Simplification of Swedish Text by Monolingual Machine Translation

Transformation Rules for Simplified Swedish

James Lindeberg
Abstract

Discussions on the concept of "Klarspråk" and "Lättläst" are major topics in Sweden at the moment with efforts towards increasing readability of government organization information in the project "Begriplig Text". In the field of natural language processing, a lot of efforts have gone into text simplification. This thesis examines the possibility of using hand crafted transfer and generation rules to increase the readability of text through monolingual machine translation.

Transfer and generation rules were developed that handle Swedish grammatical structures including explicitivity and word order. The data used in this thesis consist of manually collected data sets from government organization websites and Uppsala University. The translation results were evaluated using manual evaluation and the automatic readability evaluation methods LIX, OVIX and nominal ratio (NR).

Test results show that the developed rules can produce successful transformation on text although not significantly altering the automatic evaluation scores. The results from the manual evaluation show that a higher readability can be achieved with the developed rules although more tests on bigger data sets are needed in order to fully evaluate all rules developed in this thesis.
# Contents

Acknowledgments ................................................. 4

1 Introduction ................................................. 5
   1.1 Purpose ............................................... 5
   1.2 Outline of this Thesis ................................. 5

2 Background .................................................. 6
   2.1 Simplified Swedish ................................... 6
      2.1.1 Defining Simplified Swedish ................. 7
   2.2 Simplification by Machine Translation ........... 8
      2.2.1 Synonym Replacement Models ............... 9
      2.2.2 Alignment Models ............................. 9
      2.2.3 Summarization Models ....................... 10
      2.2.4 Neural Models ............................... 10
      2.2.5 Rule Based Models ......................... 10
   2.3 Evaluation of Simplifications ....................... 11
      2.3.1 LIX ........................................... 12
      2.3.2 OVIX .......................................... 12
      2.3.3 Nominal Ratio .................................. 12
   2.4 The Convertus Syllabus Translator .................. 13
      2.4.1 Transfer and Transformation Rules .......... 14

3 Data & Method ............................................... 15
   3.1 Data Selection ....................................... 15
   3.2 Method Details ...................................... 16

4 Results ..................................................... 20
   4.1 Test Setup & Results ............................... 20

5 Discussion .................................................. 24
   5.1 Data .................................................. 24
   5.2 Results ............................................... 24
   5.3 Rules ................................................. 25

6 Conclusion .................................................. 27

Bibliography .................................................. 28
Acknowledgments

I would first like to thank my supervisor Mats Dahllöf at Uppsala University for his support and patience. I also want to thank my supervisor Anna Sågvall Hein at Convertus AB for her guidance and insightful discussions and granting me the opportunity for this thesis. Lastly, I want to thank Sebastian Schleussner at Convertus AB, for his resourcefulness and technical support.
1 Introduction

1.1 Purpose

The purpose of the thesis is to develop and test lexical and grammatical transformation rules for a rule based translation module that translates from Swedish to simplified Swedish in order to increase readability. The module is a preprocessing module in the machine translation software, the Convertus Syllabus Translator, that translates from Swedish to foreign languages. A small amount of rules have been implemented and tested and indicate that the module can contribute to a higher translation quality. The main part of the thesis consists of increasing the number of transformation rules and evaluate their effect on machine translations. Another question that will be handled in this thesis is if the preprocessing module can increase readability and be seen as a contribution to the ongoing discussion on *Klarspråk*, simplified Swedish.

1.2 Outline of this Thesis

The structure of this thesis is as follows. Chapter 2 introduces the notion of defining simplified Swedish. It also gives a background to simplification methods in natural language processing and the evaluation methods applied to such systems. Furthermore, it includes a background to the Convertus Syllabus Translator system as well as the transfer module for Swedish simplification. Chapter 3 contains a description of the method for rule set development in the system. Data analysis and selection for the experiment are also included in chapter 3. Chapter 4 consists of a description of the test setup and test results. In chapter 5 the results are discussed and the conclusion of the thesis is found in chapter 6.
2 Background

Attempts have been made to define standardized, simplified languages in order to assist writing texts for people with disabilities, second language learners and other under represented groups (Lundberg and Reichenberg, 2008). Both independent organisations and universities as well as government organizations help enforce a good quality of readable material using writing guidelines (Språkrådet, 2014) and laws governing language and discrimination (Språkrådet, 2011). Using these resources as guidelines and inspiration, a form of simplified Swedish will be described in this thesis. These resources are very useful both for producing new material for all language users as well as forming a baseline for what can be achieved when applying simplification methods through natural language processing.

