Low-Resource Neural Machine Translation for Simplification of Dutch Medical Text

Marloes Evers
STUDENT NUMBER: 2051360

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

Dr. E.O.J. Vanmassenhove
C.D. Emmery MSc

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
May 2021

Preface

I would like to thank my supervisor Dr. E.O.J. Vanmassenhove for her ideas, advice, critical eye and supportive attitude during this thesis. I would also like to thank my parents and Marnix, for always supporting me and being able to deal with me and my missing talent for planning and structure.

Low-Resource Neural Machine Translation for Simplification of Dutch Medical Text

Marloes Evers

Automatic sentence simplification could be of great use for translation for lower-educated people, or second-language learners. In this work we therefore explore the possibilities of low-resource text simplification of medical Dutch text. As recent approaches have shown promising results for simplification using Neural Machine Translation (NMT), two different approaches are being investigated and evaluated, using a manually constructed test set. We propose a pivot and a zero-shot simplification model, trained with no parallel simplification corpus available. The dataset used to make the translation consists of parallel corpora from the European Medicine Agency (EMEA), the OpenSubtitles website and a Simple Wikipedia -Wikipedia alignment. The Sentence Transformers framework was used to create a medical subset out of the OpenSubtitles corpus to support domain-adaptation of the model. Results show that automatic sentence simplification is possible and that the zero-shot model outperforms the pivot model. However, as the currently presented research was based on a relatively small manually constructed test set, further research would need to be conducted in order to corroborate the current findings.

1. Introduction

The twenty-first century is the century of digitisation and visual media. In 2018 the average Dutch person watched more than three hours of visual media such as TV and streaming services per day, and this number is still growing (Ministerie van Onderwijs, Cultuur en Wetenschap 2020). Nevertheless, this does not mean that written language is no longer of importance. As one sees thousands of words per day and most information is still being transferred through written language (for instance product labels in the supermarket, or product instructions), being able to read is an important skill to get through daily life. However, even in First World countries, functional illiteracy is a remaining problem. In the Netherlands, around 2.5 million inhabitants from 16 years and older face difficulties reading and writing and can therefore face several problems in their daily life. These problems can differ from having difficulties finding jobs and filling out the applications, to comparing different products in a store, reading and understanding important information like bank statements, or reading food labels (Stichting Lezen en Schrijven 2020; Vágvölgyi et al. 2016). Illiteracy can also problematic when talking about health. In their 2013 report the World Health Organization concludes that more than half of the European population has "...inadequate or problematic health literacy." (World Health Organization 2013, p.15). For the Netherlands, this concerns 29 percent of the population that has a limited ability to follow health instructions, make convenient decisions regarding their health and to read and understand healthcare information (Van den Bercken, Sips, and Lofi 2019; World Health Organization 2013).

One way of dealing with functional illiteracy and limited health literacy for healthcare information, is simplification of medical texts. Text simplification can be described

as a way of making a text easier to read, while retaining the main meaning of the sentence and without losing important information. This is done by reducing the complexity of the text, for example by using techniques as sentence splitting, and by limiting the use of jargon and domain specific terms (Mallinson, Sennrich, and Lapata 2020). The idea of automatically simplifying texts has been around for some time. In their 1996 paper Chandrasekar, Doran, and Bangalore already state that text simplification could be of great use in several Natural Language Processing (NLP) related areas, such as Machine Translation (MT), information retrieval and summarisation. This prospect has been confirmed by the increasing amount of research into automatic text simplification and style transfer during the past years (Martin et al. 2020a). Nevertheless, research is still limited. The reason for this is that the availability of simplification corpora and parallel sentences is often scarce, especially for languages other than English (Bulté, Sevens, and Vandeghinste 2018). This is even the case when it concerns normally considered high-resource languages, such as Western European languages like French and German (Wu et al. 2020). Typically, a lot of data exists for high-resource languages, since they are widely spoken and studied The exact opposite is the case for low-resource languages. These are typically languages that are less studied, have limited written resources, since they are not spoken by a large population, are endangered, or are less commonly taught. Overall, they can be considered resource scarce (Mattoni et al. 2017; Magueresse, Carles, and Heetderks 2020). This makes simplification research outside the Anglophone world challenging, as the amount of available simplification data is currently often not sufficient to build a well-performing simplification model (Mallinson, Sennrich, and Lapata 2020). For the Dutch language this is also the case, although the language is usually considered high-resource, simplified Dutch is resource scarce and the few datasets that exist, are not publicly available (Bulté, Sevens, and Vandeghinste 2018). In this research we will therefore address simplified Dutch as low-resource. There are several ways known of dealing with translation models when it concerns a low-resource language. For example by using an Neural Machine Translation (NMT) or Statistical Machine Translation (SMT) system, one could still obtain promising results with sparse training data (Mattoni et al. 2017; Kim et al. 2019). In this research, we will provide a comparison between two low-resource NMT systems. Both techniques, pivot translation and zero-shot translation, have been shown to reduce the need for parallel data in a low-resource settings translation. We will therefore explore the possibilities of applying these techniques to a simplification task.

1.1 Motivations

As shown in the previous paragraphs, functional illiteracy is not a language specific, but a worldwide problem, even in highly developed countries. Especially in-domain texts containing domain-specific terms and jargon, such as medical or legal texts, can be hard to understand for non-native language speakers and people that are already dealing with limited literacy (Van den Bercken, Sips, and Lofi 2019; World Health Organization 2013). While previous research has mainly focused on general or domain-adapted simplification for the English language, other languages could also benefit from developments in this research area. Being able to automatically simplify medical text, this research could contribute to a better understanding of health related instructions and texts, making health information accessible to everyone. In addition, this research will contribute to the currently limited availability of simplification research in other languages than English. Although Dutch is not a low-resource language, research into its simplification is limited. By exploring the possibilities of domain-adapted automatic

simplification using a low-resource NMT system, we aim to contribute to the developments within the field of NLP, provide new insights and therefore expand the field.

1.2 Research Questions

In this research, we aim to explore the possibilities of generating simplified Dutch for Dutch medical text, using two low-resource translation techniques. This will be done using one main research question and two sub-questions, which will be explained below.

Main Research Question

As is previously mentioned, this research aims to address the limited research into automatic simplification for languages other than English. Even though a language might be considered high-resource, such as Dutch, automatic simplification is challenging as parallel simplification corpora are often not (publicly) available. Complexity of in-domain text that often contains jargon, limits the accessibility of information. With this research, we aim to explore the use of approaches to low-resource MT for text simplification. More specifically, our main research question is formulated as follows:

RQ: To what extent can low-resource MT approaches such as pivot-based and zero-shot approaches be useful for text simplification in the Dutch medical domain?

Sub-Questions

In our experiments, we will compare a pivot and a zero-shot model for Dutch medical text simplification. Both models will be trained on the same data set which will allow us to gain understanding on the differences between the two approaches. Our second research question can thus be summarized as follows: SQ1: How do pivot NMT and zero-shot NMT models compare when applied to the task of Dutch medical domain text simplification?

An important bottleneck in research on MT/NLP in general, is the application of automatic evaluation metrics to evaluate the performance of models. Many of these automatic evaluations are known to have several shortcomings (Štajner et al. 2016; Sulem, Abend, and Rappoport 2018). As such, an additional human evaluation has been conducted in order to either further corroborate the automatic evaluation scores, or reveal shortcomings of the automatic approaches. Given the importance of accurate translations and simplification (especially for a domain such as the medical domain), the comparison of the human and automatic evaluation could answer our second sub-question:

SQ2: To what extent can automatic approaches be used to assess automatic text simplification for the Dutch medical domain?

1.3 Findings

Results from this study suggest that in-domain text simplification for low-resource language is possible, however that the performance highly depends on the available data and computational resources. Nevertheless, there is definitely room for further research.

2. Related Work

In this section we will discuss related and relevant work regarding the simplification of text using NMT. We will start with an overview of text simplification research through the years. After that, the chapter will narrow down to simplification research for low-resource languages, and will be followed by a section addressing the topic of simplification of medical texts.

2.1 Text Simplification

As has been previously mentioned, automatic text simplification has been a subject of interest for some time. There exist several ways of making a text more simple. In general, two main approaches can be distinguished. One can make sentences less complex by using *syntactic* methods, such as splitting sentences into multiple ones, or by using *lexical* simplification methods, such as identification and substitution of difficult words for simpler ones (Chandrasekar, Doran, and Bangalore 1996; Siddharthan 2006; Shardlow 2014; Carroll et al. 1999). In some studies a combination of both approaches is used.

