huPWKP: A Hungarian Text Simplification Corpus

Noémi Prótár
Eötvös Loránd University
protarnoemi@student.elte.hu

Dávid Márk Nemeskey
Eötvös Loránd University
Department of Digital Humanities
nemeskey.david@btk.elte.hu

Abstract

In this article we introduce `huPWKP`, the first parallel corpus consisting of Hungarian standard language–simplified sentence pairs. It is the Hungarian translation of PWKP (Zhu et al., 2010), on which we performed some cleaning in order to improve its quality. We evaluated the corpus both with the help of human evaluators and by training a seq2seq model on both the Hungarian and the original (cleaned) English corpus. The Hungarian model performed slightly worse in terms of automatic metrics; however, the English model attains a SARI score close to the state of the art on the official PWKP set. According to the human evaluation, the corpus performs at around 3 on a scale ranging from 1 to 5 in terms of information retention and increase in simplification and around 3.7 in terms of grammaticality.

1 Introduction

The most important function and goal of human communication is joint meaning construction (Tolcsvai Nagy, 2017): we want every person who participates in the discourse to understand the referential scene (Tátrai, 2017, 2020) – i.e. what we are talking about – exactly (or as similarly as possible) as we intended it to be understood. In order to achieve this, we sometimes need to simplify what we are saying and how we are phrasing it: meaning, we need to reduce ”the linguistic complexity of a text, while still retaining the original information content and meaning” (Siddharthan, 2014). Simplified texts can be of use for several groups of people, e. g. for people with (communicative or other) disabilities (Maaß and Rink, 2020; Maaß and Hernandez Garrido, 2020), non-native speakers (Paetzold, 2015) or children (De Belder and Moens, 2010). However, as text simplification is a fairly time- and resource-consuming task for humans, it seems beneficial to try to automate this task. There have been multiple successful attempts at creating text simplificaton systems: most of them for English, e.g. Zhu et al. (2010) or Xu et al. (2016) or Xu et al. (2015). Less-resourced languages, such as Hungarian, have been largely ignored in the literature. In this paper, we introduce the first (albeit translated) Hungarian parallel corpus consisting of standard language – simplified sentence pairs, as well as a simplification model trained on it.

2 Related work

2.1 Text simplification in NLP

Text simplification (TS) is a fairly popular research area in NLP, especially for the English language. Most modern TS systems are capable of abstractive text simplification, meaning they can create new text on the basis of the original, usually on sentence-level units (Paetzold and Specia, 2017).

The work of Nisioi et al. (2017) has brought a breakthrough in abstractive text simplification: they used a sequence-based model, originally designed for machine translation, using standard-language material as source text and simplified texts as target text – this allowed more complex automatic changes to take place that could greatly affect the syntactic structure of the sentence. Since then, numerous different attempts were made to better the existing TS methods. These mainly focus on lexical simplification (such as Zhao et al. (2022) or Sheang et al. (2022)), however some of them concentrate on paragraph-level or document-level simplification (for example, Trienes et al. (2022) successfully attempt both document- and paragraph-level simplification). However, what seems to be similar in most – although not all – of these attempts is the need for data, as a lot of these systems are fine-tuned on large parallel corpora.

2.2 Corpora

There are not many languages that possess parallel corpora consisting of standard language–simplified

Proceedings of Recent Advances in Natural Language Processing, pages 898–907
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_097

pairs. Kajiwara and Komachi (2016) name 7 languages for which at least one TS corpus has already been created (English, German, Spanish, Portuguese, Italian, Danish and Japanese). Since 2016 such corpora have been created for a few other languages e.g. for French (Grabar and Cardon, 2018) or Basque (Gonzalez-Dios et al., 2018) – Hungarian, however, is not among these languages.

3 Creating the corpus

Due to the limited financial and human resources available to us, as well as the lack of existing Hungarian parallel data, building an original corpus was out of the scope of this research.