Several different natural language processing systems exist that in various ways try to simplify source texts for various reasons. A brief discussion on different simplification methods will be provided in this thesis with a focus on rule based machine translation, which the Convertus Syllabus Translator is mostly built upon. Further description of the Convertus system is provided in this thesis along with an in depth look at the different parts of the system in use, heavily focusing on the translation rules for simplifying text which are used to perform the experiments in this thesis.

2.1 Simplified Swedish

In order to develop a language simplification system, the term simplified needs to be defined. Different approaches have been made to create standardized languages for various reasons. Some of these approaches have been made accompanying the development of specific simplification models in natural language processing as part of the process of developing the models. The overall consensus has been that a simplified version of text is always more readable than the source material. One definition of text simplification that will be applied in this thesis is that of Siddharthan and Angrosh (2014), where the importance lies on the fact that the produced simplified version still retains the original information, content and meaning. This distinction is of great importance and will be further discussed in Section 2.2. Other sources used for defining simplified languages are Wikipedias guidelines for writing simple articles and using Simple English (Simple English Wikipedia 2018; Wikipedia: Simple English Wikipedia 2018) along with the CLOUT rules described by Muegge (2002). Both of these sources describe the importance of using correct grammar, explicitivity and direct word order in sentence structures. Some conflicting ideas of simplified languages have become apparent when reviewing such sources and will be further discussed later in this chapter.
Fortunately, when it comes to simplified Swedish, the Swedish government has two laws in place that help enforce the quality of languages in the public sector; Språklagen and Diskrimineringslagen (Språkrådet, 2011). The general notion of these two laws is that all Swedish citizens have the right to any and all information produced by Swedish governments and authorities. This means that all information produced by officials should, or must, have a readable version that people with for example, some form of reading disability, cognitive impairment or citizens that are not native Swedish speakers can understand the contents of. These laws are enforced with the help of Språkrådet which regularly give out new issues of Myndighetens Skrivregler (Språkrådet, 2014). This book contains extensive details and rules about how government employees should produce readable text. The contents of the book range from capitalization, web layout and fonts to spelling, grammatical constructions and much more. In this book, different terms are used to define a simplified version of Swedish, Klarspråk and Lättläst. The basis of these two terms is quite similar and they both try to define the same thing, in different terms. The term Klarspråk is more prominent in governmental writing guides while the term Lättläst has been used more in the public sector and by the publishing company, LL-förlaget that focuses on producing material for a target audience in need of texts with good readability (Österlund, 2011).

2.1.1 Defining Simplified Swedish

The need for simplified material has led to the installment of the project, Begriplig Text which includes parts of the Swedish government, different disability associations and several private companies. One of the biggest issues with terms such as Klarspråk and Lättläst is that they have been defined by people without reading disabilities and are somewhat outdated. This is one of the most prominent tasks for the project Begriplig Text, to listen to people that require this type of material and rewrite the rules for what simplified Swedish actually stand for. An extracted compilation of definitions from previous work on simplification methods for machine translation and the government guidelines previously mentioned will be used to redefine simplified Swedish for the purposes of this thesis. This compiled list will later help build the base for the simplification system described in this thesis. The list of what defines the simplified Swedish is provided below.

   * **Explicitivity**: When referring to people, objects or places in text, the use of additional determiners or prepositions is something that can improve readability (Muegge, 2002; Utbildningsmaterialet Skriva på myndighet 2014). Having multiple subjects and objects within clauses can easily cause confusion and being explicit helps to keep them separated. This is also true when dealing with subordinate clauses, inserting a conjunction where one is missing helps with readability.