Sentence simplification is often considered a special form of monolingual MT, where the source is the complex or regular sentence and the target side is a simplified version. Zhu, Bernhard, and Gurevych (2010) were the first ones to introduce a large simplification dataset, based on sentence alignments between Wikipedia and Simple Wikipedia, for this purpose. Using this dataset they were able to successfully train a statistical simplification model, which covered "...splitting, dropping, reordering and word/phrase substitution" (Zhu, Bernhard, and Gurevych 2010, p.1360), taking both syntactic and lexical simplification into account. This is contrary to Coster and Kauchak (2011), who solely focused on lexical phrase-based simplification using specific deletion techniques. Woodsend and Lapata (2011) presented a model based on quasi-synchronous grammar and proposed an integer linear programming model for choosing the most suitable simplification. Wubben, van den Bosch, and Krahmer (2012) took another approach and added dissimilarity-based re-ranking to their phrase-based MT (PBMT) system, yielding simplification results on automatic evaluation comparable to the state-of-the-art systems as developed by Zhu, Bernhard, and Gurevych (2010) and (Coster and Kauchak 2011), while yielding better results on fluency and adequacy in human evaluation. The model also outperforms the other models on automatic evaluation, achieving a BLEU score of 43. Automatic evaluation in simplification research is often done using the metrics BLEU and SARI. BLEU (Bilingual Evaluation Understudy) compares similarity between output en reference sentences through counting overlapping word n-grams and is a widely used metric to measure the quality of a translation. SARI (System output Against References and against Input sentences) is a text simplification metric that focuses on measuring lexical simplification by looking at addition, copying and deletion of words (Xu et al. 2016).[1] Zhang and Lapata (2017) base their simplification approach on a sequence-to-sequence model (Sutskever, Vinyals, and Le 2014), using an LSTM encoder-decoder model with attention. They also train their "Deep Reinforcement Learning" simplification model on different datasets. They use their own Simple Wikipedia dataset and the Newsela corpus, consisting of manually simplified sentences, as proposed by Xu, Callison-Burch, and Napoles (2015). The final

1 A more detailed explanation of automatic evaluation metrics in MT is provided in Section 3.5.

system yields superior performance for human evaluation on simplicity, and a comparable performance to state-of-the-art models on adequacy and fluency. The automatic evaluation yields similar BLEU and SARI scores to the existing models. Zhao et al. (2018) introduce DMASS (Deep Memory Augmented Sentence Simplification), based on the Transformer architecture using multi-headed attention (Vaswani et al. 2017), achieving current state-of-the-art performance, with a SARI score of 40.45 (Alva-Manchego et al. 2019). Martin et al. (2020b) propose a multilingual unsupervised simplification approach with controllable generation mechanisms and pre-training. The approach achieves state-of-the-art results with a BLEU score of 78.17 and a SARI score of 42.53.

2.2 Text Simplification for Low-Resource Languages

As has been mentioned in the introduction, low-resource languages are languages that are overall resource scarce. Translation between two low-resource languages is therefore considered challenging. However, this limited data combined with data of a high-resource language, has shown to give promising results (Lakew et al. 2018; Mattoni et al. 2017). Currently, two main approaches exist for translation of low-resource languages. In the pivot-approach a high-resource language is used as pivoting language between two low-resource languages. In therefore consists of two language models: source-pivot and pivot-target. As long as there is enough training data available between source→pivot and target→pivot, one could still translate source→target, even though there is little to no parallel data available between source and target (Kim et al. 2019). This approach is also referred to as *explicit bridging* (Johnson et al. 2017). A second approach was introduced by Johnson et al. (2017). In their zero-shot model all components of a single multilingual model are shared. They showed that one zero-shot NMT system, trained on all data, could translate all unseen non-English language pairs. Only one model is needed in this case, in contrast to the pivot-system. For zero-shot translation, all data in both directions is concatenated and an artificial token is added to the source sentence, indicating the language of the target sentence. Johnson et al. (2017) also refer to this approach as *implicit bridging*. Nevertheless, Johnson et al. (2017) demonstrated the system using an extensive amount of data for all languages. And although many have implemented zero-shot NMT for translation of actual low-resource languages (see for example (Mattoni et al. 2017; Lakew et al. 2018; Kumar, Jha, and Sahula 2019; Korotkova et al. 2019), the extent can be broadened. As simplification data is often also considered scarce, low-resource NMT could be of great use in fields as style transfer and text simplification as well. In their paper Mallinson, Sennrich, and Lapata (2020) explore this topic of sentence simplification for a low-resource language. They propose a zero-shot modelling framework for simplification of English sentences to a low-resource language, where a parallel simplification corpus is non-existent. Simplified German is in this study considered a low-resource language. To train the model, the English Simple Wikipedia dataset is used, as well as the parallel English-German bilingual WMT19 news dataset. The last part added to the data consists of simple German non-parallel sentences, derived from a German children's magazine. After training, it is tested on two datasets: one focused on second language learners and one on children. The zero-shot transformer model is manually evaluated and automatically evaluated, using the metrics BLEU, I-BLEU, SARI and FRE-BLEU and compared to a baseline model, a pivot based model, and an Unsupervised NMT (UNMT). The zero-shot model scores significantly better than the other models on both automatic and human evaluation for the second-language learner test set (BLEU=21.11, SARI=41.12). Johnson et al. (2017) show in their paper that zero-shot NMT does not outperform pivot NMT on the translation of

high-resource languages. However, examples like Mattoni et al. (2017) and (Mallinson, Sennrich, and Lapata 2020) show that this could be different in case of working with low-resource languages.

2.3 Simplification of Medical Texts

To achieve a robust performance, an NMT system usually requires a large amount of training data. However, large datasets are often only available out-of-domain, while the better and more consistent results could be achieved while training on in-domain data.[2] In-domain simplification datasets are often small or non-existent and this is also the case for the medical domain. Shardlow and Nawaz (2019) solve this by using an already existing neural text simplification (NTS) system (Nisioi et al. 2017) and amplify it with a newly created phrase table (PT) that can link medical terminology to more simple words by mining SNOMED-CT. The final system is manually evaluated by letting annotators rank the PT baseline, NTS, original text and NTS+PT. NTS+PT is ranked first by most annotators. Automatic evaluation scores are not provided. Van den Bercken, Sips, and Lofi (2019) create their own medical simplification dataset, by lack of an existing one. The dataset that is constructed uses fully and partially aligned Wikipedia – Simple Wikipedia corpora (Hwang et al. 2015), which are filtered by medical topics and checked for medical relevance by a healthcare expert. The final dataset that is constructed consists of 3,797 aligned sentences. An LSTM model is trained on this data, combined with a dataset that is able to replace medical concepts with corresponding codes (CUI) from the Unified Medical Language System.[3] (UMLS) The model translates these to a term in the Consumer Health Vocabulary (CHV) (Zeng and Tse 2006). This replacement happens for instance when a CUI is not present in the vocabulary of the NMT model. A baseline model with pre-trained word-embedding is also trained. The models are manually evaluated and automatically evaluated using BLEU and SARI. However the advanced model scores slightly better according to the SARI metric, the baseline model is the best performing one according to the human evaluation and the BLEU score. Van den Bercken et al. (2020) state that this might be due to the fact that the information coming from the CUI/CHV might still be too complex for the purpose of simplification and assume therefore that "... training only with our extended dataset without additional replacements should yield superior performance." (Van den Bercken, Sips, and Lofi 2019, p.3291).

In this research we aim to simplify Dutch medical text using two different NMT systems: a pivot and a zero-shot NMT system. These models will be compared on performance, using Mallinson, Sennrich, and Lapata (2020) as a rough guideline. We therefore consider simplified Dutch medical text as low-resource language. Because evaluation of text simplification systems is considered challenging, many studies use a combination of several evaluation methods. We will also adopt this approach following, among others, Van den Bercken, Sips, and Lofi (2019); Mallinson, Sennrich, and Lapata (2020). They both evaluate their systems using the automated metrics BLEU and SARI as well as human evaluation on meaning preservation, grammar and simplicity. Although not mentioned in this Section, we will also adopt METEOR as evaluation method, as

2 We use the definition of (Koehn and Knowles 2017) here: "...a domain is defined by a corpus from a specific source, and may differ from other domains in topic, genre, style, level of formality, etc." [p.29]

3 https://www.nlm.nih.gov/research/umls/index.html

different studies have shown its potential in evaluating simplification (Štajner et al. 2016; Martin et al. 2019). This will be further elaborated on in Section 3.5.

3. Methods

In this section the methods that are adopted in this research will be described, providing an overview of the main concepts. The previous section has shown that there are several ways of addressing the topic of in-domain sentence simplification. There is however no clear consensus in what works best for this specific setting, as little comparable research has been done yet. Hence, we will opt for a state-of-the-art translation method and train a transformer model for both their pivot and their zero-shot models, following Mallinson, Sennrich, and Lapata (2020).

3.1 Transformer Algorithm

The transformer architecture as proposed by (Vaswani et al. 2017), consists of an encoder-decoder structure that uses a self-attention mechanism. In the original paper, both the encoder and decoder consist of a stack of six layers, all comprising two sub-layers. The first sub-layer of the encoder is the multi-head self-attention, the second a fully connected feed-forward network. The decoder has an extra sub-layer, performing multi-head attention over the output of the encoder stack, also called the masked multi-head self-attention, or encoder-decoder attention. In the original paper, 8 attention heads are used (Vaswani et al. 2017). The non-sequential character of the transformer architecture makes it possible to decode an entire input at once, by selectively looking at certain parts of a sentence. The model can learn the context of a word by looking at its surroundings. Although transformer models are currently considered state-of-the-art, they do not always outperform traditional Recurrent Neural Networks (RNN) on every task. However, when they do match or outperform the performance of the RNN model, they do so with lower computational costs. For a schematic overview and a more in-depth explanation of the model architecture we kindly refer you to the original paper "Attention Is All You Need" by Vaswani et al. (2017).