Instead, following the already existing literature, such as Megna et al. (2021), we opted for the translation of an already existing English corpus. This obviously influences further studies on the corpus: since it does not consist of authentic Hungarian data, it cannot be used to determine e.g. the strategies that Hungarians use to simplify texts. However, assuming that the simplifications in the original English corpus are adequate and the translation is good enough, the resulting corpus can still be used to train simplification models on.

3.1 Choosing the corpus

We chose PWKP (Zhu et al., 2010) as the basis of our research. We, however, have also considered the other three most commonly used English simplification corpora: WikiSmall, WikiLarge and Newsela, but all of these corpora had downsides, that would have made the research considerably harder.

The WikiSmall and WikiLarge corpora were introduced in Zhang and Lapata (2017). These are tokenized corpora – however as modern transformer-based language models are trained on text in standard orthography,[1] a tokenized corpus is suboptimal for finetuning them.

The Newsela corpus, introduced in Xu et al. (2015), contains more than a thousand news articles with multiple levels of simplifications each. Unfortunately, the corpus is not publicly available, which would also prevent us from sharing the translation.

PWKP (Zhu et al., 2010), however, is readily available and is widely used (e.g. Omelianchuk et al. (2021); Vu et al. (2018); Zhang and Lapata (2017); Narayan and Gardent (2016, 2014)). The corpus was created by pairing more than 65,000 articles automatically from the English Wikipedia and the Simple English Wikipedia. From the article pairings more than 108,000 sentence pairs were extracted automatically. Of these, 205 and 100 sentence were set aside for validation and testing, respectively. It is important to note that the corpus consists of 1-to-n pairs, meaning that more than one simplified sentence can belong to one standard-language sentence.

Nonetheless, it has some downsides, too: as Xu et al. (2015) have shown, 17 % is not paired correctly, and in another 33 %, the "simple" sentences are not actually simpler than their standard language counterparts.

Another huge problem from the machine learning standpoint is that about 20,000 sentence pairs are duplicates, so the effective number of training instances is only about 88,000. Moreover, there is an overlap between the test and the training set, rendering the results reported on this set unreliable.

Still, despite all of these disadvantages, PWKP seemed to be the most optimal choice for our research. However, we tried to address some of its shortcomings prior to translation.

3.2 Improving the corpus's quality

Fixing all known issues with PWKP manually would have required an immense amount of work – and thus, financial resources. Lacking that, we employed a series of semi-automated steps to correct some of the most glaring (and easily fixable) problems.

3.2.1 Deduplication

First, the corpus was deduplicated. Sentence pairs were grouped by the original sentence, and of each group, only the first sentence was kept. With this step almost 20,000 sentence pairs were removed from the corpus.

Note that the method above does not take the simplified sentences into account and it filters a duplicate original even if the simplified sentences differ. Luckily, only about 1800 sentence pairs are affected by this issue; i.e. 9 % of the removed data. Because of this, and the generally low quality of the pairing (see 3.1), we decided to simply remove these pairs from the corpus. This also avoids the problem of bias that might emerge from having sentence pairs with the same original sentence in both the training and test splits.

[1] To the extent content creators adhere to it.

3.2.2 Clean-up

PWKP contains a lot of artifacts that probably stem from misparsing wiki markup or invalid markup in the source pages themselves. Some examples include empty brackets (`[ ]`, `( )`), list bullets converted to colons (`:`, `::`), URLs etc. We cleaned these up semi-automatically and deleted sentences that consisted solely of these artifacts.

3.2.3 Frequent simplifications

We also removed 3386 sentence pairs that each had the following structure: the standard language sentence states where a commune is located (e.g. "*Thiernu is a commune in the Aisne department in Picardie in northern France.*"), and the simplified version replaces the subject by *It* (e.g. "*It is found in the region Picardie in the Aisne department in the north of France.*")

Clearly, these sentences lack a contextualizing phrase (e.g. *Thiernu is a commune.*), through which *it* could be correctly interpreted. While we handle the general case of referential subjects in 3.2.5 differently, we decided to remove these sentence pairs from the corpus for two reasons. First, the simplified sentence is not simpler. Second, all 3386 sentences fall into roughly 13 different templates (with different region and department names); leaving them in would only have lead to overfitting in models trained on the corpus.

