Article

Automatic Text Simplification of Hungarian Texts

Martin Sallai [1], Márk Muliter [1], René Zinga Banda Firmino [1] and Csaba Ömböli [1]

[1] Eötvös Loránd University, ELTE, IK (Faculty of Informatics), Pázmány Péter 1/C, 1117, Budapest, Hungary ;
* Correspondence: nlp@inf.elte.hu

Version November 1, 2022 submitted to Languages

Abstract: Reducing the complexity of texts by applying an Automatic Text Simplification (ATS) system has been sparking interest in the area of Natural Language Processing (NLP). So far proposed state-of-the-art methods mainly focus on lexical, syntactic, discourse simplification and respectively machine translation. Concerning Hungarian language, thanks to the lack of large amounts of corpora (original-simplified) the development of an ATS tool has not been proposed yet. As a result of this, we present our system, which is a hybrid text simplification tool for Hungarian language based on lexical simplification and rule-based syntactic and discourse transformations. Furthermore we show automatic and human evaluations. As well as we discuss the performances of our system at the lexical, syntactic, and discourse levels.

Keywords: automated text simplification; hybrid architecture; Hungarian corpora

---

1. Introduction

Text simplification (TS) is the process of rewriting a complex text into a simpler form while preserving its meaning. The purpose of text simplification is to assist the comprehension of readers, especially language learners and children [9]. Usually unnecessary details are omitted. Another characteristic trait of simplified texts is that usually only one main idea is expressed by a single sentence. This also means that in the simplification process complex sentences are often split into several smaller sentences. The availability of a sentence-aligned corpus of original texts and their simplifications is of paramount importance for the study of simplification and for developing an automatic text simplification system [3].

There is never a halt in the swiftness of any language, it always moves forward, Hungarian, also called Magyar, traditionally belongs to the Ob-Ugric languages (e.g. Khanty and Mansi) of the Finno-Ugric branch of Uralic. Hungarian is the official language of the Republic of Hungary, and has approximately fifteen million speakers, of which four million reside outside of Hungary [11]. The fact is that the Hungarian language, unlike the Germanic, Romance and Slavic languages of Europe (but similar to Greek, Albanian, or Armenian, for instance) has no close relatives. All the other Finno-Ugric languages are geographically and genetically far away. As there are no “almost-Hungarian” languages that a Hungarian speaker can “almost” understand, there is no easy way for Hungarians to experience relatedness between languages [8].

According to our study, several works have been done on Hungarian natural language processing, yet despite its importance, there has not been any attempt to simplify linguistic complexity in Hungarian texts, therefore the goal of our case is to simplify a complex sentence and make it more readable and easily understandable. Our model is combining several text simplification approaches, namely lexical, syntactic, discourse simplification, that is why the architecture of the model is called hybrid. Firstly we discuss the state-of-the-art related works of this field in Section 2. Secondly in Section 3 we propose our model. After that we show the results our model in Section 4 and in conclusion to this paper we mention the discussion of our model in Section 5 and the conclusion in Section 6.

2. Related Work

Numerous ATS-related research has been published over the past 20 years, as reviewed by Saggion (2017)[19], and Al-Thanyyan and Azmi (2021)[1], just to mention some of them. In short, the field has mostly concentrated on creating techniques for automatically simplifying difficult words (lexical simplification) and/or difficult syntactic structures (syntactic simplification). Tong Wang (2016)[27] described a study of the LSTM based model for text simplification. This study shows several operational rules such as sorting, reversing, replacing sentence pairs, meanwhile in the same year a rule-based text simplification model for the German language was proposed by Julia Suter at the University of Zurich [24]. They use experiments to explain how RNN and LSTM operate. Their model can perform distinct sorting, reversing, and replacement processes. It requires to combine all three procedures to simplify. Later then there a new language representation model were introduced called BERT, which stands for Bidirectional Encoder Representations from Transformers. (Kristina Toutanova, 2019)[4] Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018)[16][18], BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. While some work has also been developed for languages like Spanish, Portuguese, Basque, French or even Japanese, Hungarian has been hardly researched. One of these few Hungarian-related research were Dávid Márk Nemeskey's, published in 2020 [13]. It contains huBERT: a variation of BERT models especially for the Hungarian language, using the WebCorpus 2.0, which is worth to mention, because this is the largest Hungarian NLP dataset with over 9 billion words in it. It has begun to be questioned how historically, simplifications are assessed using automated criteria like BLEU[15] or the Flesch-Kincaid Reading Grade Level[10] (Sulem et al., 2018; Tanprasert and Kauchak, 2021; Alva-Manchego et al., 2021)[23][25][2]. The goals of ATS research have also been questioned. Stajner (2021)[22] highlights how prospective target groups characteristics have not been considered and urges the creation of more modular ATS systems that can be tailored for certain populations. Some of the most recent approaches, like Maddela et al. (2021)[7] or Sheang and Saggion (2021)[20] for English. Some of the above-mentioned publications have been used as resources to help us better grasp text-simplification techniques.

