total, 6 workers have participated in 50% of 1000 HITs where on average 198 assignments are completed by native speakers. Around 61% of CPs among all annotations have been selected by at least two annotators. The total Turkish CWI dataset consists of 1000 paragraphs (HITS). Each HIT is required to be annotated by 5 workers. After applying preprocessing steps that include data cleaning and removal of unnecessary information such as timestamps, the information belonging to rejected HITs, etc., the dataset is prepared in the same format as the CWIG3G2 dataset [16]. The dataset consists of 5 genres, and a total of 18 subcategories of periodicals, news, WikiNews, Wikipedia and book summaries.

Table 1 shows the distribution of selected complex phrases across all annotators. The percentages of book summaries and periodical categories yielded higher results than the others for the multiple-selection case. These values account for complete annotations in all batches. Table 2 and Table 3 display the distribution of HITS and annotated CPs across genres respectively. In Table 4, the ratio of complex phrases across genres and categories that are selected by at least two annotators has been shown for unique instances.

Table 1: Distributions of selected CPs (in %) across all (native) annotators, The Sing. column stands for annotations selected by only one annotator while the Mult. column stands for annotations selected by at least two annotators.

Dataset	All	
	Sing.	Mult.
News	47	53
Wikipedia	46	54
WikiNews	41	59
Book Sums.	35	65
Periodicals	36	64

Table 2: Distribution of HITs (in %) across genres

Wikipedia	News	WikiNews	Periodicals	Book Sums.
35	25	19	11	10

Table 3: Distribution of annotated CPs (in %) across genres

Wikipedia	News	WikiNews	Periodicals	Book Sums.
36	22	20	12	10

Table 5 summarizes the statistics of annotated words. Among 9229 complex words, the average frequency of the same words in different HITs is 5.56. The average number of syllables, length and non-vowels are 3.75, 8.84, and 5.15. Since Turkish is an agglutinative language, words may take inflectional and derivational suffixes in a flexible way. DB stands for “derivational boundary” and indicates that the word takes a new form by changing its structure and the main tag. Words might have new forms more than once, and the number of DBs indicates the average number of such transformations for annotated complex words. Among all annotated complex words, approximately 29% of these words have one or more DBs, and ~71% have no DBs. Table 6 shows the distribution of word types among annotated complex words. These main tagsets consist of nouns, verbs, adverbs, adjectives, as well as the other group (i.e., Conjunctions, Duplications, Pronouns, Postposition, Numbers, and Questions) [27].

Table 7 shows the samples of complex words in surface and root forms together with their morphological analysis. The root of the word “dondurma” (ice cream) is a verb known as “don-mak” (to freeze). Since the word is transformed two times as don→ dondur → dondurma, there are two DBs. The tagset after the last DB belongs to the final form of the word. In this case, it is initially a verb and transformed to some causative form (to make it frozen), and a noun at the end. Table 8 summarizes the statistics of all batches undertaken with Amazon MTurk.

Table 4: Ratio of CPs (annotated at least 2 times or more – in %) across genres and categories

Genres	Categories	CP Ratio
Book Sums.	Novel-1	38
	Novel-2	36
Wikipedia	Sports	34
	History	31
	Science	33
	Wiki-Exclusive	33
	Society	30
	Technology	29
	Wikigen2	32
	Wikigen1	31
News	TurNews-1	28
	TurNews-2	29
	World News	35
	Tur News Final	38
WikiNews	Wikinews-1	35
	Wikinews-2	31
Periodicals	Periodicals-1	38
	Periodicals-2	37

Table 5: Average numbers for word frequencies, DBs, Syllables, length, and non-vowels among annotated complex words.

Freqs.	#DBs	#Syllable	Length	Non-vowel
5.56	1.41	3.75	8.84	5.15

Table 6: Distribution of word types among complex words

Word Type	Distribution
Adjective	10.4%
Adverb	3.4%
Noun	68.3%
Verb	17.0%
Others	0.9%

Table 7: Morphological Analysis of Complex Words

word	root	analysis of word
izinsiz *(unauthorized)*	izin *(permission)*	Noun+A3sg+Pnon+Nom+**^DB**+Adj+Without
inip *(after going down)*	in (-mek) *(going down)*	Verb+Pos+**^DB**+Adverb+AfterDoingSo
etkileyici *(impressive)*	etkile(-mek) *(to impress)*	Verb+Pos+**^DB**+Noun+Agt+A3sg+Pnon+Nom
dondurma *(ice cream)*	don(-mak) *(to freeze)*	Verb+**^DB**+Verb+Caus+Pos+**^DB**+Noun+Inf2+A3sg+Pnon+Nom

Table 8: Statistics of all batches undertaken with Amazon Mechanical Turk

Number of Annotators	25
Number of Instances	13,837
Number of Annotations	21,436
Annotations per Instance	1.55
Instances per Annotator	857.44

5 CONCLUSION AND FUTURE WORK

This paper presents the CWITR – a Turkish CWI dataset - preparation steps using the MTurk crowdsourcing platform. In the scope of this work, we included new genres to the dataset to provide a broader and more reliable CWI system. These cover several data sources in varying complexity levels. Both complex words and word phrases were annotated by MTurk workers. Although our tasks have been prepared for both native and non-native Turkish speakers, only native speakers took part in the annotations. Because less-resourced languages are used less frequently as a second language, the scarcity of non-native speakers during the experiments is evaluated as an expected outcome.