There are other frequent simplifications: in fact, about 2200 simplified sentences occur more than once. For some of them, all occurrences are valid; for others, only one has a matching standard pair and the rest are just pairing mistakes. Due to our limited resources, we did not pursue this path further, but filtering out the invalid pairs manually could significantly benefit the corpus.

3.2.4 Header removal

Working closely with the data made it clear that the automatic collection of the sentences was not completely without issues: if the sentence was the first in a Wikipedia subsection, the subsection title was also included:

(1) **Career** In 1905, Cortot formed a trio with Jacques Thibaud and Pablo Casals, which established itself as the leading piano trio of its era, and probably of any era.

The removal of these subsection titles was done in two parts. First, sentence pairs which did not come from the main text of the Wikipedia articles were removed completely. To identify such pairs, we checked if either the standard language sentence or any of the simplified sentences started with "*References*", "*Sources*", "*Notes*", "*Properties*", "*Bibliography*", "*Further reading*", "*See also*", "*External links*", "*External references*" or "*Other websites*", followed by a capital letter (which was the start of the actual sentence or the reference). With this simple, heuristic method about 650 sentence pairs were filtered from the corpus.

The remaining sentence pairs were cleared up with the help of the Wikimedia Dumps of February 2023 (2023). We filtered out the subsection titles from the dump and listed them in descending order of frequency. As the vast majority of these titles were single occurrences, we used the first 2000 subcategory titles from this list, except for *The*, *In*, *Out* and *President*, which are usually valid parts of the sentence and not subcategory titles.

Again, the filtering was applied to sentences that started with a subcategory title followed by a capitalised word; only this time, only the titles were removed. In total, 8704 sentences in 6500 sentence pairs were changed.

After these steps a total of 85,226 sentence pairs remained in the corpus.

3.2.5 Referential subjects

Even aside from the template sentences mentioned in 3.2.3, the corpus contained a relatively large number of sentence pairs in which the subject of the standard language sentence with a specific referent was replaced in the simplified sentence by the third person neutral singular pronoun *it*.

As mentioned in 3.2.3, the second sentence of such a pair is not a valid simplification of the first due to lack of context. Therefore in sentence pairs where the standard sentence begins with a noun + *is* construction and the simplified version begins with the construction *It is*, the word *it* has been replaced with the noun in the standard sentence.

At this time, we did not attempt to resolve *it* in more complicated sentences, or other referential subjects, as handling them in each case would require manual supervision or semi-automatic scripts based on dependency parsing or machine learning. We leave this task for future work.

3.2.6 Sentence swapping

Another common phenomenon in the corpus is that the simplified sentences were longer and contained more information than their standard language counterparts (this problem has also been previously raised by Xu et al. (2015)). In some cases this could mean that the simplified sentence is longer because it explains a hard-to-understand concept in the standard-language sentence (see Shardlow (2014)). However, after examining a few of the affected sentence pairs, it seemed that this was not usually the case in PWKP.

Therefore, we decided to create a version of the corpus where the standard-language and the simplified sentences are swapped if the latter was longer than the former by at least 20 characters. The limit was introduced to allow minor stylistic differences.

This affected a total of 5057 sentences. We refer to this version of the corpus as SWAPPED.

While splitting the standard sentence into multiple sentences is a valid simplification technique, based on a cursory glance at the examples, we conjectured that in PWKP, such pairs are mostly pairing artifacts. To test this hypothesis, we created another version of the corpus, called SWAPPED (SINGLE ONLY). This version has 79,953 sentence pairs, 5273 less than the full corpus.

3.2.7 Train–validation–test split

The corpus was split into train, validation and test splits, approximately 90 %–5 %–5 %. We ended up with 76,801–4188–4237 sentence pairs in the three splits, respectively. This allows for a more robust evaluation than PWKP / WikiSmall's 205-long validation and 100-long test sets.