3. Method

3.1. Introduction to methods

Nowadays most systems do not focus on introducing new techniques for text simplification but instead focus on implementing existing techniques in their own language.[1] Since Hungarian language does not have a large corpus for natural language processing and concerning the fact that during our research we did not encounter any automated text simplification method for Hungarian language, we decided to establish the foundation of this field of science. Generally, automatic text simplification approaches are classified into four classes: lexical, syntactic, monolingual machine translation, and hybrid techniques. As far as our proposed method is considered for Hungarian text simplification we tried to follow HECTOR's [26] framework, which is a hybrid text simplification tool for raw french texts. This simplification architecture consist of 4 steps, as illustrated on Figure 1: preprocessing, syntactic simplification, discourse simplification, lexical simplification. We kept the architecture and built the models based on Hungarian language corpora and on some points we suggested different simplification techniques in order to increase the evaluation results of the architecture.

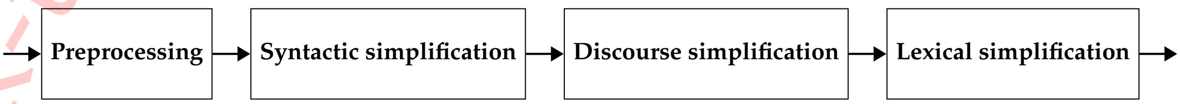

Figure 1. Architecture of the proposed method

3.1.1. Corpus and lexicon

First of all, concerning our corpus we manually simplified original literary (tales and stories) and scientific (documentary) texts and created a database called HUNALECTOR, modelled on the French text simplification corpus ALECTOR [6].Examples and the structure of the dataset can be seen in Table 1, respectively an English translated example in Table 2. It consists of complex sentences along with their simplified equivalents. In this developed corpus, we manually identified the simplification operations at the syntactic and discourse levels. Furthermore we created a lexicon of 10,000 unique words from our corpus labelled with complex or non-complex tag. The creation of the lexicon and the corpora was a combined effort by volunteers and ourselves. The volunteers consisted of native and non-native Hungarian speaking people. Having constructed the corpus and the lexicon, we spent quite amount of time proofreading and making sure that there are no spelling errors. Moreover both the corpora and the lexicon were also verified by the Hungarian Research Centre for Linguistics.

Original	Simplified
"A barlang lassan megtelt a széttépett liba belső részeinek nehéz szagával, s erre megmozdult az egész halom kisróka."	"A barlang lassan megtelt a széttépett liba belső részeinek nehéz szagával. Megmozdult az egész halom kisróka."
"A barlang lassan megtelt a széttépett liba belső részeinek nehéz szagával,..."	"A széttépett liba belső részeinek nehéz szaga megtöltötte a barlangot,..."
"A barlang lassan megtelt a széttépett liba belső részeinek nehéz szagával, s erre megmozdult az egész halom kisróka."	"Megmozdult az egész halom kisróka."
...	...

Table 1. Example of the HUNALECTOR dataset. Original(complex) and simplified Hungarian sentences in comparison.

