How to Simplify Law Automatically? A Study on South Korean Legislation and Its Simplified Version

Stefanie Urchs[a], Akshaya Muralidharan[b] and Florian Matthes[c]
Chair of Software Engineering for Business Information System, Technical University of Munich, Garching, Germany

Keywords: Text Simplification, Legal Tech, South Korean Legilsation, Natural Laguage Processing.

Abstract: People with a low literacy level have problems understanding complex texts. Especially legal texts can be challenging. Automatic Text Simplification (TS) can help to make the legal text more accessible. However, most TS research is based on Wikipedia articles and newspaper articles. To be able to use automatic TS on the legal text we have to understand what constitutes simple legal text. Therefore, we examine the English translation of South Korean legislation and its official simplification. Subsequently, we use state of the art TS models on the legislation text. The models simplify the text only quantitatively lacking in retaining the context of the original text.

1 INTRODUCTION

Text Simplification (TS) is a Natural Language Processing task that aims to automatise the tedious work attached to manual text simplification. Al-Thanyyan and Azmi (Al-Thanyyan and Azmi, 2021) define TS as "*[...] the process of reducing the linguistic complexity of a text to improve its understandability and readability, while still maintaining its original information content and meaning.*". Generally, TS includes the following tasks: simplifying the form and content of a text, reducing redundancies and not needed information from a text and summarisation of a text (Siddharthan, 2015). TS opens documents up to people with a lower literacy level. A study by the "Program for the International Assessment of Adult Competencies"[1] showed that 32% of U. S. adults are able to read and extract information from simple texts but are overwhelmed by more difficult tasks. Being unable to understand more complex texts is especially critical when legal texts are involved. If a person does not understand the law they will have problems adhering to it. Furthermore, people who are not able to understand the rights, granted to them by the law, are not fully able to use them. Rubab et al. (Rubab et al., 2020) show that simplifying the wordings of law increases the comprehensibility of the texts for law professionals, law teachers and laymen alike, making it more accessible.

However, simplifying legal text is no easy task, as all the meaning of the source text has to be preserved. Therefore, TS in the legal field is in need of carefully constructed training data and metrics to detect if a text is successfully simplified while retaining all crucial information. Unfortunately, training corpora for normal TS task are already scarce. To the authors' knowledge currently, no corpus of normal legislation aligned to a simplified version of this legislation exists. Hence research on automatic simplification of legislation is scarce too. Normal TS measures do not necessarily apply to the legal field and its sensitivity to semantics in text. Consequently, TS models that are optimised towards common TS measures do not necessarily apply to the legal field.

The South Korean government provides its citizens with an online version of their legislation[2] (hereafter referred to as E-Law) and a translation into English. Moreover, the South Korean government provides a simplification of its legislation[3] (hereafter referred to as Easy-Law) which is translated into twelve different languages, among them English. We align Easy-Law with its corresponding E-Law parts. Subsequently, we analyse how an easy to read legal text, that contains all necessary information, compares to its standard law counterpart. Therefore, we examine what constitutes easy legal language. Additionally,

[a] https://orcid.org/0000-0002-1118-4330
[b] https://orcid.org/0000-0003-2495-4717
[c] https://orcid.org/0000-0002-6667-5452

[1] https://nces.ed.gov/surveys/piaac/ accessed on 2021-10-15

[2] https://elaw.klri.re.kr/ accessed on 2021-10-28

[3] https://www.easylaw.go.kr accessed on 2021-10-28

Urchs, S., Muralidharan, A. and Matthes, F.
How to Simplify Law Automatically? A Study on South Korean Legislation and Its Simplified Version.
DOI: 10.5220/0010894000003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 697-704
ISBN: 978-989-758-547-0; ISSN: 2184-433X

we use state of the art TS models on E-Law and compare them to the Easy-Law baseline. Due to copyright issues, we can not publish the parallel dataset.

2 RELATED WORK

This paper analyses the simplification of South Korean legislation. Therefore, this chapter introduces related work from the fields of text simplification (TS), TS corpora and metrics to analyse the simplicity of the text.