   * **Word Order**: Using basic and correct grammatical structures and keeping the events within sentences chronological improves readability (Språkrådet, 2014). Reordering and rewriting text to fit with what is the basic Swedish word order (SVO - subject, verb and object) makes the text more understandable.

   * **Main & Subordinate Clauses**: Having long, cramped sentences with several subordinate clauses can make it difficult to keep the entirety of the sentence together. It is sometimes wiser, when dealing with very long and complex sentence
structures, to simply split the subordinate clauses into individual main clauses (Utbildningsmateriale Skriva på myndighet 2014) or include further explicitivity to assist the reader (Språkrådet, 2014).

**Compound Structures/Abbreviations:** In Swedish, some splitting of compound structures can be hard to interpret as in the Swedish example "läs- och hörförståelse" referring to "reading and listening comprehension" (Utbildningsmateriale Skriva på myndighet 2014). This type of construction showcases a difficulty that can occur if a reader cannot discern that the initial word läs- is part of a longer, compound construction "läsförståelse". It is important however to make the distinction between this type of construction and proper name constructions, for example: "Ansvarsnämnden för djurens hälso- och sjukvård" where a simplification of this would be an incorrect representation of the named entity.

**Adverbials:** Making use of adverbials is not something that inhibits readability in itself but following the theme of previous definition on word order, placing them after the main clause is something that makes text more readable (Språkrådet, 2014; Utbildningsmateriale Skriva på myndighet 2014).

**Nominalization:** In Swedish there exists a transformative construction where a verb or adjective can be nominalized into a noun which is referred to as nominalization. This type of construction sometimes makes it hard for the reader to figure out the actions performed in the clause and rewriting such sentences is something that can improve readability (Utbildningsmateriale Skriva på myndighet 2014).

Some general concepts have been excluded from this list due to contrasting findings where no unified consensus on the effects on readability has been reached. One of these contrasting findings is passive to active transformation, which is proposed to increase readability by Språkrådet (2014) while it is mentioned in Utbildningsmateriale Skriva på myndighet (2014) that writing in passive tense is sometimes needed and fitting for some situations. Another example of such findings is that using shorter sentences improves readability (Muegge, 2002; Wikipedia: Simple English Wikipedia 2018). Contradicting to this point, the Swedish writing guides proposes that a writer should alternate between longer and shorter sentences as this improves readability (Språkrådet, 2014; Utbildningsmateriale Skriva på myndighet 2014). Another simplification technique that is applicable is using simple words and a standardized vocabulary (Stadsrådsberedningen, 2011). However due to the limitations of the thesis work, none of the more general concepts mentioned here will be attempted during the development stages.

### 2.2 Simplification by Machine Translation

The demand for high performance machine translations and simplification of texts has resulted in different approaches to improve readability through natural language processing. Some of these methods include synonym replacement models (Abrahamsson et al., 2014), alignment models (Saggion et al., 2011), neural models (Nisioi et al., 2017), summarization models (Smith and Jönsson, 2011) and rule based models (Decker, 2003; Rybing et al., 2010). The different methods all build on the same principle of simplifying original text and produce a more readable version of said text. The distinction of having a simplification model
that retains the original information, content and meaning can be described as being conservative (Feblowitz and Kauchak, 2013) while simplification models that reformat or remove substantial parts of the text can be considered aggressive (Štajner and Popovic, 2016). The different methods are varying in terms of resources needed to fulfill the task of simplification and will be briefly described in this thesis with a more in depth description on rule based systems. The models presented in this chapter all have possibilities of being implemented as either conservative or aggressive while the system in which the experiments performed in this thesis is defined as purely conservative.