3.2 BERT and Sentence-BERT

Devlin et al. (2018) base their work on the transformer architecture with multi-headed self-attention, by introducing BERT, Bidirectional Encoding Representations from Transformers. BERT is a language representation model that is based on the transformer architecture, but does only use the encoder side to create the representations. Another difference with the original transformer architecture is that BERT uses more attention heads (Devlin et al. 2018). BERT can handle a wide variety of representation learning tasks, such as Named Entity Recognition (NER), question-answering and sentence classification (Reimers and Gurevych 2019). Many variations on BERT have been developed since the original BERT model was released. A few examples are RoBERTa, which is an optimized BERT training approach (Liu et al. 2019), DistilBERT, a distilled version of BERT with half the number of the original parameters, what makes the model faster and lighter, and language-specific pre-trained BERT models, such as BERTje (de Vries et al. 2019) and RobBERT (Delobelle, Winters, and Berendt 2020) for Dutch. The BERT modification we use for this research is Sentence-BERT. Sentence-BERT (SBERT) is developed by Reimers and Gurevych (2019) and is another modification of the original

BERT model. The model is called a twin or siamese network and is designed to be able to process two sentences at the same time. It is considered a fine-tuned version of BERT/RoBERTa, that adds a pooling operation to make is possible "... to derive a fixed size sentence embedding." (Reimers and Gurevych 2019, p.3984). It is therefore used to compute different sentence embeddings and compare them using cosine-similarity. For a schematic overview of this process, we refer you to the original paper by Reimers and Gurevych (2019). The cosine similarity between both embeddings is calculated using the following formula:

$$\text{cosine similarity: } \cos(\theta) = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|} = \frac{\sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{v}_i}{\sqrt{\sum_{i=1}^{n} \boldsymbol{u}_i^2}\sqrt{\sum_{i=1}^{n} \boldsymbol{v}_i^2}} \tag{1}$$

As the model is optimized for finding most similar pairs and for semantic search, the amount of time to perform these tasks is significantly smaller than when using the original BERT model. Within this research the SBERT SentenceTransformers framework is used to extract in-domain related sentences from the OpenSubtitles dataset.[4] This issue will be further elaborated on in section 4, Experimental Setup.

3.3 Byte Pair Encoding

BPE (Byte Pair Encoding) is a subword tokenization technique. Although it was originally introduced by Gage (1994) as a data compression technique, nowadays it's mainly popular for it's use in NMT systems to build word segments. These word segment (or subword units) are the most frequent pairs of characters or character sequences within the corpus where BPE is applied (Wu and Zhao 2018; Sennrich, Haddow, and Birch 2016; Liu et al. 2019). Low frequency words or word combinations are split up in more frequently occurring combinations. This has as result that rare subword tokens are no longer being replaced with "unknown", but that the system is able to use them for training, possibly leader to better translation results (Sennrich, Haddow, and Birch 2016). The use of BPE is therefore especially useful in corpora where the occurrence of rare terms is frequent. Hence, we apply BPE within this research, to ensure no rare, but domain-specific terms are excluded. For an example of BPE, consider the words and the possible segmentations below:

Example BPE

Word	**BPE subwords**
appendicitis	ap-pen-di-citis
cervicitis	cer-vi-citis

3.4 Translation Models

In this research, a comparison is made between two low-resource NMT models. As there is no direct parallel data available between medical Dutch and simplified Dutch,

4 https://www.sbert.net/index.html

we adopt the following two models for our experiments.

Pivot NMT

The first model is the pivot translation model. As was already briefly discussed in Section 2, the pivot model uses a *pivoting* or additional language to make the final translation between source and target. Figure 3.4 shows the set-up of our pivot model. Although we solely use English as a pivoting language, our translation requires two pivoting steps. Every arrow between the rectangles represents a one-to-one NMT model, trained on its own specific data. The first language model translates Dutch medical sentences into English medical sentences. These English medical sentences are then fed to the second model, that aims to simplify them into simple English sentences. Finally, the last model translates these simple English sentences back to simple Dutch. Hence, the three models combined should be able to eventually simplify the Dutch medical text that was inserted into the first model.

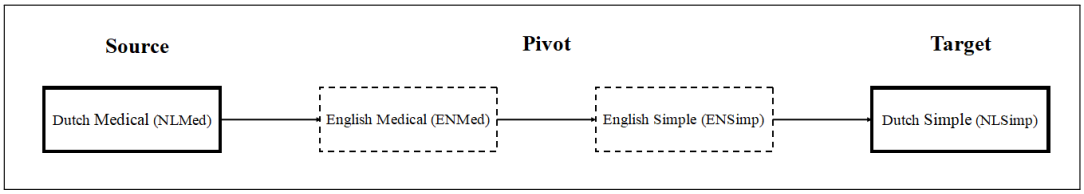

Figure 1
Schematic representation of the pivot NMT model. Every arrow represents a language model between the two connected rectangles. The dotted rectangles represent the pivoting steps within the translation.

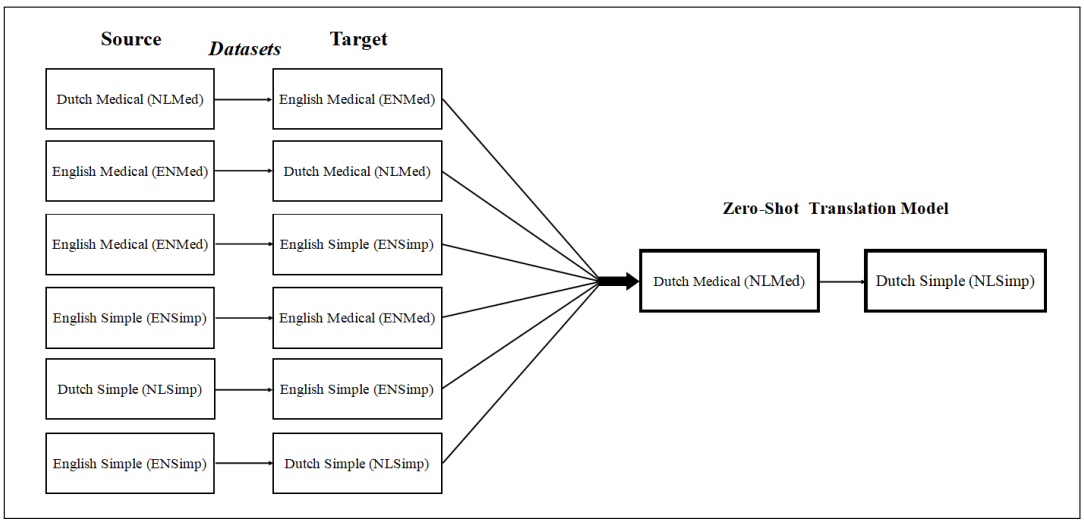

Figure 2
Schematic representation of the zero-shot NMT model. All data is concatenated and used to train the translation model.

Zero-Shot NMT

The second translation model is the zero-shot NMT model. With this zero-shot technique, originally proposed by Johnson et al. (2017), all components of a single multilingual model are shared. This enables the translation between multiple languages within one model, and also enables translation between unseen language pairs. In contrast to the pivoting method, only one model is needed to make the desired translation. All data that was used to train the pivoting models is used here as well. The datasets are used in both directions and all source sentences are modified with an artificial token, which indicates the target language. The datasets are then concatenated, as can be seen in Figure 3.4. The full dataset is then used to train the zero-shot NMT model. An example of how the artificial token is added is, provided below:

Example token zero-shot

Source sentence (NLMed)
"De verandering in de symptomen werd gemeten aan de hand van een standaardschaal voor schizofrenie."

Source sentence (NLMed) with token
"<2ENMed> De verandering in de symptomen werd gemeten aan de hand van een standaardschaal voor schizofrenie."

Target sentence (ENMed)
"The change in the symptoms was measured using a standard scale for schizophrenia."

3.5 Evaluation Methods

Evaluation in simplification research is often done using a combination of automatic evaluation metrics and human evaluation. In the latter, human annotators are asked to score the simplified text on the categories meaning preservation, grammar and simplicity. For automatic evaluation, often several metrics are used, the most commonly applied being BLEU and SARI (Janfada and Minaei-Bidgoli 2020).

Automatic Evaluation

BLEU (Bilingual Evaluation Understudy) is a metric that is used in different MT tasks to measure the quality of a translated text. It compares the similarity between the output (or candidate) translations and the reference translations. BLEU does this through counting matching word n-grams, often varying from n=1 to n=4 (Papineni et al. 2002). However, even though it's widely used, BLEU is not always considered the best evaluation metric for all MT tasks. It for instance does not take the meaning of a sentence into account, as it solely looks at exact matches of the n-grams. Synonyms and paraphrases are not allowed for and all n-grams are equally weighted. Hence, the results often correlate poorly with human judgement, especially on the category of simplicity (Sulem, Abend, and Rappoport 2018; Xu et al. 2016). A metric that is therefore often used to measure the degree of simplification in a sentence is SARI (System output Against References and against Input sentences), as proposed by Xu et al. (2016).

SARI focuses on measuring lexical simplification, on the grounds of the addition, copying and deletion of words. It for instance rewards addition operations if they relate

to a simplification and it penalizes additions that were neither in the input sentence, nor in the reference translations. Figure 3.5 shows the example adapted from Xu et al. (2016). In this example the word "now" would be rewarded, as it occurs in one of the reference sentences as well as one of the output sentences. The word "you" in Output-1 is penalized, as it did not occur in the input, nor in the reference sentences. For a more in-depth explanation regarding the internal calculations of all categories (add, keep and delete), we kindly refer you to the original paper "Optimizing Statistical Machine Translation for Text Simplification" by Xu et al. (2016).