The splits are the same across all corpus versions.

3.3 Translating the corpus

Due to the size of the corpus, manual translation was not a feasible solution, so we opted for machine translation. We experimented with both Opus-MT's (Tiedemann and Thottingal, 2020) `en-hu` model from the Hugging Face Hub (2023) and DeepL (2023).

An evaluation of the translation was conducted by the first author and an independent annotator. First, we calculated the BLEU-score (Papineni et al., 2002) for each sentence with the help of NLTK's BLEU-calculator (Bird et al., 2009). As there is no gold-standard translation for this corpus, the two translations were compared to each other: with using DeepL as a gold standard, we were able to get higher scores, so we used this distribution for the evaluation.

We randomly selected 5 sentence pairs from each BLEU-percentile, and evaluated their translations on a Likert-scale ranging from 1 to 5 in the following aspects:

- **Meaning preservation:** Checking wheteher the Hungarian sentence means the same as the original.
- **Grammaticality:** Evaluating if the translation was grammatically correct and whether it sounded ”natural”.
- **Identical word use for coreferential nouns:** Checking whether when the same English word appeared multiple times in a pair of sentences, it was translated in the same way in the Hungarian translation, or the translator used synonyms. (see Section 4.1.3 for an example).

The results can be found in Table 1. Although both systems performed adequately, both annotators agreed that DeepL provided a better translation. Therefore we used this translation in our research.

	Meaning pres.	**Grammaticality**	**Indentical w. use**
	OpusMT		
1^{st} ann.	4.04	4.37	4.75
2^{nd} ann.	4.32	4.56	4.45
	DeepL		
1^{st} ann.	4.32	4.69	4.92
2^{nd} ann.	4.71	4.86	4.57

Table 1: The scores of the two translations in meaning preservation, grammaticality and identical word use.

4 Evaluation

We evaluated the translated corpus in two different ways. First, we trained a seq2seq model on both the English and the Hungarian corpora and compared the results. Second, we conducted a questionnaire study in order to include the human perspective in the evaluation.

4.1 Seq2seq models

4.1.1 Setup

For the model-based comparison, we trained encoder-decoder models with the `transformers` (Wolf et al., 2020) library. We used the code published with Barta et al. (2023), originally for text summarization, with slight modifications, such as using SARI (Xu et al., 2016) as the evaluation metric. The models were trained with the default parameters on an A100 GPU.

An encoder-decoder model in `transformers` is a sequence-to-sequence model that initializes its encoder and/or decoder from pretrained models. There are two ways to achieve a fair comparison between the English and Hungarian models: use native pretrained models with the same model architecture and parameter budget for both languages, or initialize the weights from a multilingual model that supports both languages.

At the time we ran our experiments, only a few Hungarian models were available, each of them a variant of the BERT architecture (Devlin et al., 2019). The then-best model was the cased BERT-Base model `huBERT` (Nemeskey, 2021) with 110M parameters. Our Hungarian seq2seq model uses `huBERT` to initialize both the encoder and the decoder. On the English side, `bert-base-cased` was used.

Of the multilingual models, we experimented with mT5 (Xue et al., 2021), as the `base` model previously performed comparably to, or even slightly better than, `huBERT` for summarization in Hungarian (Barta et al., 2023). Unfortunately, on our much smaller simplification dataset, mT5 failed to achieve a meaningful SARI score. Hence, we only report results for the native models.

We trained an English model on the cleaned PWKP and three Hungarian models: one each on the translated corpus and its two swapped versions.

4.1.2 Results

Table 2 presents the SARI scores achieved by the English and Hungarian seq2seq models. The upper half of the table compares the performance of the English and Hungarian models; the lower half shows the effect of training on the two swapped versions of the corpus. We used EASSE (Alva-Manchego et al., 2019) to compute the SARI scores.