Original	Simplified
"The cave was slowly filled by the heavy smell of the insides of the torn goose, and to this the whole pile of little foxes moved."	"The cave was slowly filled by the heavy smell of the insides of the torn goose. The whole pile of little foxes moved."
"The cave was slowly filled by the heavy smell of the insides of the torn goose,..."	"The heavy smell of the insides of the torn goose slowly filled the cave,..."
"The cave was slowly filled by the heavy smell of the insides of the torn goose, and to this the whole pile of little foxes moved."	"The whole pile of little foxes moved."
...	...

Table 2. Example of the HUNALECTOR dataset. Original(complex) and simplified translated English sentences in comparison.

3.2. *Hybrid simplification architecture approach*

Our hybrid system, capitalizes on lexical re-sources available, and builds linguistically grounded rules for syntactic and discourse transformations. It integrates a data-driven lexical simplification module with a hand-crafted rule-based syntactic simplification module supplemented with preprocessing and discourse simplification.

3.2.1. Preprocessing

The preprocessing module aims to provide the additional information required by the syntactic simplification. We used HunTag3 [5] to recognize the entity of the words and to identify the structure of the sentence, an example can be seen on Table 3.

Noun	Verb
Vuk	megmozdult
Vuk	moved

Table 3. Example sentence "Vuk megmozdult"("Vuk moved") tagged by HunTag3.

3.2.2. Syntactic simplification

Syntactic simplification (SS) is the task of simplifying the complex syntactic structures in a text while preserving its information content and original meaning. For example coordination, subordination, relative clauses, and passive relative clauses can be considered as complex syntactic structures. Syntactic simplification is mostly done in three stages: analyzation, transformation, regeneration. [1]

In analyzation phase the sentence's complexity is determined, which decides if it requires simplification. We automated the determination of complexity using the combination of matching rules and an Support Vector Machine binary classifier based on the structure of the sentence. [1]

In transformation phase, the modifications are made to the parse tree according to a set of pre-written rules. These rules perform the simplification operations, e.g., sentence splitting,clause rearrangement, and clause dropping. All together we used 32 syntactic transformation rules consisting of hand-crafted Hungarian language specific rules and general rules collected from different literatures. [1] [26] [21] The three most general transformation rules are discussed below with an example:

Sentence splitting: Relative clauses are extracted and transformed into main clauses.

Hungarian: *"A barlang lassan megtelt a széttépett liba belső részeinek nehéz szagával, s erre megmozdult az egész halom kisróka."* → *"A barlang lassan megtelt a széttépett liba belső részeinek nehéz szagával. Megmozdult az egész halom kisróka."*

English: *"The cave was slowly filled by the heavy smell of the insides of the torn goose, and to this the whole pile of little foxes moved."* → *"The cave was slowly filled by the heavy smell of the insides of the torn goose. The whole pile of little foxes moved."*

Sentence structure adjustments: The passive voice is transformed into active voice form.

Hungarian: *"A barlang lassan megtelt a széttépett liba belső részeinek nehéz szagával,..."* → *"A széttépett liba belső részeinek nehéz szaga megtöltötte a barlangot,..."*

English: *The cave was slowly filled by the heavy smell of the insides of the torn goose,..."* → *"The heavy smell of the insides of the torn goose slowly filled the cave,..."*

Secondary information suppression: The adverbial,past and present participle clauses are removed.

Hungarian: *"A barlang lassan megtelt a széttépett liba belső részeinek nehéz szagával, s erre megmozdult az egész halom kisróka."* → *"Megmozdult az egész halom kisróka."*

English: *"The cave was slowly filled by the heavy smell of the insides of the torn goose, and to this the whole pile of little foxes moved."* → *"The whole pile of little foxes moved."*

In the literature of syntactic simplification we can find some methods where generalization phase is made after transformation phase, but in our case since we used discourse simplification as well therefore we transferred this step to next section.

3.2.3. Discourse simplification

Syntactic simplifications might suppress important information for textual cohesion: suppressing pronouns or some secondary clauses might cut or mix up the co reference chains. In order to reduce the amount of these inferences, we applied discourse simplification rules maintaining the structure of co reference chains. In few words, discourse simplification helps and improve the readability of syntactically simplified sentences. We created our model for discourse simplification based the method proposed in HECTOR [26] and DisSim [14].

Replace new or repeated entities: Reduces the quantity of processing inferences done by the reader.