INPUT: *About 95 species are currently accepted .*
REF-1: *About 95 species are currently known .*
REF-2: *About 95 species are now accepted .*
REF-3: *95 species are now accepted .*
OUTPUT-1: *About 95 you now get in .*
OUTPUT-2: *About 95 species are now agreed .*
OUTPUT-3: *About 95 species are currently agreed.*

Figure 3
Example of SARI, where the word "now" would be rewarded and the word "you" would be penalized. Figure adapted from Xu et al. (2016, p.404)

Two metrics that are less commonly used for evaluation text simplification are FKGL (Flesch Kincaid Grade Level) (Kincaid et al. 1975) and METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Denkowski and Lavie 2011). FKGL is originally designed to measure the readability of English text and because it mainly relies on measuring the length of the generated sentences, it has limited use. Output sentences of short length could still obtain good scores, even though the meaning and grammatical structure might be unrelated to the original sentence (Wubben, van den Bosch, and Krahmer 2012; Alva-Manchego et al. 2019). If implemented, it should be used with caution, according to Wubben, van den Bosch, and Krahmer (2012).

The METEOR metric works in a different way. It aligns the output of a system translation to one (or more) reference translations and calculates the sentence-level similarity scores. This is done based on four types of matching: exact, stem, synonym and paraphrase (Denkowski and Lavie 2014). Several studies argue that the metric is suitable for evaluating text simplification, as the metric highly correlates with human evaluation scores and often outperforms BLEU here. This is especially the case for the categories meaning preservation and grammar (Lavie and Agarwal 2007; Štajner et al. 2016; Martin et al. 2019). Nevertheless, it has not gained much attention in text simplification literature (Martin et al. 2019). An important remark for METEOR is that not all languages are fully supported by the system.[5].

As we are interested in the extent to which automatic approaches can evaluate text simplification, we will include three automatic evaluation metrics in this research. BLEU and SARI will be used, as they are considered state-of-the-art evaluation metrics for simplification research. However, as the suitability of BLEU for simplification research is debated (Sulem, Abend, and Rappoport 2018; Xu et al. 2016), we will also include METEOR as different studies have shown its potential for simplification

5 As stated in the official documentation
https://www.cs.cmu.edu/~alavie/METEOR/README.html#languages

evaluation (Štajner et al. 2016; Martin et al. 2019)

Human Evaluation
Automatic evaluation is often combined with human evaluation. It is used to fill the shortcomings of automatic metrics, which are often insufficient to fully describe a model's potential (Van den Bercken, Sips, and Lofi 2019). With human evaluation in text simplification research, annotators (often natives) are asked to score the generated sentences on three categories, meaning preservation, grammar and simplicity. Meaning preservation and grammar are usually scored on a 1-5 Likert scale (1=very bad, 5=very good), while simplicity is scored on a 5-points scale from -2 to 2, where -2 stands for "much more difficult", 0 for "equally difficult/simple" and 2 for "much simpler" (Štajner et al. 2016; Nisioi et al. 2017; Van den Bercken, Sips, and Lofi 2019). However, as human annotation is costly and time consuming, usually only a small part of the test set is evaluated this way (Van den Bercken, Sips, and Lofi 2019). This will also be the case in the human evaluation part of this study. We will ask three native speakers with no medical knowledge to score 101 sentences on the three categories meaning preservation, grammar and simplicity.

To measure *inter-annotator agreement*, quadratic Cohens kappa(Cohen 1960) is used, following (Nisioi et al. 2017). Cohen's (weighted) kappa, is a statistic for measuring the agreement between two annotators, who scored on an ordinal scale (Warrens 2012). Weighted kappa variants as linear and quadratic penalize the scores if they are extremely different. As in this study three annotators are evaluating the test set and Cohen's kappa is designed to compare two, the mean will be taken out of all annotator combinations. This means that for every category (meaning preservation, grammar and simplicity), the mean of three quadratic kappa scores will be calculated. Quadratic Cohen's kappa is calculated using the following formula:

$$\text{quadratic Cohen's kappa: } \boldsymbol{w_i} = 1 - \frac{i^2}{(k-1)^2} \tag{2}$$

4. Experimental Setup

This section will address the experimental set-up of the project. It will start of with a description of the data that is being used. The subsequent section will discuss the cleaning and pre-processing steps that are taken to prepare the data. A brief description of the final test set construction will also be included. Subsequently the design and procedure of the model construction and training will be elaborated on. The chapter will end with a brief overview addressing the packages, programs and algorithms used.

4.1 Data

Table 1 shows the contents of the datasets that are used within this research. As stated before, we are working within a low-resource domain, which means that there is no direct parallel data available for the final translation goal: Dutch medical text $\rightarrow$ simplified Dutch For the sake of testing however, a small dataset of this pair will manually be created. The three main datasets used for training simplification models, will be discussed in the following part.

Table 1
Overview of the datasets, size in number of parallel sentences.

Datasets	Size	Authors
EN-NL		
EMEA	1,090,893	Tiedemann (2012)
OpenSubtitles	37,200,621	Lison and Tiedemann (2016)
EN-ENSimple		
Wikipedia	284,738	Hwang et al. (2015)
Fully aligned	*154,805*	
Partially aligned	*129,933*	
ENMed-ENSimple		
Wikipedia Medical	3,390	Van, Kauchak, and Leroy (2020)
Wikipedia Medical*	5,415	Van den Bercken, Sips, and Lofi (2019)
Fully aligned	*2,267*	
Partially aligned	*3,148*	

**is a subset of the dataset that was created by Hwang et al. (2015).*

EMEA dataset

The EMEA dataset is a parallel corpus that is created from PDF documents provided by the European Medicines Agency (EMA). It is constructed and sentence-level aligned by Tiedemann (2012).[6] Because the EMA is an agency established by the European Union, the corpus is available in every language spoken within the European Union. The sentences are derived from medicine prescriptions, therefore including domain-specific (medical) jargon. Hence, this dataset will be used to make the translation between Dutch medical text and English medical text.[7] The dataset will therefore be referred to as NLMed→ENMed.

Wikipedia Simplification dataset

The Wikipedia dataset constructed by Hwang et al. (2015) is used to make the translation between English medical text and English simple text (ENMed→ENSimp). The original idea was to solely use the medical subset of this dataset, extracting by Van den Bercken, Sips, and Lofi (2019). However, this subset (that is discussed below) appeared to be too small to train a NMT model and led to poor preliminary results. Therefore has been decided to use the full simplification dataset provided by Hwang et al. (2015), consisting of fully and partially aligned sentences. In this way the models can be trained on as much simplification data as possible.[8] This dataset will be referred

6 https://opus.nlpl.eu/EMEA.php
7 However the most recent and correct abbreviation of the European Medicine Agency is EMA, the dataset is still called EMEA. The abbreviation EMEA will therefore be used within this research.
8 https://ssli.ee.washington.edu/tial/projects/simplification/

to as ENMed→ENSimp.

Medical Simplification subset
A freely accessible English medical simplification dataset has been issued by Van den Bercken, Sips, and Lofi (2019). Van den Bercken, Sips, and Lofi (2019) created a subset out of the Simple Wikipedia dataset (Hwang et al. 2015) by using QuickUMLS, to check whether sentences actually contained medical topics. After this, the sentences were checked by a health expert to make sure that they were actually health related. The fully and partially aligned medical datasets combined, comprises 5415 sentences.[9] This dataset is used to be able to extract the simplified medical sentences from the entire Simple Wikipedia dataset. This is done to create a small medical subset that can be manually translated to create a test set (NLMed→NLSimp). Van, Kauchak, and Leroy (2020) created their own Wikipedia - Simple Wikipedia alignments and constructed a medical subset checking whether sentences would contain four of more medical words from the UMLS database.[10] The dataset therefore partly differs from the Wikipedia medical subset created by Van den Bercken, Sips, and Lofi (2019). This dataset is used to create sentence embeddings to extract medical in-domain sentences from the OpenSubtitles dataset. This issue will be further elaborated on in Section 4.2

OpenSubtitles dataset
The OpenSubtitles dataset is one of the largest available parallel datasets. It was constructed by Lison and Tiedemann (2016) and is available in 62 languages.[11] It consists of aligned sentences from movies subtitles and therefore mainly contains conversational language. This dataset will therefore be used to make the translation between English simple (conversational) and Dutch simple (conversational) (ENSimp→NLSimp). As the dataset is quite large and everyone is allowed to upload subtitles on the website of opensubtitles.org, the dataset is diverse in grammatical constructions and word use. As the size of the dataset is disproportionate in comparison to the other datasets, and to make the dataset suitable for medical domain-adaptation, this dataset requires some extra pre-processing steps. These will be further explained in the following section.

4.2 Cleaning and Pre-processing

Before the data can be used in the models, execution of some pre-processing steps is required. First, all data is converted to lower case and all non-necessary punctuation, such as quotations, are removed. Other punctuation, such as commas and periods are kept, as context can be lost when removing them. This is also the case for stopwords, which are not removed as well. The data is furthermore checked on duplicates and empty lines, which are both removed as well. Also lines with $\leqslant$ six characters on both target as source side are removed from the datasets. Long sentences are not removed yet, as they will be excluded during the training phase.