The models were evaluated on the test split of our corpus (3rd column), as well as on ASSET (Alva-Manchego et al., 2020). We translated ASSET to Hungarian with DeepL, but did not manually review the product, so the Hungarian results on that set should be taken with a pinch of salt. Similarly, scores on the test set of the corpus cannot be directly compared to numbers reported on the official PWKP test set, which is only available in tokenized format, although they are probably much more robust (see 3.2.7). The results of the English model on ASSET (bold) can be reliably used to compare our model to those in the literature.

With that said, our English model attains a competitive score on PWKP, even though no external training corpora were used; the best model we know of scores at 44.67 (Omelianchuk et al., 2021), and the second best at 32.35 (Dong et al., 2019) (results from Ruder (2023)).

4.1.3 English vs Hungarian

It can be seen that the performance of the models trained on the cleaned PWKP are slightly higher than on its Hungarian translation. Since there are many free parameters (the translation, the original pretrained models, the training process itself, Hungarian being agglutinative, etc.), it is hard to pinpoint the exact cause. We theoretize that there are two main reasons for the decreased SARI score.

The first one is inconsistencies in the translation of source and target sentences. As an example, there are several pairs in which the English word "*hill*" is translated as "*hegy*" ("*mountain*") in the source and as "*domb*" ("*hill*") in the target sentence. If the model predicts "*hegy*", it will be penalized for a perfectly valid output.

The second reason is that word n-grams work better for analytic languages, such as English, and peculiarities in Hungarian orthography and morphology are thus penalized by SARI. Agglutination and the preference for closed compounds mean that Hungarian has a higher morpheme-to-word ratio, and so a higher probability of a word being "wrong". Also, the EASSE implementation gives out higher scores for longer sentences, which works against Hungarian for the same reason.

4.1.4 Corpus versions

As for the different corpus versions, SWAPPED outperforms the regular corpus by 1 point. This implies that the swapped version is easier to learn, suggesting that in at least some of the swapped sentence pairs, the simplified sentence originally was actually more complex.

The SWAPPED (SINGLE ONLY) version performs even slightly better on the test set, but not on ASSET. This is because while it has an even more consistent training corpus, the task it actually trains for, 1-to-1 sentence simplification, is simpler, and cannot handle the 1-to-N examples in ASSET.

Based on these results, we recommend the SWAPPED version of the corpus for training, even

Language	Corpus version	SARI	SARI on ASSET
English	Final	42.32	**38.05**
Hungarian	Final	38.75	35.37
Hungarian	SWAPPED	40.06	36.82
Hungarian	SWAPPED (SINGLE ONLY)	40.41	36.61

Table 2: SARI scores achieved by the seq2seq models trained on the final corpora and on the two modified Hungarian versions.

	Inform. retention	Grammaticality	Degree of simpl.
Mean	2.99	3.69	2.82
Median	3	4	3
Highest mean	4.5	4.71	4.25
Lowest mean	1.43	1.46	1.68
St. dev.	1.50	1.47	1.42
Cohen's kappa	0.11	0.16	0.07

Table 3: The scores of the human evaluation.

though the human evaluation seems to suggest to use the original version (see 4.2.1).

4.2 Human evaluation

4.2.1 Choosing the corpus version to use

In order to be able to conduct a questionnaire study, first we needed to evaluate the three corpus versions. As no clearly best performing model could be deduced from the automatic scores (see 4.1.2) we decided to include human annotators in the evaluation. First, we randomly selected 20 sentences from the test set of SWAPPED (SINGLE ONLY), then included these sentences from SWAPPED's and the original corpus' test set in our evaluation system. Then two independent annotators and the first author evaluated the sentences by choosing the one they thought was the best simplification. All three annotators preferred the original (non-swapped) version. The inter-annotator agreement based on Cohen's kappa was 0.77.[2] We therefore proceeded with this version.