Hungarian: *"Kag megállt mellette. Jól emlékezett.→ Kag megállt mellette. Kag jól emlékezett"*

English: *"Kag stopped next to it. He remembered well.→ Kag stopped next to it. Kag remembered well."*

Specify entities: Replace the demonstrative determiner by a definite one.

Hungarian: *".., mert titokzatos volt neki ez a ház,... → ..., mert titokzatos volt neki a ház,... "*

English: *"..., because this house was mysterious to him ,... → ..., because the house was mysterious to him ,..."*

3.2.4. Lexical Simplification

As far as lexical simplification is concerned we followed the structure of LSBert [17], which includes the following three steps: complex word identification, substitute generation, filtering and substitute ranking. LSBert simplifies one complex word at a time, and is recursively applied to simplify the sentence.

Complex world identification

We trained bi-directional long short-term memory units predict the binary complexity of words as annotated in the dataset. Since Hungarian language does not have a complex word identification model we had to build one. We decided to build the model based on the proposed method in LSBert lexical simplifier. [17] [31] This method takes a sequence of words as an input and outputs a sequence of classification tags. Given a predefined threshold p, if the lexical complexity of one word is greater than the threshold, it will be treated as a complex word. The algorithm starts with the world which has the highest likelihood of belonging to complex class.

The tags can be Complex(C) or Non-Complex(NC):

$$T = s1, s2, ..., sn$$

$$s1 = w1, w2, ..., wn$$

"..., ki tudja nem őrzi-e a házat valami vahúr!"

$$wi = [NC, NC, NC, NC, NC, NC, NC, C]$$

Substitute generation

Given a sentence *si* and the complex word *wi*, the aim of substitution generation (SG) is to produce the substitute candidates for the complex word *wi*. We trained the model BERT on our database. [4] [17] After the BERT evaluation we select the top 5 words from as substitution candidates, excluding the morphological derivations. The structure of BERT and an example of substitute generation can be seen on Figure 2.

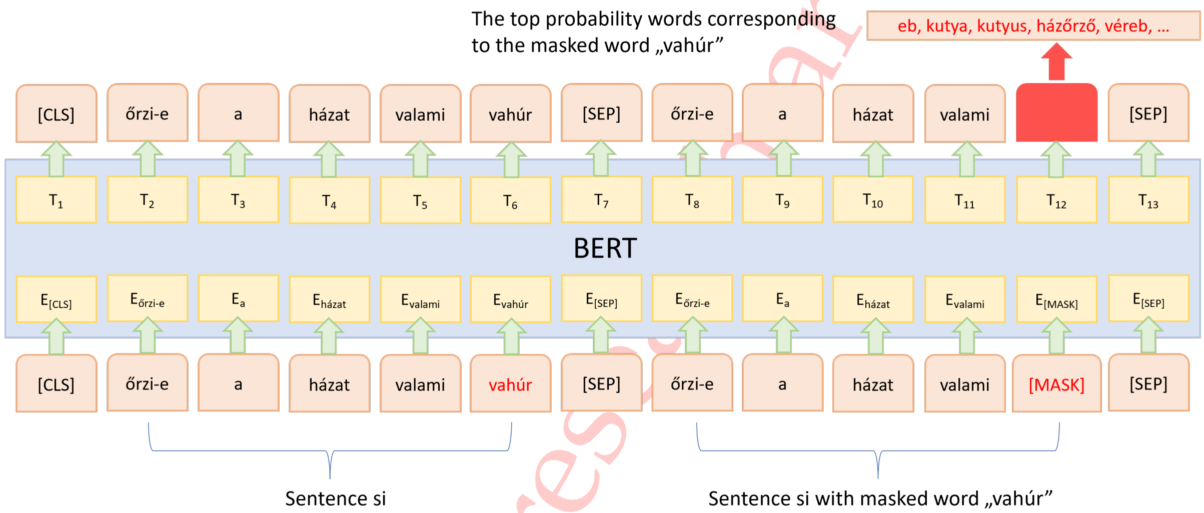

Figure 2. Example of substitute generation using BERT

Substitute ranking

Giving substitute candidates:

$$C = c1, c2, ..., cn,$$

the substitution ranking of the lexical simplification framework is to decide which one of the candidate substitutions that fits the context of complex word is the simplest. For this task frequency-based candidate ranking strategies are one of the most popular choices and they are quite effective. In general, the more frequently a word is used, the most familiar it is to readers. Unfortunately a general Hungarian frequency dictionary currently does not exist, thus we used the Wikipedia's frequency list of Hungarian words.