OpenSubtitles Data Selection and Additional Cleaning
To make the OpenSubtitles corpus suitable for medical domain-adaptation, a form of data selection is executed on the dataset using the function SentenceTransformer,

9 https://github.com/myTomorrows-research/public/tree/main/WWW2019
10 https://github.com/vanh17/MedTextSimplifier/tree/master/data_processing
11 https://opus.nlpl.eu/OpenSubtitles.php

paraphrase-distilroberta base 1 from SBERT. With this framework it is possible to map all sentences to sentences that are close in vector space, using cosine-similarity. To create in-domain sentence embeddings, the simple side of the medical subset by Van, Kauchak, and Leroy (2020) is used. This results in 3,390 sentence embeddings created out of 3,390 simple medical sentences. The OpenSubtitles file however is too large to be fed to the model at once. Therefore is decided to randomly sample 20 million sentences from the corpus, and split this sample into 20 subsets of 1 million sentences each. All these subsets are encoded into sentence embeddings as well. After this step, the cosine similarity is calculated between every OpenSubtitles subset and the medical subset. K is set to 15, extracting the 15 most similar OpenSubtitles embeddings to every embedding in the medical subset. This initially leads to a new OpenSubtitles corpus consisting of 3,390*15*20=1,017,000 sentences.[12] An illustration of the process can be found in Figure 4.2. The sentence similarities of the sentences that are included in the new OpenSubtitles corpus vary roughly from 0.25 to 0.50.[13]
As the OpenSubtitles corpus is often considered quite noisy (Vinyals and Le 2015; Ghosh and Kristensson 2017) subtitle specific cleaning and extra alignment steps are executed after retrieving the new corpus. This is done using the revised version of the OpenSubtitles cleaning script by Rachel Bawden.[14] Duplicate sentences are again removed as well. The final dataset will be referred to as ENSimp→NLSimp.

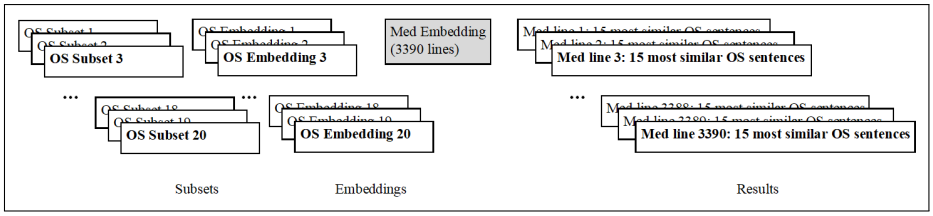

Figure 4
Illustration of the sentence similarity process

Test set

As mentioned before, is there no direct parallel data available for Dutch medical to Dutch simple text. Therefore, a small dataset is constructed to make evaluation of the models possible. For this testset, a random sample of 100 sentences is extracted from the Van den Bercken, Sips, and Lofi (2019) fully aligned medical Wikipedia subset. We choose not to include partially aligned sentences, as we want the test set to be as accurate as possible. The test set is translated using a pre-trained online translation service. After this, all sentences on both sides of the corpus are manually checked and adjusted when necessary. This is for example the case when a word is not correctly translated, or when a sentence contains severe grammatical errors. One of the sentences is split, as it actually contains of two sentences. This leads to a final Dutch test corpus of 101 sentences. After the test set is constructed, the original (randomly sampled) English sentences are

12 3,390 sentences (medical subset), 15 most similar OS sentences, 20 OS subsets
13 Although we realize there exist way more sophisticated methods for retrieving similar sentences, due to time constraints and it not being the scope of this research, other methods have not been further explored.
14 The original script can be found at `https://github.com/rbawden/PrepCorpus-OpenSubs`. Additions and revisions made by Dr. Vanmassenhove are currently not publicly available.

deleted from the main simplification corpus (ENMed→ENSimp). The final test set will be referred to as (NLMed→NLSimp).

4.3 Design & Procedure

After the execution of all steps described in Section 4.1, the final sizes of the EMEA, OpenSubtitles and Simple Wikipedia corpora are respectively 279,097, 539,235 and 283,117, as can be seen in Table 2. Further processing and analysis is executed using Google Colab Pro. The training and translation pipelines are based on the open-source NMT framework OpenNMT (Klein et al. 2017), version 2.1.2., PyTorch implementation.[15]

Table 2
Final datasets and sizes after cleaning and pre-processing. Sizes in parallel sentences

Dataset		Origin	Size
NLMed	-> ENMed	EMEA	279,097
ENMed	-> ENSimp	Simple Wikipedia	283,117
ENSimp	-> NLSimp	OpenSubtitles (med subset)	539,235
NLMed	-> NLSimp	Simple Wikipedia (med subset)	101

Pivot Translation

Three models are build for the pivot translation system, NL-EN Medical, EN-EN Simplification and EN-NL Simplified. In every model, the pre-processed data is treated the same way and trained on same parameters, to keep consistency within the system. A full overview of all settings is provided in Appendix A.

First, the lines in the datasets are tokenized using the tokenizer provided by the ONMT framework. Multithreading (16) is using to boost performance. The datasets are then split into train test en development sets, using size 1000 for both development and testing.[16] As the final testing of the models will happen using a different dataset than the one split here, the size is kept small, to train the models on as much data as possible. After splitting, a cut-off value of 160 is used to exclude sentences with more than 160 tokens. Subsequently BPE (Sennrich, Haddow, and Birch 2016) is used to build the word segments, using an merge operation size of 50,000.

Training is initially done using the freely available version of Google Colab, enabling the possibility to train the models on GPU for free. However, as the GPU time in the free version of Colab is limited and use is not guaranteed, training is during the project switched to Colab Pro, providing a faster, more stable and therefore more reliable GPU connection. For training, the maximum input length of a sentence is 150, following the approach by Mattoni et al. (2017). The models are trained using 30,000 train steps, and training is stopped after these steps have passed or when validation accuracy goes down for more than three times in a row. Validation is done every

15 The configuration files that were used and adapted for implementation in the OpenNMT framework were originally created and provided by Dr. Vanmassenhove and Dr. Shterionov

16 For training and testing, the following scripts were adopted:
https://ilk.uvt.nl/~shterion/scripts/train_test_dev.py
https://ilk.uvt.nl/~shterion/scripts/bpe.sh

500 steps. The optimizer used is Adam (Kingma and Ba 2014), adopting the settings from Vaswani et al. (2017), β_1=0.9, β_2=0.998. The learning rate is varied over the training, using noam as decay method and warm_up steps=4000 (Vaswani et al. 2017). Word_vec_size and RNN_size are both set to 256, to avoid overloading the system. Drop-out(Srivastava et al. 2014) is set to 0.1, even as label smoothing. Every 500 training steps, the model is saved, making it possible to evaluate different models afterwards. The models with the highest automatic evaluation scores are used for the final pivot translation.

Zero-Shot Translation

For the zero-shot model, all data is tokenized as well using the same settings on the same datasets as for the pivot translation models. After tokenization, all datasets in both directions (for example NL-EN Medical and EN-NL Medical), are concatenated and artificial tokens are added to the source sentences, indicating the language of the corresponding target sentence. After adding the tokens, the concatenated dataset is split in a train, development and test set, again using size 1000 for both and a cutoff value of 160 to exclude long sentences. BPE is executed as well, using a merge operations size of 50,000. The training process of the zero-shot model uses the same settings as the pivot models and models are evaluation and saved every 500 steps. Finally, the model that is performing best on the automatic metrics BLEU and SARI, is selected.[17]

Translation and Evaluation

For the pivot translation of the final test set (NLMed→NLSimp), the best performing models are combined to create the pivot translation pipeline. The best performing zero-shot model is chosen to represent the zero-shot translation pipeline. For translation of both models beam size=6 is used. After translation, all data is detokenized, as the used evaluation metric require detokenized data as input (Alva-Manchego et al. 2019). Evaluation is done using corpus-level BLEU (sacreBLEU Post (2018)) and corpus-level SARI using the EASSE framework (Alva-Manchego et al. 2019). Corpus-level METEOR is implemented using the NLGEval framework (Sharma et al. 2017). In addition to the automatic evaluation, human evaluation is executed, as automated metrics have been considered insufficient for fully describing the potential of simplification using (Neural) MT (Van den Bercken, Sips, and Lofi 2019; Mallinson, Sennrich, and Lapata 2020). For the human evaluation part, ratings are acquired using a Likert scale to measure the performance of the model. Three native Dutch speakers score the generated translations on the categories meaning preservation (1-5), grammar (1-5) and simplicity (-2-+2).[18]

4.4 Algorithms and Packages

For this research, several algorithms, programs and packages will be used. This subsection provides a short overview of these, including the versions:

- EASSE (Easier Automatic Sentence Simplification Evaluation) 0.2.4
- NLGEval (Natural Language Generation Evaluation) 2.3.0

17 As METEOR was not yet included during the first phase of the research, this metric was not used to select the best performing translation models.

18 The (Dutch) instructions that were given to the annotators can be found in Appendix A

- OpenNMT-py 2.1.2
- Python 3.7.10
- Python Numpy 1.20.0
- Python Pandas 1.1.5
- Python ScikitLearn 0.24.2
- Python SciPy 1.6.2
- Sacrebleu 1.5.1
- Sentence Transformers 1.1.1

5. Results

This Section will address the results of the described models. First, a brief overview of the performances of the individual models will be given. After this, the final models will be evaluated. This will be done using the automatic evaluation metrics BLEU, METEOR and SARI and additionally human evaluation for scores on meaning, grammar and simplification of the sentences in the test set (NLMed→NLSimp). This Section will also provided some example translations that are generated by the pivot and zero-shot models.