2.1 Text Simplification

In one of the earliest works on TS Chandrasekar and Srinivas (Chandrasekar and Srinivas, 1997) approach the task on a sentence level. The authors propose a two-step process: first, the input is **analysed** and its structure is described in a dependency tree with the Tree Adjoining Grammar (TAG). In the second step this description is used to **transform** the input into a simpler version, basically flattening the dependency trees. The transformation rules are automatically learned on an aligned corpus of sentences and their manual simplification. Dars (Dras, 1999) addresses TS on a text basis and defines the problem as "reluctant paraphrase". His base assumption is that an author produced a text that contains everything they want to express. Due to external factors like length constraints, readability issues or style guides authors have to paraphrase their text. Therefore, a "best solution" for the paraphrase, one that adheres best to the factors and still retains the most textual meaning, exist. Like Chandrasekar and Srinivas, Dars uses the TAG to represent the input. He then proceeds to map between two TAG grammars, combined with integer programming, to adhere to the previously mentioned factors. At the same time, he tries to paraphrase as little as possible. Both approaches concentrate on the syntactic simplification of text. Other researchers interpret TS as a monolingual translation task. For example, Wubben et al. (Wubben et al., 2012) integrate a re-ranking heuristic into the translation task and train their system on a corpus that aligns Wikipedia articles with their corresponding Simple Wikipedia articles. A third way to approach TS is to lexically simplify the text at hand. One of the current state of the art TS systems by Qiang et al. (Qiang et al., 2021) is a lexical simplification model based on BERT. The system pipeline consists of three steps: first identifying complex words in the text, second generating words to replace these complex words and third filtering the generated words and ranking them according to the best replacement fit. The resulting model is called LSBert. Martin et al. (Martin et al., 2020a) approach TS as a mixture of simplifying the structure and grammar of a text. Their TS system ACCESS enables the user to tune the text simplification to the following attributes: sentence length or amount of compression, level of paraphrasing and reduction of lexical and syntactic complexity. A Sequence-to-Sequence model is then conditioned on these attributes. In later work, Martin et al. (Martin et al., 2020b) extend their ACCESS model by the MUSS method. This method trains controllable TS models in an unsupervised way, by utilising sentence-level paraphrase data. Therefore, excluding the need for labelled training data.

At the time this paper is written LSBert and ACCESS in combination with MUSS can be considered state of the art. Therefore, this paper uses these models on the E-Law corpus and compares them to the Easy-Law baseline, investigating their compatibility to legal language.

2.2 Text Simplification Corpora

TS literature uses parallel text corpora. These are corpora where a complex source is aligned with its simplification. In the field of English TS corpora mostly use a combination of the English Wikipedia with Simple Wikipedia. For example Coster and Kauchak (Coster and Kauchak, 2011) aligned sentences from English Wikipedia with their simplification from Simple Wikipedia resulting in 137 thousand sentence pairs. Zhu et al. (Zhu et al., 2010) do the pairing of the Wikipedias on an article basis and obtain 65,133 paired articles. A third paired Wikipedia corpus is generated by Kajiwara and Komachi (Kajiwara and Komachi, 2016) by performing a match on the article titles. Their corpus consists of 126,725 article pairs. Xu et al. (Xu et al., 2015) argue that the combination of Wikipedias is insufficient as TS resource and propose the Newsela corpus. This corpus consists of 1,130 news articles, that are professionally rewritten in four different simplification grades. The OneStopEnglish corpus by Vajjala and Lučić(Vajjala and Lučić, 2018) aligns 189 texts over three reading levels, resulting in 567 texts that are aligned at text and sentence level.

Unfortunately, none of these TS corpora covers legal text.