4. Results

There are various techniques existing for judging the quality of the text simplification model output as well for comparing the performance of different text simplification models. These methods can be divided into two groups: automatic and manual evaluation techniques. Usually these two are both used to measure the model performance, as both have their own advantages. In this section we describe the methods used in our solution for evaluating outputs. We tested our model for three tasks: *lexical*, *syntactic* and *discourse* simplification.

In this section we first assess our system's performance using human evaluation, following the methods and aspects described in [26], and then introduce our results achieved by using automated measures, following the steps of [32].

The test were carried out on our own dataset, HUNALECTOR.

4.1. Human evaluation

Human evaluation is the more straight-forward way of measuring a text simplification model's performance, as the aspects these rating are based on correspond to concepts familiar to our way of thinking. Human evaluation was done by native Hungarian speakers, mostly by volunteers. They were asked to rate the simplified sentences on three dimensions, which are described below. All rating answers were given on a 1-5 Likert scale. The outputs were evaluated based on the following three aspects:

- **Simplicity:** Measures how simple the simplified sentence is.
- **Fluency (grammatically):** Measures the grammatical correctness.
- **Adequacy (meaning preservation):** Measures how well the original meaning is preserved.

The final scores come from the mean of scores on all used datasets for each of the three criteria.

We used 100 sentences from our dataset for human evaluation, while we ignored sentences without changes from the examples. Sentences from literary and scientific texts from our dataset were used during the assessment.

The given scores from all participants in the evaluation were taken into account to compute the final scores for all aspects. For these final results, we computed the mean of scores for each aspect. The results are described in Table 4.

Name	Fluency	Adequacy	Simplicity	All
DRESS	3.72	**3.65**	2.53	3.30
PBMT-R	2.85	2.78	**3.11**	2.92
Hybrid	3.65	2.94	3.10	3.23
Our model	**3.75**	3.52	3.19	**3.49**

Table 4. Other models' human evaluation results based on their own dataset, compared to our results. These results can be better compared as a score given by humans is less dependent on the used dataset. The values for other datasets are the mean of their performance on all the datasets they were evaluated on.

4.2. Automatic evaluation

For automatic evaluation we used the most commonly used metrics for this task. During the assessment, 3 different scores were taken into account: SARI[30], BLEU[15] and the Flesch-Kincaid Grade Level index (FKGL)[10]. We used our own dataset HUNALECTOR for running the automatic measurements.

Following [32] we used BLEU[15] to assess the degree to which generated simple sentences differed from ground truth samples and the Flesch-Kincaid Grade Level index (FKGL[10]) to measure the readability of the output where a lower FKGL[10] score implies simpler output. In addition, we used SARI[30] which provides a score computed from comparing the output against the source and reference simplifications.

We compared our model's performance with 3 reference models: DRESS[32], a reinforcement learning-based simplification system, PBMT-R[28], a monolingual phrase-based machine translation system with a re-ranking post-processing step and Hybrid[12], a model which first performs sentence splitting and deletion operations over discourse representation structures and then further simplifies sentences with PBMT-R. The reference models' performance was measured in [32], these are the values we compared our model to.

The final scores for these 3 models were computed by taking the mean of scores these models achieved on 3 different datasets, namely Newsela[29], WikiSmall[33] and WikiLarge[32] as described in [32]. The results of automatic evaluation are presented in Table 5.

Name	SARI	BLEU	FKGL
DRESS	48.54	9.11	23.43
PBMT-R	39.12	**17.77**	30.62
Hybrid	44.97	6.06	30.64
Our model	**49.78**	12.32	**31.87**

Table 5. Other models automatic evaluation scores based on their respective datasets compared to our results. These results can not be compared by these metrics alone, as the datasets the models were evaluated on differ. But these results show anyway, that our model achieves state of the art performance. The values for other datasets are the mean of their performance on all the datasets they were evaluated on.