5.1 Results of Individual Models

Table 3 gives an overview of the models that were included in the final translation pipelines of the pivot and zero-shot model. The first model for the pivot approach was trained on a cleaned EMEA dataset, translating NLMed to ENMed. The second model was trained on a Wikipedia - Simple Wikipedia alignment, ENMed-ENSimp. For this model, use was made of the entire simplification dataset and not only the medical subset, to be able to train the model with as much data and examples as possible. The third model, ENSimp-NLSimp, used a subset of the OpenSubtitles dataset. This subset was created using the SBERT framework to measure sentence similarity between sentences in the OpenSubtitles dataset and those in an existing simplified English medical corpus. The three models combined represent the pivot translation pipeline, or pivot model. The performance of the models in the pivot pipeline range in BLEU score from 57.74 (NLMed-ENMed) to 24.28 (ENSimp-NLSimp) and in METEOR score from 42.28 to 23.14. SARI is only calculated for ENMed-ENSimp, as this is the only model that performs a simplification.

The zero-shot model is trained on the exact same data as the pivot model. However, all data in both directions is concatenated and an artificial token indicating the target language has been added to the source sentences. The performances of the zero-shot model on its own test set vary from a BLEU score of 30.90 to a METEOR score of 26.92. The SARI score of 57.51 is relatively high, probably because it is not only being calculated on the simplification corpus (ENMed-ENSimp), but on the entire concatenated dataset that was used to train the zero-shot model.

Overall can be seen that the individual models perform well on their own test set. Some examples translations of the individual models can be found in Appendix C.

Table 3
Train steps, perplexity, BLEU, METEOR and SARI scores of the models on their own test sets

		Perplexity scores				
Model	**Train steps**	**Train**	**Translation**	**BLEU**	**METEOR**	**SARI**
Pivot model						
NLMed - ENMed	15,000	1.36	1.30	57.74	42.28	
ENMed - ENSimp	8,500	3.70	1.55	28.34	23.14	39.89
ENSimp - NLSimp	11,000	2.53	2.53	24.28	25.21	
Zero-shot model						
Zero-shot	15,000	6.36	1.72	30.90	26.92	57.51

Table 4
Automatic evaluation using BLEU and SARI for the pivot and the zero-shot translation models. Best scores are boldfaced

Model	**BLEU**	**METEOR**	**SARI**
Pivot	5.28	10.63	33.32
Zero-Shot	**27.19**	**31.19**	**40.04**

5.2 Automatic Evaluation

As stated before, the main goal of this research is to compare the performances of two low-resource NMT models, pivot and zero-shot, on the task of simplification of Dutch medical text. Table 4 summarizes the performances of both models on automatic evaluation metrics when presenting them the unseen test set (NLMed-NLSimp). As the combination of datasets used in this study has not been used before in similar research, it is difficult to compare the scores in Table 4 to the state-of-the-art BLEU and SARI scores for simplification research. As METEOR has limited use as metric for simplification, it is hard to compare the scores for this metric as well. We can however compare the BLEU and SARI scores to a research with similar set-up by Mallinson, Sennrich, and Lapata (2020), who investigated simplification models for low-resource languages. Important to note is that they did not adapt a specific domain. The SARI scores Mallinson, Sennrich, and Lapata (2020) report for their pivot and their zero-shot model are respectively 38.64 and 41.12, on a second-language learners dataset. The pivot and zero-shot model in this research obtain similar automatic evaluation scores, as our pivot model achieves a SARI of 33.32, and our zero-shot model outperforms this score with SARI=40.04. Although was stated before that zero-shot NMT usually does not outperform pivot NMT (Johnson et al. 2017), Mattoni et al. (2017) and Mallinson, Sennrich, and Lapata (2020) already showed that this could be different in case of dealing with low-resource languages. This is also apparent from our research, where the zero-shot NMT scores higher on all three automatic evaluation metrics.

The results of our research implicate that simplification using a pivot model is challenging. Despite the shortcomings of the BLEU metric, for instance that is does not accept synonyms or paraphrases and does not take the meaning of a sentence into account, one could conclude that a BLEU score of 5.28 is extremely low. This indicates

the model can be considered useless, as the meaning of the sentences does not relate at all to the references sentences anymore. We can also conclude this when looking at the METEOR score of 10.63. When comparing these scores to BLEU and METEOR scores of the zero-shot model, one might conclude that this model produces more accurate translations and simplifications. The BLEU score of 27.19 and the METEOR score of 31.19 of the zero-shot model indicate that these outputs are more related to the original reference sentences. Table 5 shows three examples of source sentences, reference sentences and the translation that is made by both models. The results indeed differ per model, but overall both outputs still seem to address the same topics. Human evaluation is therefore applied to further analyze the performances of both models.

Table 5
Example sentences from the models outputs

Model	Sentence
Source	Wanneer muizen blootgesteld worden aan onvoorspelbare chronische milde stress, beginnen zij symptomen te vertonen die doen denken aan een zware depressieve stoornis bij de mens.
Reference	Muizen die lijden aan onvoorspelbare chronische lichte stress vertonen symptomen die lijken op depressie bij mensen .
Pivot	Als muizen zijn blootgesteld aan chronische stress ... hebben ze symptomen die in de mens kunnen zijn .
Zero-Shot	Als muizen worden blootgesteld aan onvoorspelbare chronische lichte stress , beginnen ze symptomen te vertonen die denken aan een zware depressie bij de mens .
Source	De meeste kleurenblindheid is permanent, maar sommige aandoeningen kunnen ook tot tijdelijke kleurenblindheid leiden. Kleurenblindheid kan daarnaast ook erfelijk zijn.
Reference	Kleurenblindheid is vaak blijvend, maar sommige aandoeningen kunnen tot tijdelijke kleurenblindheid leiden. Kleurenblindheid kan ook erfelijk zijn.
Pivot	Blindheid is permanent... Maar sommige aandoeningen kunnen leiden tot tijdelijke blindheid.
Zero-Shot	De meeste kleurenblindheid is permanent, maar sommige aandoeningen kunnen ook tot tijdelijke kleurenblindheid leiden.
Source	De spieren verzwakken en vertonen atrofie.
Reference	De spieren verzwakken.
Pivot	Spierskelet en verband.
Zero-Shot	De spieren verzwakt en vertonen atrofie.

5.3 Manual Evaluation

For the human evaluation part, all 101 sentences in the test set are evaluated by three annotators on a 5-point Likert scale, for the categories meaning preservation, grammar and simplicity. They did this using two surveys, one showing the simplifications of the

Table 6
Human evaluation mean scores. Meaning and grammar on scale 1-5, simplification scale -2-+2. Best scores are bold-faced.

Model	Meaning	Grammar	Simplification
Pivot	2.017	2.825	-0.264
Zero-Shot	**3.686**	**3.525**	**0.010**

** Models score significantly different on all categories ($p<0.001$). Significance tests performed using student's t-test.*

Table 7
Inter-annotator agreement scores per category and model. IAA is calculated using quadratic Cohen's kappa.

Model	Meaning	Grammar	Simplification
Pivot	0.693	0.423	0.293
Zero-Shot	0.250	0.222	0.209
Mean	*0.471*	*0.322*	*0.251*

pivot model, compared to the source sentences, the other one showing the simplifications made by the zero-shot model, compared to its source sentences. The mean results of these evaluations can be found in 6. As the results show, the zero-shot NMT model again outperforms the pivot model NMT, on all categories. In the pivot model, both meaning and grammar are not well captured, as they score below 3 ("not good/not bad"). The score 0.26 for simplification implies that the model made the sentences harder to understand, not simpler. The zero-shot model performs slightly better, with scores of respectively 3.69 and 3.52 for meaning preservation and grammar. Simplification is however close no non-existent, with a score of 0.01, where 0 indicates "not more difficult, not more simple". To further examine the scores given by the annotators, the inter-annotator agreement (IAA) is calculated using quadratic Cohen's kappa. As Cohen's kappa is originally designed for IAA between two annotators, the score is calculated for every pair of annotators and of this, the mean is calculated. The scores for Cohen's kappa vary from 0 to 1, respectively indicating no agreement to full agreement. The results in table 7 show the IAA scores per model and per category. Surprisingly, the annotators did not agree as much on meaning preservation and grammar for the zero-shot model (0.250, 0.222), as they did for the pivot model (0.693, 0.423).

As the automatic evaluation metrics suggest that the zero-shot system is best is capturing meaning preservation, grammar and generating a simplification, the output of the zero-shot model is manually analyzed. It appears that a large amount of the simplified sentences are left unchanged by the zero-shot model and don't differ from the source sentence. It is therefore possible that these negatively affect the mean simplification score. For further analysis of the zero-shot NMT system, the duplicate sentences are therefore removed. This results in a corpus of 88 sentences. Table 8 shows the mean human evaluation scores after removal of the duplicate sentences. The scores for meaning preservation and grammar are slightly lower than before removing the duplicates. The improvement in simplification in nil. The IAA scores for the zero-shot model, after removal of the duplicates are respectively

Table 8
Further analysis of the performance of the zero-shot model, after removing duplicates. Meaning and grammar on scale 1-5, simplification scale -2-+2, IAA in quadratic Cohen's kappa.