2.3 Text Simplification Metrics

To evaluate TS systems metrics have to be used. Dars (Dras, 1999) names three formulas as commonly known readability metrics for adult reading material:

- Flesch Reading Ease
  reading ease $= 206.835 - 0.846w_t - 1.015s_t$
  Subtracts the average number of words per sentence(s_t) from the number of syllables per 100 words (w_t), the constants norm the reading ease to a grade based number.
- Dale-Chall Formula
  $x_{c50} = 0.1579x_1 + 0.0496x_2 + 3.6365$
  x_{c50} is the reading grade score of a student who knows half of the answers for questions about a text paragraph. It is calculated by adding the percentage of words outside a predefined easy to read list (x_1) to the average sentence length in words (x_2).
- FOG Index
  reading level $= 0.4x_1 + 0.4x_2$
  Adds the average sentence length x_1 to the percentage of words with more than three syllables x_2.

Al-Thanyyan and Azmi (Al-Thanyyan and Azmi, 2021) extend this list by the following formulas:

- SMOG grading score
  SMOG $= 3 + \sqrt{x_p}$
  Looks at the average of polysyllable words (x_p) (words with more than two syllables) in 30-sentence-long text segments.
- Flesch-Kincaid Grade Level Index (FKGL)
  FKGL $= 0.39x_1 + 11.8x_2 - 15.59$
  Combines the average length of a sentence (x_1) and the average number of syllables per word (x_2) and uses the constant to calculate a grade level.

This paper uses these metrics to understand what constitutes easy legal language, using the Easy-Law corpus. Furthermore, the metrics are used to analyse the text generated by LSBert and MUSS in combination with ACCESS.

3 ENGLISH KOREAN LEGISLATION

The Korean Legislation Research Institute (KLRI)[4] provides a global platform for legislative research. The KLRI cooperates with the Korean Law Translation Center (KLT)[5] to offer an English translation of current Korean statutes (referred to as E-Law in the remainder of this paper). KLRI and KLT aim to promote understanding of Korean Legislation for a global audience and non-Korean speaking residents. The Korean Ministry of Government Legislation[6] expands this online resource with a website[7] that explains Koran legislation in an easy to understand way. The website (further referred to as Easy-Law) provides information in Korean and twelve different translations, among them English.

This chapter introduces the E-Law and Easy-Law websites and a parallel corpus we created by aligning acts from these two websites.

3.1 E-Law

The E-Law website provides a translation of current legislation and historic legislation. We concentrate on the current legislation, where 2,230 statutes and regulations are provided. Users can search the website in different ways:

- By Legal Field
  Twelve different fields, from the constitution to foreign affairs, are offered.
- By Ministry
  Users can search for legislation for one of 23 Ministries, one of 17 Agencies or Administration, one of six commissions, one of six constitutional institutions or one of five miscellaneous options.
- By Subject
  The website offers 16 different subjects including family law, launching business, employment/labor, culture and environment/energy.
- All Overview of all legislation with keyword search. The sidebar enables searching by legal field, ministry and subject as well as validity (current and historical legislation) and type of statutes and regulations.
- Recently Translated
  Listing of all legislation ordered by date of the translation.
- Most viewed
  Listing of all legislation ordered by the number of total hits either by week, month or year.
- Law in News
  List of all legislation that is featured in the news.

In addition to the translation of legislation, the KLT offers a glossary of legal terms which lists certain keywords and provides a full-text search for these in all statutes. Furthermore, the Korean legislative system and legislative procedures are explained.

[4]https://www.klri.re.kr/eng.do accessed on 2021-10-30
[5]https://elaw.klri.re.kr/eng_service/introduction.do accessed on 2021-10-30
[6]https://www.moleg.go.kr/english/ accessed on 2021-10-30
[7]https://www.easylaw.go.kr/CSM/SubMainCmd.laf?langCd=700101 accessed on 2021-10-30

3.2 Easy-Law

In contrast to the E-Law website, the Easy-Law website is less structured. Therefore, all information can be accessed directly from the front-page without navigating through complex menus. Users can select one of sixteen categories. Each category contains one or more subcategories, which contain one or more key contents. By clicking the key contents the user arrives at a text page where the first subcategory of the chosen key content is explained in an easy to understand way. Users can select different subcategories and key contents in a side-bar.