2.2.1 Synonym Replacement Models

With synonym replacement as a text simplification method, parameters that define which words are difficult need to be determined. This can be done manually by for example consulting specialized dictionaries containing entries that should be avoided when writing simplified versions of text (Keskisärkkä, 2012). To determine the complexity of words automatically in a text, evaluation methods that calculate sentence complexity such as LIX (Björnsson, 1968) can be used. Evaluation methods such as this help determine which words might be subject to synonym replacement. Another common method of tracing which words in the original text appear as difficult is word frequency lists (Abrahamsson et al., 2014). By order of probability, words that appear more often should in turn be more readable as they are more frequent. When developing text simplification methods that use frequency lists, a large amount of data is needed to firstly gather word sets from a more general data set and then either developing a synonym data set from those words or using previously created synonym dictionaries. The method is then based on using that synonym data set to replace words of low frequency with words of highest frequency and in that way increase readability. Limitations to a synonymity replacement model may become apparent when words initially appearing as difficult do not have a matching simple, more frequent replacement.

2.2.2 Alignment Models

Sentence alignment models (or phrase based alignment models) make use of aligned parallel data sets to find simplified versions of original text (Coster and Kauchak, 2011). Due to the nature of parallel data sets, the data is never a 100 per cent match, since it is meant to work as a substitution for the original text. To counteract this, models are weighted based on probability and matching substrings when attempting to generate sentences or phrases with higher readability (Quirk et al., 2004). Such data sets are however not very common so the initial step in creating an alignment model is often times to procure or create the parallel data sets. This can be done through various methods generally by making use of individual, separate data sets (Bott and Saggion, 2011) or by collecting large amounts of data over time containing similar information (Glavaš and Štajner, 2015). Having to go through this initial step of data collection can make alignment models quite time consuming. This is especially true when creating entirely new systems for previously unexplored languages where the availability of data might be limited.
2.2.3 Summarization Models

With the area of natural language processing always evolving, there is no surprise when certain ideas from different parts of the field overlap. One of these overlaps is the idea of using automatic summarization for text simplification (Smith and Jönsson, 2011). Summarizing models are used to mainly create summarizations of text and extracting important information from larger amounts of data (Jurafsky and H Martin, 2014). The different kinds are mainly distinguished as extract and abstract summaries. Extract summarization models analyses the entirety of the text and extracts the main themes and produces the most important aspects of the original data while abstract summaries build new content from the existing information in the original data. To make use of a summariser as a text simplification method has been done by Smith and Jönsson (2011) where an increase in readability has been reached. A method of summarizing is using vector models to achieve higher accuracy when producing the output data with proximity and distributional relations representing semantic and linguistic relations between word and/or sentence vectors. With this type of method not being initially developed for text simplification, it is important to predefine which parts of the original text is allowed to be summarized and which parts are of most importance to the content. This is especially true when applying abstract summarizing models.

2.2.4 Neural Models

Another attempt at a cross over between parts of the field of natural language processings have been made, including neural models. Neural models were initially developed in the field of biology but were later adapted for natural language processing (Graves, 2012). Neural models today are applied to various parts of natural language processing including text synthesize, speech recognition and machine translation (Jurafsky and H Martin, 2014). A form of neural model that has been used for text simplification is a sequence to sequence neural network model by Nisioi et al. (2017). The model is first trained on training data and later applied to previously unseen data. The neural network model is implemented using several different algorithms, optimization steps and functions to build up neural networks that in turn are used to simplify text. In the model described by Nisioi et al, additional training data consisting of already simplified texts helped improve the overall output of the system. The results of the model adapted by Nisioi et al show good initial results. However with the claims that are made that the model developed is the first one applied for text simplifications, there is great room for development in the specific area.

2.2.5 Rule Based Models

Machine translations can be split into two distinct main categories, statistical and rule based. In both statistical and rule based machine translations, an integral part of translation systems is syntactic parsing. Syntactic parsing processes the input text as a way of understanding the input material and produce translations. Statistical machine translations are built using probability based language models to generate the most probable output text based on applied functions on the input text (Jurafsky and H Martin, 2014). In order to improve results for statistical
machine translation system, large amounts of training and testing the systems, reevaluating and optimizing language models help improve the output translations.