4.3. Error Analysis

In this section we would like to list some common mistakes our model made and show example outputs for these errors.

Syntactic errors. A common mistake our model made was giving the output in the wrong word order. For example:

Hungarian:*"A barlang lassan megtelt a széttépett liba belső részeinek nehéz szagával,..." -> "A liba belső széttépett részeinek nehéz szaga megtöltötte a barlangot,..."*

English:*"The cave was slowly filled by the heavy smell of the insides of the torn goose,..." -> "The heavy smell of the torn insides of the goose slowly filled the cave,..."*

In this example we can see, that the change of word order changes the meaning of the sentence. In the original example the goose is torn, while in the wrong simplification the insides of the goose are torn.

Discourse errors. A regularly occurring error was the usage of the wrong form of inflection on certain words, These changes can change the meaning of sentences, so these should also be punished during assessment. An example for this kind of error:

Hungarian:*"Se nappal, se éjjel nem nyugszik a rókák népe ilyenkor[...]" -> "Ilyenkor a rókák népe nem nyugodni[...]"*

English:*"In such times the nation of foxes have no rest neither at day or night." -> "In such times the nation of foxes no rest."*

Lexical errors. In scientific texts the change of certain words to their synonyms can change the meaning of the sentence as scientific words usually have distinct meanings. For example:

Hungarian:*"Nevezzétek meg a halmazokat!" -> "Nevezzétek meg a kupacokat!"*

English:*"Name the sets!" -> "Name the heaps!"*

This error in simplification can occur in Hungarian language as the meaning of the word "halmaz" and "kupac" can be synonyms in everyday language, but have different meaning in a scientific environment.

5. Discussion

The aim of the research was to establish the foundation of Hungarian text simplification. We created our simplification tool for a specific domain, since our corpora is based on Hungarian children books, namely better understanding for disabled children. The models on the dataset performed quite good, but for other domains the models should be learned on different datasets. Our errors were mainly caused by the fact that the size of our corpora is rather small comparing to for example English language corpora. Despite of this we showed similar results to other non English language text simplification methods.

6. Conclusion

Automatic text simplification is far from perfect. Although inventing new approaches and methods is not really trending nowadays, since current state-of-the-art methods are just using already proposed methods on different languages. Concerning the next break-through of automated text simplification we support the idea of reverse engineering how children learn complex linguistic structures as it was suggested in this state-of-the-art survey of automated text simplification. [1]

References

[1] Suha Althunayyan and Aqil Azmi. "Automated Text Simplification: A Survey". In: *ACM Computing Surveys* 54 (Mar. 2021), Article no. 43. DOI: 10.1145/3442695.

[2] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. "The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification". In: *Computational Linguistics* 47.4 (Dec. 2021), pp. 861–889. ISSN: 0891-2017. DOI: 10.1162/coli_a_00418. eprint: https://direct.mit.edu/coli/article-pdf/47/4/861/1979827/coli_a_00418.pdf. URL: https://doi.org/10.1162/coli%5C_a%5C_00418.

[3] Stefan Bott and Horacio Saggion. "An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction". In: June 2011, pp. 20–26.

[4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805.

[5] István Endrédy and Balázs Indig. "HunTag3, a general-purpose, modular sequential tagger - chunking phrases in English and maximal NPs and NER for Hungarian". In: 2015.

[6] Núria Gala et al. "Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1353–1361. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.169.

[7] Sebastian Gehrmann et al. *The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics*. 2021. DOI: 10.48550/ARXIV.2102.01672. URL: https://arxiv.org/abs/2102.01672.

[8] Daniela Haarmann. "The Hungarian Language Issue in Hungary and Transylvania Before 1795". In: 57 (Jan. 2018), pp. 385–402.

[9] Tomoyuki Kajiwara and Mamoru Komachi. "Text Simplification without Simplified Corpora". In: *Journal of Natural Language Processing* 25 (Mar. 2018), pp. 223–249. DOI: 10.5715/jnlp.25.223.