The explanation starts with the most general concepts and expands on them or delivers additional context. Furthermore, references to the explained legislation are provided.

3.3 Parallel Corpus

To understand what differentiates *easy* law from *normal* law the legal information in the subcategories of the key facts (further referred to as simple articles) provided on the Easy-Law website is aligned with one or more corresponding sentences in the statutes provided by the E-Law website. More precisely: given a list of simple articles S_1, S_2, S_3, ..., S_n that consists of stand alone legal information $s_{x,1}, s_{x,2}, s_{x,3}$, ..., $s_{x,m}$ and a list of *normal* acts A_1, A_2, A_3, ..., A_o consisting of sentences $a_{y,1}, a_{y,2}, a_{y,3}$, ..., $a_{y,p}$. A mapping of the form

$$
\begin{gathered}
s_{x,1}, s_{x,2}, s_{x,3}, \ldots, s_{x,m} \\
\Updownarrow \\
a_{y,1}, a_{y,2}, a_{y,3}, \ldots, a_{y,p} \\
\ldots \\
a_{z,1}, a_{z,2}, a_{z,3}, \ldots, a_{z,q}
\end{gathered}
$$

is generated, such that the information contained on both sides of the mapping is the same. n, x, m, o, y, p, z, q indicate control variables. To create this mapping the following pipeline is executed: identification of stand-alone information in Easy-Law, referencing to corresponding E-Law act and extraction of the referenced sentence(s) from the E-Law act.

Every paragraph in the key facts document of the Easy-Law website that contains a reference to a legal act is considered a stand-alone piece of legal information. The reference is extracted with a simple regex/-pattern search approach that searches the text for all mentions of acts.

It is not possible to directly map the references of the Easy-Law website to legislation on the E-Law website, because of different providers and different translations of the legislation names. Thus, BERT word embeddings are used to find a match with a minimal cosine distance between the act names.

The reference from the Easy-Law is mapped to the full act from the E-Law. This means, that the E-Law side of the mapping can contain more information than the Easy-Law side. However, to analyse the textual differences between *easy* law text and *normal* law text this additional information on the *normal* side is negligible.

717 simple articles are mapped to 2,183 structured acts from the E-Law website leading to a parallel corpus of 922 aligned samples.

4 ANALYSIS OF ENGLISH KOREAN LEGISLATION

To understand the textual difference between *easy* and *normal* English Korean legislation both sides of the parallel corpus are analysed. At first simple count based analysis are performed, in a second step reading formulas introduced in chapter 2.3 are used and thirdly structural features are examined. In the end the results are discussed.

4.1 Count Based Analysis

For each sample the following count based metrics are processed:

- number of syllables
- number of words (excluding punctuation)
- number of sentences
- number of characters (including punctuation)
- number of letters (excluding punctuation)
- number of polysyllables (words with a syllable count greater or equal to three)
- number of monosyllables

Resulting in the mean values shown in table 1. The scores of simple law are always lower than the scores of normal law. Interestingly, simple law uses on average about half of the measure that is used for normal law. Polysyllables are generally scarce which is a characteristic of the English language.

However, these results have to be interpreted as a trend. As mentioned in subsection 3.3 the *normal* side of the parallel corpus can contain more information and therefore more sentences.

Table 1: Count based analysis of Korean legislation, all values represent the mean over all samples.

	simple	normal
syllables	139	271
words	81	164
sentences	1	3
characters	424	849
letters	411	812
polysyllables	17	30
monosyllables	48	100

4.2 Readability Scores

The following readability scores are computed on each sample of the parallel corpus, using the python library textstats[8]:

- Flesch Reading Ease
  The higher the score the more easy a text is to read, the highest possible score is 121.22, negative values are valid.
- Dale Chall Formula
  The higher the score the more complicated a text is to read. Values above 9.0 indicate that readers need at least a college education to understand the text.
- FOG Index
  Grade based score, that indicated which level of school education a reader needs to understand the text. E. g. a score of 9.3 indicated that readers with a ninth-grade education could understand the text.
- Flesch-Kincaid Grade Level Index
  Grade based score, that indicated which level of school education a reader needs to understand the text. E. g. a score of 9.3 indicated that readers with a ninth-grade education could understand the text.