4.2.2 The questionnaire

For the questionnaire, we generated a 50-sentence-long sample from the test dataset, and from this we chose 25 sentences whose original, standard-language version seemed the most intelligible and "authentic" in Hungarian and whose simplification differed from the standard-language sentence, as well as five sentences where the simplified version was the same as the original. The questionnaire consisted of three sections. After the respondents agreed to a consent form, they proceeded to the second section, where we used the 25 differing simplifications. The respondents were asked to give a score on a Likert-scale ranging from one to five, for the following three aspects (based on Alva-Manchego et al. (2020)):

- The simplified sentence adequately expresses the original meaning, possibly omitting the least important information.
- The simplified sentence seems to be an authentic Hungarian text and does not contain any grammatical errors.
- The simplified sentence is easier to understand than the original sentence.

The respondents saw the sentences in a randomized order within the sections of the questionnaire.

In the third section, the respondents were asked whether the sentences which were not simplified by the model could have been simlified more. This section, however, has produced indecisive results, mostly because of the small amount of data that has been seen by the participants. Therefore we decided not to discuss it here, but rather conduct a specific research on this topic in the future.

4.2.3 Results

A total of 27 people completed the questionnaire between 08.04.2023 and 13.04.2023. The respondents were aged between 22 and 60 years, 8 men and 19 women. It is important to note that this questionnaire is not representative, it serves merely for us to gain some insight into the real-life usability of the corpus.

Asking laymen to rate the outputs of the model was a conscious choice from our side: while filling

[2] We took the mean of the pairwise scores. Cohen's kappa was calculated using Scikit-learn (Pedregosa et al., 2011).

Name	Description	License
ELTE-DH/PWKP_cleaned ELTE-DH/huPWKP	The English corpus The Hungarian corpus	CC BY-SA 4.0
ELTE-DH/simplification-pwkp-en ELTE-DH/simplification-pwkp-hu	The English model The Hungarian model	Apache 2.0

Table 4: Availability of the datasets and models on the Hugging Face Hub.

out the questionnaire we wanted to activate the participants' intuitive concept of SIMPLIFIED TEXT, that is probably possessed by most of the prototypical adult population, even if it differs by each person. We decided not to give the participants any guidelines about what a SIMPLIFIED TEXT is, because we wanted to know whether they really believed the model output to be simpler, and not them solving a "sorting task" according to what we or the literature considers simplified.

Table 3 represents the results of the human evaluation. The model performs best in terms of grammaticality, with a mean of 3.69 and a median of 4. It should be noted that standard deviation is relatively high and inter-annotator agreement is relatively low for all three aspects.This suggests that the intuitive concept of SIMPLIFIED TEXT varies greatly by each person.

The model produces a mean of around 3 and the same median in terms of information retention and increase in the degree of simplicity. It is worth noting that for some sentences the model can achieve a mean of 4.5 or above for information retention and grammaticality, and a mean of 4.25 for the increase in the degree of simplicity. On the other end of the spectrum are sentences with average scores of around 1.5. In these cases, the model either returns factually wrong information, or renders the simplified sentence unintelligible.

To summarise, the results of the questionnaire show that, although the responses have a relatively large standard deviation and an exceptionally low inter-annotator agreement score, the model can produce averages of around 3 for all aspects of the survey. It is worth noting that the mediocre scores from human annotation stand in contrast to the competitiveness of the automatic metrics (4.1.2). This seems to validate the criticism SARI receives for its low accuracy and correlation with human judgement (Alva-Manchego et al., 2021).

4.2.4 Availability

Both the corpora and the models are available in the Hugging Face Hub under the organization ELTE-DH. See Table 4 for details. The code is on GitHub[3].

5 Conclusion

In this paper, we have introduced huPWKP, a Hungarian translation of the PWKP corpus. The translation was performed automatically, based on a cleaned version of PWKP, which we also publish.

The translation was evaluated both manually and automatically: the latter by training a seq2seq simplification models initialized from native BERT-Base checkpoints for both languages. The English and Hungarian models performed similarly, at around the best SARI score reported by other models on the official PWKP test set.

The manual evaluation was carried out using a questionnaire survey. It shows that the model can produce averages of around 3 for meaning preservation and increasing the degree of simplicity, and 3.7 for grammaticality.