[10] J Peter Kincaid et al. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Tech. rep. Naval Technical Training Command Millington TN Research Branch, 1975.

[11] Beáta Megyesi. "The Hungarian Language". In: (July 2001).

[12] Shashi Narayan and Claire Gardent. "Hybrid Simplification using Deep Semantics and Machine Translation". In: *the 52nd Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. ACL. Baltimore, United States, June 2014, pp. 435–445. URL: https://hal.archives-ouvertes.fr/hal-01109581.

[13] Dávid Márk Nemeskey. "Natural language processing methods for language modeling". In: (2020).

[14] Christina Niklaus et al. "DisSim: A Discourse-Aware Syntactic Text Simplification Framework for English and German". In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, Oct. 2019, pp. 504–507. DOI: 10.18653/v1/W19-8662. URL: https://aclanthology.org/W19-8662.

[15] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

[16] Matthew E. Peters et al. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association

for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

[17] Jipeng Qiang et al. "LSBert: Lexical Simplification Based on BERT". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3064–3076. DOI: 10.1109/TASLP.2021.3111589.

[18] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[19] Horacio Saggion. "Automatic text simplification". In: *Synthesis Lectures on Human Language Technologies* 10.1 (2017), pp. 1–137.

[20] Kim Cheng Sheang and Horacio Saggion. "Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer". In: *Proceedings of the 14th International Conference on Natural Language Generation*. Aberdeen, Scotland, UK: Association for Computational Linguistics, Aug. 2021, pp. 341–352. URL: https://aclanthology.org/2021.inlg-1.38.

[21] Advaith Siddharthan. "Syntactic Simplification and Text Cohesion". In: *Research on Language Computation* 4 (July 2004). DOI: 10.1007/s11168-006-9011-1.

[22] Sanja Stajner et al. *Lexical Simplification Benchmarks for English, Portuguese, and Spanish*. 2022. DOI: 10.48550/ARXIV.2209.05301. URL: https://arxiv.org/abs/2209.05301.

[23] Elior Sulem, Omri Abend, and Ari Rappoport. *BLEU is Not Suitable for the Evaluation of Text Simplification*. 2018. DOI: 10.48550/ARXIV.1810.05995. URL: https://arxiv.org/abs/1810.05995.

[24] Julia Suter, Sarah Ebling, and Martin Volk. "Rule-based Automatic Text Simplification for German". In: Sept. 2016.

[25] Teerapaun Tanprasert and David Kauchak. "Flesch-Kincaid is Not a Text Simplification Evaluation Metric". In: *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1–14. DOI: 10.18653/v1/2021.gem-1.1. URL: https://aclanthology.org/2021.gem-1.1.

[26] Amalia Todirascu et al. "HECTOR: A Hybrid TExt SimplifiCation TOol for Raw Texts in French". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 4620–4630. URL: https://aclanthology.org/2022.lrec-1.493.

[27] Tong Wang et al. *An Experimental Study of LSTM Encoder-Decoder Model for Text Simplification*. 2016. DOI: 10.48550/ARXIV.1609.03663. URL: https://arxiv.org/abs/1609.03663.

[28] Sander Wubben, Antal van den Bosch, and Emiel Krahmer. "Sentence Simplification by Monolingual Machine Translation". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 1015–1024. URL: https://aclanthology.org/P12-1107.

[29] Wei Xu, Chris Callison-Burch, and Courtney Napoles. "Problems in Current Text Simplification Research: New Data Can Help". In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 283–297. DOI: 10.1162/tacl_a_00139. URL: https://aclanthology.org/Q15-1021.

[30] Wei Xu et al. "Optimizing Statistical Machine Translation for Text Simplification". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 401–415. DOI: 10.1162/tacl_a_00107. URL: https://aclanthology.org/Q16-1029.

[31] Seid Muhie Yimam et al. "CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups". In: *IJCNLP*. 2017.

[32] Xingxing Zhang and Mirella Lapata. "Sentence Simplification with Deep Reinforcement Learning". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 584–594. DOI: 10.18653/v1/D17-1062. URL: https://aclanthology.org/D17-1062.

[33] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. "A monolingual tree-based translation model for sentence simplification". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010, pp. 1353–1361.