The SMOG grading score is excluded, because it is normed for text sequences of thirty sentences. As discussed in subsection 4.1 the samples in the normal part of the parallel corpus average about three sentences.

Table 2 shows the result of the calculation. According to readability scores simple law is harder to read than normal law. Additionally, all scores indicate that both kinds of law are hard to read.

4.3 Structural Features

Two kinds of structural features are examined for each kind of law: the grammatical structure, represented as

[8]https://pypi.org/project/textstat/

Table 2: Calculation of readability scores for Korean legislation, all values represent the mean over all samples.

	simple	normal
Flesch Reading Ease	-3.36	6.82
Dale Chall Formula	12.06	11.69
FOG Index	31.73	29.27
Flesch-Kincaid Grade Level Index	29.69	27.53

pares tree expansion at a depth of level three and the number of references to other articles.

Kauchak et al. (Kauchak et al., 2017) discuss that parse trees at level two are too generic for in-depth analysis of the grammatical structure because only one expansion is done. This first expansion mostly consists of a noun phrase in combination with a verb phrase. Expansions after the third level become very specific to the sentence at hand, complicating generalisations over the texts. Table 3 shows the five most frequent grammatical structures in simple and normal. The following tags are used:

- DT - determiner
- IN - preposition/complementiser
- LST - list item marker
- MD - modal
- NN - singular noun
- NP - noun phrase
- PP - prepositional complement
- S - sentence
- SBAR - subordinate clause
- VBN - past participle
- VBZ - 3rd person singular present tense verb
- VP - verb phrase
- WHNP - wh-noun phrase

Both, normal and simple law sentences, mostly follow the grammatical structure of S [NP [NP PP] VP [MD VP]]. An example of this structure is “[(An individual [NP]) (under influence [PP]) (may [MD]) (not drive [VP]).]”. In the second rank grammatically structures begin to drift apart. Normal law favours noun phrases without whole sentence constructs in the second rank and later on rank four lists of singular nouns. Simple law mostly adheres to full sentences with noun phrases followed by verb phrases and favours modals verbs in verb phrases. An in-depth analysis of modal verbs shows that 69% of simple sentences use modal verbs and only 21% of normal sentences. In third rank, easy law also abandons the conventional sentence construct in favour of a noun phrase.

Table 3: Five most frequent grammatical structures of simple and normal law.

rank	simple	normal
01	S [NP [NP PP] VP [MD VP]]	S [NP [NP PP] VP [MD VP]]
02	S [NP [NP SBAR] VP [MD VP]]	NP [NP [DT NN] SBAR [WHNP S]]
03	NP [NP PP [IN NP]]	S [VBN PP [IN NP]]
04	S [NP [NP VP] VP [MD VP]]	LST [NN]
05	S [NP [DT NN] VP [MD VP]]	S [NP [NP NP] VP [VBZ NP]]

Going one level deeper into the parse tree enables the detection of passive voice. Normal law uses passive voice in 43% of the parsed sentences and simple law only 27%.

Law often references other parts of the law in one article. The frequency of these references is an interesting indicator of the readability of the text. Two percent of simple law and twelve percent of normal law reference to other articles. Though, if normal law references other articles it references multiple ones whereas simple law only references a few.

4.4 Discussion

Simple law tends to use fewer words and characters to explain concepts, indicating a more direct communication style. The use of readability formulas does not lead to useful results. These metrics mostly try to predict the school education, in other words grade level, a reader needs to understand the text, but legislation addresses complex topics that are above the knowledge of an average school child. These readability formulas perform well on average adult reading material. However, to perform well on legal text these formulas need domain-specific adjustments. These adjustments have to be determined by linguistic experts. The structural analysis indicates that increased usage of modal verbs and full sentences as well as the reduced usage of passive voice increase the readability of legal text. Furthermore, excessive referencing to other acts decreases the readability of a law.