The dataset has been shared with appropriate licensing. It will be utilized in future experiments to investigate the impact of complex word annotations in Turkish, and to predict complexity scores for the single words and MWEs.

REFERENCES

[1] Paetzold, G., and Specia, L. .2016a. Unsupervised lexical simplification for non-native speakers. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1).

[2] Shardlow, M. .2013a. The CW corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations* (pp. 69-77).

[3] Shardlow, M. .2014. Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* (pp. 1583-1590).

[4] Ziegler, W., and Aichert, I. 2015. How much is a word? Predicting ease of articulation planning from apraxic speech error patterns. Cortex, 69, 24-39.

[5] Shardlow, M. .2013b. A Comparison of Techniques to Automatically Identify Complex Words. In 51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop (pp. 103-109).

[6] Paetzold, G. H., & Specia, L. .2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60, 549-593.

[7] Bott, S., Rello, L., Drndarević, B., and Saggion, H. .2012. Can spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proceedings of COLING 2012,* (pp. 357-374).

[8] Collins-Thompson, K. .2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2), 97-135.

[9] Al-Thanyyan, S. S., and Azmi, A. M. .2021. Automated text simplification: A survey. *ACM Computing Surveys* (CSUR), 54(2), 1-36.

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

[11] Rello, L., Baeza-Yates, R., Dempere-Marco, L., and Saggion, H. (2013). Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction* (pp. 203-219). Springer, Berlin, Heidelberg.

[12] Siddharthan, A. .2002. An architecture for a text simplification system. *In Language Engineering Conference, 2002. Proceedings* (pp. 64-71). IEEE.

[13] Siddharthan, A. .2010. Complex lexico-syntactic reformulation of sentences using typed dependency representations. In *Proceedings of the 6th International Natural Language Generation Conference*.

[14] Siddharthan, A. .2011. Text simplification using typed dependencies: A comparison of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation* (pp. 2-11).

[15] Narayan, S., and Gardent, C. .2016. Unsupervised sentence simplification using deep semantics. *In Proceedings of the 9th International Natural Language Generation conference*, pages 111–120, Edinburgh, UK.

[16] Yimam, S. M., Štajner, S., Riedl, M., & Biemann, C..2017. CWIG3G2-complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Volume 2: Short Papers) (pp. 401-407).

[17] Horn, C., Manduca, C., and Kauchak, D. .2014. Learning a lexical simplifier using Wikipedia. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (pp. 458-463).

[18] Paetzold, G., & Specia, L. .2016b. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 560-569).

[19] Shardlow, M., Evans, R., & Zampieri, M. .2021a. Predicting lexical complexity in English texts. arXiv preprint arXiv:2102.08773.

[20] Shardlow, M., Cooper, M., and Zampieri, M. .2020. Complex: A new corpus for lexical complexity prediction from Likert scale data. *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)* (pp. 57-62).

[21] Zaharia, G. E., Cercel, D. C., and Dascalu, M. .2021. UPB at SemEval-2021 Task 1: Combining Deep Learning and Hand-Crafted Features for Lexical Complexity Prediction. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* , (pp. 609-616).

[22] Shardlow, M., Evans, R., Paetzold, G. H., & Zampieri, M. .2021b. Semeval-2021 task 1: Lexical complexity prediction. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 1-16).

[23] Thomas, S. R., & Anderson, S. .2012. WordNet-based lexical simplification of a document. *In Proceedings of KONVENS 2012*, (pp. 80-88).

[24] Aroyehun, S. T., Angel, J., Alvarez, D. A. P., and Gelbukh, A..2018. Complex word identification: Convolutional neural network vs. feature engineering. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications,* (pp. 322-327).

[25] Sheang, K. C. .2019. Multilingual complex word identification: Convolutional neural networks with morphological and linguistic features. In *Proceedings of the Student Research Workshop Associated with RANLP 2019* (pp. 83-89).

[26] Hartmann, N., and Dos Santos, L. B..2018. NILC at CWI 2018: Exploring feature engineering and feature learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 335-340).

[27] Oflazer, K. .1994. Two-level description of Turkish morphology. Literary and linguistic computing, 9(2), 137-148.