On the other side of the spectrum of machine translation lies rule based models. Rule based machine translations applies a set of defined lexical and/or structural rules used to generate translations (Jurafsky and H Martin, 2014). The lexical rules mainly make up the translation of individual words or phrases while the structural rules can be applied to reformat sentences and clauses in order to for example rewrite sentences where the subject predicate order differs in the source language and the target language. In addition to syntactic parsing as a step in rule based machine translation, some rule based systems rely on statistical machine translation models as a form of fallback for instances where no rules are applied (Sågvall Hein et al., 2003). These rule based machine translation systems mostly rely on linguistic capabilities during the development and implementation of language models to capture sentence structures in source and target languages. There exists cases where rule based language models consist of automatically generated translation rules (Siddharthan and Angrosh, 2014). The possibility of introducing new rules to already existing language models makes rule based models adaptable. The use of rule based models in terms of text simplification has seen a lot of attention for Swedish in particular (Rennes and Jönsson, 2015). The process of simplification using rule based models builds on the notion of applying rules for monolingual translations, meaning translating from difficult language into a simplified version of the same language. This can be achieved through both structural and lexical rules, replacing individual words with an easier counterpart or reformatting of entire sentences and phrases.

2.3 Evaluation of Simplifications

Evaluating machine translation results can be done by implementing automatic evaluation methods or manual evaluation methods. Different manual evaluation techniques have previously been applied when measuring readability and translation quality, including aspects such as fluency, grammaticality, readability and adequacy (Fomicheva and Specia, 2016; Štajner et al., 2015). For the purpose of this thesis, a binary metric will be applied to the produced texts from the tests representing if the result can be distinguished as more readable or not. One of the more prominent automatic methods applied in machine translation is the BLEU method (Papineni et al., 2002). This method makes use of reference translations to measure precision, recall and a combined F-score. This has been applied to methods previously mentioned in this thesis where such reference data exists (Coster and Kauchak, 2011) but will not be applied here. Instead, automatic evaluation methods based on readability will be implemented to analyze the results produced during the work conducted in this thesis along with manual evaluation. These methods include LIX, OVIX and nominal ratio (NR) and will be described in this chapter.
LIX value | Text genre
--- | ---
<25 | Children’s books
25-30 | Easy texts
30-40 | Normal text/fiction
40-50 | Informative text
50-60 | Specialist literature
>60 | Research, dissertations

Table 2.1: LIX-value and genre correlation.

2.3.1 LIX
The LIX evaluation method was developed in the late 60’s by Björnsson (1968) as a means to measure sentence complexity. Although the method has been around for some time, it is still extensively used when evaluating text simplification models in natural language processing. LIX, or läsbarhetsindex (readability index) calculates the amount of words \(w\), sentences \(s\) and words containing more than 6 characters through the formula:

\[
LIX = \frac{n(w)}{n(s)} + \frac{n(words > 6\text{chars})}{n(w)} \times 100
\]

By multiplying the final score by 100, a number is produced that corresponds to a predefined list which is based on the difficulty of different types of texts. A translation of this list into English adapted by Mühlenbock and Johansson Kokkinakis (2009) is presented in this chapter. The list provided shows the general idea of the LIX method, conveying that texts containing shorter words and fewer sentences are primarily defined as more readable.

2.3.2 OVIX
Another type of automatic evaluation model that is used to analyze readability for Swedish is OVIX, “ordvariationsindex” or word variation index (Hultman and Westman, 1977). The formula is as follows:

\[
OVIX = \frac{\log(n(w))}{\log(2 - \frac{\log(n(uw))}{\log(n(w))})}
\]

where "\(w\)" denotes the amount of words and "\(uw\)" denotes the amount of unique words. The produced number correlates to the amount of unique words and indicates the readability of the overall text. A lower number indicates that the text contains few unique (difficult) words making it more readable.