Therefore, readability measures that fit legal text need to focus on shortness of text, while using complete sentences and favouring modal words. Moreover, references in stead of direct explanations should be penalised.

5 AUTOMATIC TEXT SIMPLIFICATION

Automatic Text Simplification (TS) offers the opportunity to make the legislative text more accessible. Manually reformulating laws, regulations and announcements is a tedious and labour intensive task. This chapter investigates how two state of the art TS models perform on legal text by using the parallel corpus described in subsection 3.3.

Both models are used in the pre-trained versions. The *easy* law part of the corpus is used as ground truth for evaluating the results.

5.1 Models

LSBert (Qiang et al., 2021) specialises in lexical simplification. At first complex words are identified and a list of possible replacements is generated. This list is then ranked and the most suitable replacement chosen. The automatic simplification is conducted with the pre-trained model published by Qiang on GitHub[9]. Due to the sequential lookup, the model performs and limited resources it was only possible to generate simplifications for 1500 lines of text of the *normal* law part of the parallel corpus.

The second model ACCESS in combination with MUSS (Martin et al., 2020a) is more sophisticated and paraphrases the input instead of just exchanging words. For this experiment, the pre-trained version provided on GitHub[10] by Facebook research is used. For comparability reasons, the same 1500 lines of the *normal* law of the parallel corpus are simplified.

5.2 Evaluation and Discussion

As discussed in subsection 4.4, conventional readability formulas are insufficient to express the readability of simplified law. Therefore, the evaluation concentrates on count-based measures as shown in table 4.

Overall both models seem to improve the readability of the text, even slightly outperforming the baseline of the simple law. As a lexical simplification model LSBert is able to reduce the mean character, letter and polysyllable count, ACCESS in combination with MUSS uses on average fewer words and therefore syllables. Both models quantitatively reduce the complexity of the normal law texts.

Just reducing the complexity of words used in the law could be misleading and change the meaning of

[9]https://github.com/qiang2100/BERT-LS accessed on 2021-11-01

[10]https://github.com/facebookresearch/muss accessed on 2021-11-01

Table 4: Comparison of count based measures for normal law, simple law and the simplifications with LSBert and ACCESS in combination with MUSS. All values depict the mean over all lines.

	normal	simple	LSBert	ACCESS + MUSS
syllable	56	53	52	**48**
words	34	31	32	**30**
sentence	1	1	1	1
char	173	161	**147**	150
letter	167	158	**142**	145
poly-syllables	6	6	**4**	5
mono-syllables	21	**19**	21	**19**

the law. Therefore, samples of the simplification results are explored and compared to the original law. The main problems of both systems can be expressed in the following two samples.

First article 5 of the civil act:

- normal:
  A minor shall obtain the consent of both parents in order to marry. If one parent is unable to exercise the right of consent, the minor shall obtain the consent of the other parent, and if neither parent is able to exercise the right of consent, the minor shall obtain the consent of his or her guardian of minor.
- simple:
  In case a minor or an adult ward marries, he/she shall go through the following procedures for consent. A minor (under the age of 19) shall obtain the consent of both parents in order to marry. If one parent is unable to exercise the right of consent, the minor shall obtain the consent of the other parent, and if neither parent is able to exercise the right of consent, the minor shall obtain the consent of his/her guardian of minor.
- LSBert:
  A child will receive the permission of both parents in order to meet. If one parent is unable to use the right of permission, the child must get the permission of the other parent, and if neither parent is able to use the right of permission, the child will get the permission of his or her father of major.
- ACCESS in combination with MUSS:
  A minor must get the consent of both parents in order to marry. If one parent is not able to give consent, the minor must get the consent of the other parent. If neither parent is able to give consent, the minor must get the consent of the minor's guardian.

The simplification shows shortcomings of the LSBert model. The word *minor* is simplified to *child* and *marry* to *meet*, changing the expressed meaning of the text significantly. ACCESS in combination with MUSS is able to simplify the text without compromising the actual content.