While some of the most glaring issues in PWKP have been addressed, the corpus could be improved further by tackling the more involved cases of referential subjects and simplified sentence duplication. We plan to incorporate such changes in future releases of the corpus.

Acknowledgements

Research was funded by programs of the Ministry for Culture and Innovation, Hungary via the National Research, Development and Innovation Fund. The first author was supported by the ÚNKP-22-2 New National Excellence Program; the second author was supported by the National Laboratory for Digital Heritage.

[3] https://github.com/DavidNemeskey/PWKP_hun

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Konor Nyist, and Judit Ács. 2023. HunSum-1: an Abstractive Summarization Dataset for Hungarian. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*, pages 231–243, Szeged, Magyarország. Szegedi Tudományegyetem, Informatikai Intézet.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc., Sebastopol, CA.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19 – 26. ACM; New York.

DeepL. 2023. https://www.deepl.com/. (Online; accessed 10-May-2023).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz De Ilarraza. 2018. The corpus of basque simplified texts (cbst). *Lang. Resour. Eval.*, 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.

Hugging Face Hub. 2023. Online; accessed 2023-05-16. [link].

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan. The COLING 2016 Organizing Committee.

Christiane Maaß and Sergio Hernandez Garrido. 2020. Easy and plain language in audiovisual translation. In Christiane Maaß Silvia Hansen-Schirra, editor, *Easy Language Research: Text and User Perspectives*, 1 edition, volume 2 of *Easy – Plain – Accessible*, pages 131–161. Frank & Timme.

Christiane Maaß and Isabel Rink. 2020. Scenarios for easy language translation: How to produce accessible content for users with diverse needs. In Christiane Maaß Silvia Hansen-Schirra, editor, *Easy Language Research: Text and User Perspectives*, 1 edition, volume 2 of *Easy – Plain – Accessible*, pages 41–56. Frank & Timme.

Angelo Megna, Daniele Schicchi, Giosuè Lo Bosco, and Giovanni Pilato. 2021. A controllable text simplification system for the italian language. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 191–194. IEEE.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, pages 435–445.

Shashi Narayan and Claire Gardent. 2016. Unsupervised sentence simplification using deep semantics. In *Proceedings of the 9th International Natural Language Generation conference*, pages 111–120, Edinburgh, UK. Association for Computational Linguistics.

Dávid Márk Nemeskey. 2021. Introducing `huBERT`. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021)*, page TBA, Szeged.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text simplification by tagging. *CoRR*, abs/2103.05070.

Gustavo Paetzold. 2015. Reliable lexical simplification for non-native speakers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 9–16, Denver, Colorado. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sebastian Ruder. 2023. Nlp-progress. Online; accessed 2023-05-16.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, 4(1).

Kim Cheng Sheang, Daniel Ferrés, and Horacio Saggion. 2022. Controllable lexical simplification for English. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 199–206, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

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

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Gábor Tolcsvai Nagy. 2017. Bevezetés. In Gábor Tolcsvai Nagy, editor, *Nyelvtan*, pages 23–71. Osiris Kiadó, Budapest, Hungary.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: Towards a new text simplification dataset. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 19–27, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Szilárd Tátrai. 2017. Pragmatika. In Gábor Tolcsvai Nagy, editor, *Nyelvtan*, chapter Pragmatika, pages 899–1058. Osiris Kiadó, Budapest, Hungary.

Szilárd Tátrai. 2020. On the perspectival nature and the metapragmatic reflectiveness of contextualization. *Studia Linguistica Hungarica*, 32:109–120.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85, New Orleans, Louisiana. Association for Computational Linguistics.

Wikimedia Dumps of February 2023. 2023. https://dumps.wikimedia.org/. (Online; accessed 10-May-2023).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Sanqiang Zhao, Rui Meng, Hui Su, and Daqing He. 2022. Divide-and-conquer text simplification by scalable data enhancement. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 166–172, Abu Dhabi, United Arab Emirates (Virtual). 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 (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.