	Meaning	**Grammar**	**Simplification**
Zero-Shot	3.598	3.413	0.011
IAA	0.239	0.190	0.211

6. Discussion

In this Section, we will discuss the results as presented in the previous chapter, while we will attempt to answer the research questions that were introduced in Section 1.2. The limitations of the study will also be discussed.
The goal of this study was to explore the possibilities of in-domain text simplification for a low-resource language using an NMT system, focused on simplifying Dutch medical text. Although NMT of low-resource languages and automatic text simplification have been relevant research topics for years, the potential of using both techniques in the same system has not been fully explored yet, in our opinion. Simplification can be of great use in every language, to simplify text for lower educated people, or second-language learners. This is especially the case for health care related information, as this information is important for the population's well-being and the ability to make healthy decisions, as several studies have pointed out (Van den Bercken, Sips, and Lofi (2019); Van, Kauchak, and Leroy (2020); Aluísio et al. (2008) and many others).
For the purpose of this goal, two low-resource NMT models were trained, yielding different performances. Table 4 shows that both models score well on the automatic simplification metric SARI, yielding scores of respectively 33.32 and 40.04 for the pivot and the zero-shot model. The pivot model however, scored extremely low on both automatic evaluation metrics BLEU (BLEU=5.28) and METEOR (METEOR=10.63), what might indicate that the simplifications that were generated are no longer related to the original sentences. The zero-shot model scored better on all automatic evaluation metrics. Yielding a BLEU score of 27.19 and a METEOR score 31.19, we can say that the overall meaning of the sentence has stayed the same in the process of simplification. However, human evaluation showed that both models barely succeed in simplifying the source sentences. In addition, the scores that were given for the categories meaning preservation and grammar, showed that there is enough room for improvement. The average score given by three annotators for meaning preservation was 2.02 for the pivot model and 3.69 for the zero-shot model. The average scores for grammar were respectively 2.83 and 3.53.
As the performance of a model can highly depend on the data on which it is trained, one would prefer to train NMT models on as much and as high qualitative data as possible. This is however often not possible. Especially when training a low-resource translation model, data is often scarce and the quality often highly depends on the source. In this study, direct parallel data was not available at all. Therefore has been decided to manually construct a test set, to be able to compare both models. As a small test set is better than no test set, we based our evaluation on this small dataset. However, one could argue that for proper evaluation a larger and higher qualitative dataset would be needed. Creating a more extensive test set was unfortunately not possible within the time and resource constraints of this research. The data that was

used to train all models also highly differed in quality. This could somewhat be inferred from Table 3, where the EMEA dataset NLMed-ENMed achieves much higher BLEU and METEOR scores on their own test set than for example the OpenSubtitles based dataset ENSimp-NLSimp. As the latter is consisting of sentences that are randomly taken from movie subtitles, context is often lacking. As the subtitles can be uploaded by anyone, grammatical structure and word use can differ a lot as well. When we step-by-step examine the translations of the models in the pivot NMT pipeline, we see that even though the EMEA dataset might perform well on its own test set, a significant amount of context is lost, especially during the first step (NLMed-ENMed). Two examples are provided below:

Example 1:
Original: De spieren verzwakken en vertonen atrofie.

NLMed→ENMed: Musculoskeletal and connective tissue disorders.

ENMed→ENSimp: Musculoskeletal and connective tissue disorders.

ENSimp→NLSimp: Spierskelet en verband.

Example 2:
Original: Sikkelcelziekte komt vaker voor bij mensen van wie de voorouders in tropische en sub-tropische sub-sahara regio's leefden, waar malaria veel voorkomt of voorkwam.

NLMed→ENMed: Sickle cell syndrome is more common in people who have experienced the onset of understanding and papulopustular lesions in opposite part of a higher survival time or prevented .

ENMed→ENSimp: Sickle cell syndrome is more common in people .

ENSimp→NLSimp: Sikkelcelanemie .

As can be seen from the examples, a lot of context is indeed lost already in the first translation step, as the translation of the original sentence in example 1: *The muscles weaken and show atrophy."*, is entirely different from the translation made by NLMed→ENMed: *" Muscoloskeletal and connective tissue disorders."*. This is also the case in example 2, where the NLMed→ENMed translation system does not translate well and even adds part to the sentence that are not related to the original. The translation of the original sentence in example 2: *" Sickle cell disease occurs more commonly among people whose ancestors lived in tropical and subtropical sub-Saharan regions, where malaria is or was common."*. As the EMEA dataset only contains highly in-domain specific data, while the test set also contains more general sentences, this could be a reason that the pivot translation model performs a lot worse than the zero-shot model, as context is lost already in the first translation step. The zero-shot model, that is trained on all data at the same time, sees more examples. This probably makes it possible to translate the sentences while staying closer to the original meaning and grammatical structure. This however also led to the zero-shot exactly copying the source sentence and presenting it

as the simplified version. An example is provided below:

Example 3:
Original: Vrouwen blijken een verhoogd risico te lopen op een stille bereorte, waarbij hoge bloeddruk en roken tot de predisponerende factoren horen.

Zero-shot simplification: Vrouwen blijken een verhoogd risico te lopen op een stille beroerte, waarbij hoge bloeddruk en roken tot de predisponerende factoren behoren.

As can be seen from the example, the zero-shot model duplicated the original sentence. This happened in 13 sentences, what is a quite extensive amount in a corpus with a total size of 101 sentences. As the automatic evaluation metric BLEU looks at the extent to which the exact n-grams of a sentence are similar to the original, this could explain the high score for this metric. The zero-shot model also achieved a high METEOR score. METEOR calculates sentence similarity scores based on four matching types exact, stem, synonym and paraphrase. However, as the metric is for Dutch only optimized for the first two, exact and stem, this could also explain the relatively high score here. [19].

6.1 Limitations

Several limitations were encountered during this research. First, the data processing and cleaning took much longer than expected, leaving less of the (already limited) time to train extra models and explore different settings of the parameters. For simplification, one extra model was trained, mimicking the exact baseline set-up for medical text simplification as proposed by Van den Bercken, Sips, and Lofi (2019). They used a two-layered LSTM model with pre-trained Word2Vec embeddings. The results achieved by us gave a completely different BLEU score (22.92) than the one reported in the original paper (53.07). As transformer models often outperform LSTMs and this was also the case for us, this experiment was not continued. Another limitation that was encountered was the availability of simplification data. Two main simplification datasets are know, the Simple Wikipedia dataset and the Newsela corpus. Xu, Callison-Burch, and Napoles (2015) among others however, report that the Wikipedia simplification dataset is suboptimal, because "...It is prone to automatic sentence alignment errors; 2) It contains a large proportion of inadequate simplifications; 3) It generalizes poorly to other text genres." (Xu, Callison-Burch, and Napoles 2015, p.283). Access to the higher qualitative Newsela corpus was unfortunately not granted.[20] The only simplification dataset available was therefore the Simple Wikipedia alignment. Due to time constraints, further exploration of different data selection techniques was not possible. The OpenSubtitles medical subs-selection made using sentence alignments might have not been the most sophisticated approach, as every sentence in the medical subset by Van, Kauchak, and Leroy (2020) had an equal contribution to the final OS medical subset. A better option would also have been to look at the similarity

19 As stated in the official documentation
https://www.cs.cmu.edu/~alavie/METEOR/README.html#languages
20 https://newsela.com/data/

scores themselves instead of taking the top 15 similarities per sentence. However, within this research, this approach was too computationally expensive for the resources available. As stated before, was the size of the test set also a limitation. As a random sample of 100 taken from the English medical subset, the final performances of both models can heavily fluctuate, as not every sentence from the automatic alignment is from the same quality. Although we only included so-called fully aligned sentences, the simplifications did not always fully contain the same information as the original sentences. An example of this can be found below:

Example 4
Original sentence: Een plotselinge afsluiting van een kransslagader resulteert in een myocardiaal infarct of hartaanval.

Reference simplification: Een hartaanval wordt ook wel een myocardinfarct genoemd.

which translate to:

Original: A sudden occlusion of a coronary artery results in a myocardial infarction or heart attack .

Reference simplification: A heart attack is also called a myocardial infarction .

6.2 Contribution

Overall, this research has shown the possibilities and pitfalls of automatic simplification of domain-specific text, in a low-resource setting. Although we discussed that Dutch is not an actual low-resource language, simplification data to train a model is scarce. In Section 1, we introduced the research question: "To what extent can low-resource MT approaches such as pivot-based and zero-shot approaches be useful for text simplification in the Dutch medical domain?" and the sub-questions: "How do pivot NMT and zero-shot NMT models compare when applied to the task of Dutch medical domain text simplification?" and "To what extent can automatic approaches be used to assess the automatic simplification of Dutch medical text?". In this study we showed that there are possibilities to make an automatic simplification of Dutch medical text. We showed that a zero-shot model performs better than a pivot-shot model in a low-resource setting, in the footsteps of Mattoni et al. (2017); Mallinson, Sennrich, and Lapata (2020), but that performance is still highly dependent on the availability and quality of the data, of source, target, but especially pivot-language. When working with in-domain data, one should realize that translation can get lost, especially in a pivot model setting. Compared to a pivot model, the zero-shot model therefore automatically performs better, as it sees all data during the same training and therefore has a higher chance of correctly translation and simplifying the source sentence. To evaluate our models, we made use of three automatic evaluation metrics. The zero-shot model outperformed the pivot model on all metrics, BLEU, METEOR and SARI, as well on the human evaluation. However, although one might say that the metrics therefore correctly correlate with the human evaluation, we cannot say this with certainty. We argue that a test set of size 101 cannot fully indicate whether this is actually the case.

7. Conclusion

This research tried to answer the question to what extent Neural Machine Translation can be used to simplify Dutch medical information. Automatic text simplification can especially be useful for people dealing with (health) illiteracy and second-language learners, as information regarding healthcare should be accessible to everyone.

This research tried to answer the question *"To what extent can Neural Machine Translation (NMT) be used to simplify Dutch medical information?"*. Results of this study suggest that it is possible to do so. However, with limited in-domain simplification data available, there is room for improvement. The transformer-based zero-shot model that was trained in this research, showed promising results and therefore could be the starting point of further exploring the possibilities of text simplification for low-resource languages. The construction of a more extensive and higher qualitative dataset could contribute to a stronger foundation of the evaluation of the models.