Second article 633 of the civil act:

- normal:
  Rent shall be paid at the end of each month in the case of a movable, a building, or a building site, and at the end of each year in the case of any other land: Provided, That in the case of a leased object which has a harvest season, rent shall be paid, without delay, upon the close of such a season.
- simple:
  If the lessee and the lessor have not determined the timing to pay rent, rent shall be paid at the end of each month.
- LSBert:
  it will be paid at the end of each month in the case of a fixed, a building, or a building site, and at the end of each year in the case of any other land: Provided, That in the case of a used subject which has a crop year, rent will be paid, without cause, upon the close of such a season.
- ACCESS in combination with MUSS:
  Rent shall be paid at the end of each month in the case of a movable object, a building or a building site, and at the end of each year in the case of any other land. The rent shall be paid, without delay, upon the close of such season in the case of a leased object.

ACCESS in combination with MUSS fails to retain the complete meaning of the original text. The model excludes the word *harvest* and, therefore, an important part of the information. LSBert exchanges the word *harvest* with the word *crop*. However, harvest season is an established concept describing a specific time of year. The crop season might refer to the season of a specific crop that does not necessarily span the same time period. Both models are not able to do the complete paraphrase the baseline, done by a human professional, suggests for this sample.

Both examples show that simple paraphrasing is insufficient for legal text, which relies on established concepts. If words are simplified with no regard of their domain the content of the text can change, rendering it unsuitable to explain the legislation to a broader audience. The results of both models can be improved by adding context-awareness.

6 CONCLUSION

We examined the English translation of South Korean legislation and its official simplification and produced a parallel corpus by aligning both sources. Subsequently, we explored the parallel corpus and investigated how the normal legalisation differs from the simple one. We concluded that simple legislation generally uses fewer and shorter sentences. Furthermore, complete sentences, fewer passive voice and modal verbs are favoured in simple law. Common Readability measures lead to insufficient results and were deemed unusable for legal texts. State of the art Text Simplification models were able to quantitatively reduce the complexity of the *normal* legal text. However, the models had problems retaining all information when used on the *normal* legal text in our parallel corpus. Awareness to the domain of the words the models paraphrase would improve the results.

ACKNOWLEDGEMENTS

This paper is based on the master thesis "Automatic English Text Simplification for Statutes" by Akshaya Muralidharan[11].

REFERENCES

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

Chandrasekar, R. and Srinivas, B. (1997). Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190.

Coster, W. and Kauchak, D. (2011). Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669.

Dras, M. (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. PhD thesis, Citeseer.

Kajiwara, T. and Komachi, M. (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.

Kauchak, D., Leroy, G., and Hogue, A. (2017). Measuring text difficulty using parse-tree frequency. *Journal of the Association for Information Science and Technology*, 68(9):2088–2100.

[11]https://wwwmatthes.in.tum.de/pages/1pwcti6a1ymz0/Master-Thesis-Akshaya-Muralidharan

Martin, L., de la Clergerie, É., Sagot, B., and Bordes, A. (2020a). Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Martin, L., Fan, A., Éric de la Clergerie, Bordes, A., and Sagot, B. (2020b). Muss: Multilingual unsupervised sentence simplification by mining paraphrases.

Qiang, J., Li, Y., Zhu, Y., Yuan, Y., Shi, Y., and Wu, X. (2021). LSBert: Lexical simplification based on BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

Rubab, I., Khan, M. Y., and Asgher, T. (2020). Transformation of legal texts into simplified accounts to make the justice accessible. *Pakistan Social Sciences Review*, 4(1):141–153.

Siddharthan, A. (2015). A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, 165(2):259–298.

Vajjala, S. and Lučić, I. (2018). Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In Li, H., Lin, C.-Y., Osborne, M., Geunbae, G., and Park, J., editors, *Proceedings of the 50th annual meeting of the Association for Computational Linguistics (ACL), Jeju, Repulic of Korea*, volume 1, pages 1015–1024. Association for Computational Linguistics.

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

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. COLING '10, page 1353–1361, USA. Association for Computational Linguistics.