For further research, we suggest to explore the possibilities of a low-resource simplification system for another domain.

References

Aluísio, Sandra M, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 240–248.

Alva-Manchego, Fernando, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. Easse: Easier automatic sentence simplification evaluation. *arXiv preprint arXiv:1908.04567*.

Van den Bercken, Laurens, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.

Bulté, Bram, Leen Sevens, and Vincent Vandeghinste. 2018. Automating lexical simplification in dutch. *Computational Linguistics in the Netherlands Journal*, 8:24–48.

Carroll, John A, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270.

Chandrasekar, Raman, Christine Doran, and Srinivas Bangalore. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Delobelle, Pieter, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Denkowski, Michael and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Association for Computational Linguistics, Edinburgh, Scotland.

Denkowski, Michael and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gage, Philip. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2).

Ghosh, Shaona and Per Ola Kristensson. 2017. Neural networks for text correction and completion in keyboard decoding. *arXiv preprint arXiv:1709.06429*.

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

Janfada, Behrooz and Behrouz Minaei-Bidgoli. 2020. A review of the most important studies on automated text simplification evaluation metrics. In *2020 6th International Conference on Web Research (ICWR)*, pages 271–278, IEEE.

Johnson, Melvin, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kim, Yunsu, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. 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)*, pages 865–875.

Kincaid, J Peter, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Kingma, Diederik P and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017,*

System Demonstrations, pages 67–72, Association for Computational Linguistics, Vancouver, Canada.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Korotkova, Elizaveta, Agnes Luhtaru, Maksym Del, Krista Liin, Daiga Deksne, and Mark Fishel. 2019. Grammatical error correction and style transfer via zero-shot monolingual translation. *arXiv preprint arXiv:1903.11283*.

Kumar, Rashi, Piyush Jha, and Vineet Sahula. 2019. An augmented translation technique for low resource language pair: Sanskrit to hindi translation. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 377–383.

Lakew, Surafel M, Marcello Federico, Matteo Negri, and Marco Turchi. 2018. Multilingual neural machine translation for low-resource languages. *IJCoL. Italian Journal of Computational Linguistics*, 4(4-1):11–25.

Lavie, Alon and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

Lison, Pierre and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. *International Conference on Language Resources and Evaluation*.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Magueresse, Alexandre, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Mallinson, Jonathan, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126.

Martin, Louis, Éric Villemonte de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020a. Controllable sentence simplification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4689–4698.

Martin, Louis, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020b. Multilingual unsupervised sentence simplification. *arXiv preprint arXiv:2005.00352*.

Martin, Louis, Samuel Humeau, Pierre-Emmanuel Mazaré, Antoine Bordes, Éric Villemonte de La Clergerie, and Benoît Sagot. 2019. Reference-less quality estimation of text simplification systems. *arXiv preprint arXiv:1901.10746*.

Mattoni, Giulia, Pat Nagle, Carlos Collantes, and Dimitar Shterionov. 2017. Zero-shot translation for indian languages with sparse data. *Proceedings of the 16th machine translation summit (MTSummit 2017)*, 2:1–10.

Ministerie van Onderwijs, Cultuur en Wetenschap. 2020. Tijdsbesteding aan media.

Nisioi, Sergiu, 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.

Papineni, Kishore, 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.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Association for Computational Linguistics, Brussels, Belgium.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Association for Computational Linguistics, Hong Kong, China.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Association for Computational Linguistics, Berlin, Germany.

Shardlow, Matthew. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Shardlow, Matthew and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389.

Sharma, Shikhar, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

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

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Štajner, Sanja, Maja Popovic, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016. Shared task on quality assessment for text simplification. *Training*, 218(95):192.

Stichting Lezen en Schrijven. 2020. Factsheet laaggeletterdheid in Nederland.

Sulem, Elior, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Vágvölgyi, Réka, Andra Coldea, Thomas Dresler, Josef Schrader, and Hans-Christoph Nuerk. 2016. A review about functional illiteracy: Definition, cognitive, linguistic, and numerical aspects. *Frontiers in psychology*, 7:1617.

Van, Hoang, David Kauchak, and Gondy Leroy. 2020. Automets: The autocomplete for medical text simplification. *arXiv preprint arXiv:2010.10573*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vinyals, Oriol and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Warrens, Matthijs J. 2012. Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables. *Statistical Methodology*, 9(3):440–444.

Woodsend, Kristian and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420.

World Health Organization. 2013. Health literacy: The solid facts.

Wu, Anne, Changhan Wang, Juan Pino, and Jiatao Gu. 2020. Self-supervised representations improve end-to-end speech translation. *arXiv preprint arXiv:2006.12124*.

Wu, Yingting and Hai Zhao. 2018. Finding better subword segmentation for neural machine translation. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, pages 53–64.

Wubben, Sander, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024.

Xu, Wei, 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.

Xu, Wei, 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.

Zeng, Qing T and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29.

Zhang, Xingxing 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, Association for Computational Linguistics, Copenhagen, Denmark.

Zhao, Sanqiang, Rui Meng, Daqing He, Saptono Andi, and Parmanto Bambang. 2018. Integrating transformer and paraphrase rules for sentence simplification. *arXiv preprint*

arXiv:1810.11193.
Zhu, Zhemin, 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.

Appendices

A. Parameter settings of the Transformer models

train steps: 30000
valid steps: 500
warmup steps: 4000
report every: 100
decoder type: transformer
encoder type: transformer
word vector size: 256
rnn size: 256
layers: 6
transformer ff: 2048
heads: 8
accum count: 8
optim: adam
adam beta1: 0.9
adam beta2: 0.998
decay method: noam
learning rate: 2.0
max grad norm: 0.0
batch size: 4096
batch type: tokens
normalization: tokens
dropout: 0.1
label smoothing: 0.1
max generator batches: 2
param init: 0.0
param init glorot: 'true'
position encoding: 'true'

Appendix B: Instructions human evaluation

Bedankt dat u wil deelnemen aan dit onderzoek naar de automatische versimpeling van tekst ten behoeve van mijn masterscriptie onderzoek. U gaat straks twee zinnen vergelijken. De eerste zin is het origineel, de tweede zin is een versimpelde versie van het origineel. U zal voor drie categorieën gaan aangeven in hoeverre zin 2 een versimpeling is van de zin 1. Dit wordt gedaan door middel van het geven van een score op een 5-puntsschaal. Probeert u de categorieën alstublieft los van elkaar te beantwoorden. De drie categorieën en hun bijbehorende puntenschalen zijn als volgt:

1. **Betekenis** – In hoeverre is de betekenis van beide zinnen hetzelfde?

1 = zeer slecht
2 = slecht
3 = niet slecht/niet goed
4 = goed
5 = zeer goed

2. **Grammatica** – In hoeverre is de grammatica van zin 2 correct?
Met grammatica wordt met name de zinsbouw bedoeld, niet zozeer spelling.

1 = zeer slecht
2 = slecht
3 = niet slecht/niet goed
4 = goed
5 = zeer goed

3. **Versimpeling** – In hoeverre is zin 2 een versimpelde versie van zin 1?
U kunt hierbij denken aan de lengte van de zin, het gebruik van makkelijkere woorden, het weglaten van overbodige woorden, etc.

-2 = veel moeilijker
-1 = ietwat moeilijker
0 = even moeilijk
+1 = ietwat simpeler
+2 = veel simpeler

Het invullen van de gehele enquête zal ongeveer 20 minuten in beslag nemen. U kunt op elk moment stoppen, als u niet langer wenst deel te nemen. Dit kan zonder hiervoor een reden op te geven. Uw antwoorden worden opgeslagen ten behoeve van dit onderzoek, en zullen bewaard worden voor eventueel vervolgonderzoek. Bij het inleveren van de enquête geeft u toestemming voor het gebruik van uw antwoorden.Uw antwoorden kunnen op verzoek altijd verwijderd worden door een e-mail te sturen naar het volgende e-mailadres: m.e.j.evers@tilburguniversity.edu.

Zie voor verdere informatie met betrekking tot uw privacy ook https://www.tilburguniversity.edu/nl/over/gedrag-integriteit/privacy-en-security.

Alvast bedankt voor uw deelname. Met vriendelijke groet, Marloes Evers

C. Outputs of the individual models

Model	Sentence
NLMed-ENMed	
Reference	Controle van hun nierfunctie en plasma kalium spiegels voor aanvang en tijdens de behandeling met deze gecombineerde therapie wordt aanbevolen .
Target	Monitoring of their renal function and plasma potassium levels is recommended before initiation and during treatment with combined therapy .
Prediction	Monitoring of renal function and plasma potassium levels prior to initiating and during treatment with this combination therapy is recommended.
ENMed-ENSimp	
Reference	The primary symptoms of cholera are profuse diarrhea and vomiting of clear fluid.
Target	The most common symptoms of cholera are dehydration and fever.
Prediction	The primary symptoms of cholera are insectivores and vomiting of clear fluid .
ENSimp-NLSimp	
Reference	Give me a large diet malt liquor and a popcorn with extra motor oil.
Target	Geef me een grote graanjenever light en een popcorn met extra motorolie.
Prediction	Geef me een grote cola light en een popcorn met extra motorolie.
Zero-Shot	
Reference	This is a piece of the xindi probe that crashed on earth.
Target	Dit is een stuk van de xindi-sonde die op aarde is neergestort.
Prediction	Dit is een stuk van de sonde op aarde.