2.3.3 Nominal Ratio
The final automatic evaluation method that will be implemented in this thesis is nominal ratio (NR). NR indicates the ratio between certain types of part of speech tagged words included in the text (Hultman and Westman, 1977). The formula is constructed to divide the words tagged as nouns, prepositions and participles from
pronouns, adverbs and verbs making up an easy category of words and one difficult. The formula to calculate nominal ratio is provided here where $NN = \text{nouns}$, $PP = \text{prepositions}$, $PC = \text{participles}$, $PN = \text{pronouns}$, $AB = \text{adverbs}$ & $VB = \text{verbs}$:

$$\text{NominalRatio} = \frac{NN + PP + PC}{PN + AB + VB}$$

The normal value for NR is 1.0 which represents newspaper text. If the value produced through the evaluation has a lower value, it then corresponds to easier text.

2.4 The Convertus Syllabus Translator

The system that will be used in the work conducted in this thesis is the Convertus Syllabus Translator. The system builds on a previous research project, MATS, conducted at Uppsala University (Weijnitz et al., 2004) which in turn is the product of an upscaled version of the MULTRA system (Beskow, 1994; Sägvall Hein et al., 2003). The original system is a rule based system but has since been implemented with a number of fall back methods for cases where rules do not apply. The main application of the system is the translation of University course syllabi from Swedish to English which is accomplished through a graphical interface on the client side. Other applications of the system include domain specific language models along with further development in an experimental model for Finnish to English using a malt-parser.

The Convertus system is classified as a modular hybrid, where all parts that make up the entire system can be configured individually. The core of the Syllabus Translator is made up of the lexicon system, parsing and transfer and generation rule systems with additional modules for tokenizing, part-of-speech tagging and post-editing. All parts of the translation process (analysis, transfer and generation) basically rely on unification of feature-value structures (Beskow, 1993). Unification-based parsing analyses the input as feature-value structures where an example of one such structure might be the feature "gender" with a value "neutrum". The transfer rules in the system makes use of these feature-value structures to express unification between a set amount of feature-value structures on the source and target side. With these unifications, generation rules are applied to determine the order of the individual constituents to generate a target sentence structure. A more in depth description of feature-value structures is provided further in Section 2.4.1.

The previously mentioned fall back methods basically consist of two parts, syntactical partial parsing and a statistical method. During the parsing process of the input, if no full parse can be achieved, the system generates a full parse through a set amount of partial parses. In edge cases where no full parse can be achieved through the partial parsing, the syntactic representation of the source is projected to the target representation (Weijnitz et al., 2004). The lexical rules that are applied in this process are context based and have priority over the lexical dictionary entries. The statistical fall back system is built in such a way that if no transfer or generation rules can be applied and no parsing can be achieved (neither full or partial) a statistical language model is applied and produces the most probable target sentence.

13
2.4.1 Transfer and Transformation Rules

The module in the Syllabus Translator that handles transfer and generation rules is the previously mentioned MULTRA system. The formalism in which these transfer and generation rules are expressed is based on the formalism defined by Beskow (1993). As it is originally defined for unification based parsing, the formalism is constructed using feature-value structures that represent the source and target text. Additional parts of the rules consist of a label and a possible transfer part. The label of a rule usually denotes either the name of the source text it is set to transform or the type of reformatting the transfer is developed to achieve. The transfer part of the rule contains values from the feature-value structure contained inside of the rule along with the relation between the two, presented with "\(<=>\)". An example of a rule containing such a transfer relation, presented by Beskow (1993) depicting the transfer rule of the Swedish to English relation between "gärna" and "like to" is provided in table 2.2. As depicted in the table, the example shows the relation between the Swedish adverb "gärna" and the English translation with the verb "like" and verb complement "to". Along with the defined tense, "\(<\text{Tense}>\)" for the a full sentence "S", the example also shows that the translation rule contains feature-value structures corresponding to the subject "\(<\text{subj1} & \text{subj2}\)" and the transfer relation between them: "\(<\text{subj1} <= \text{subj2} \& \text{pred1} <= \text{pred2}\)". The example presented here shows the basis of what will later be the simplification rules developed in this thesis.
3 Data & Method

In the previous chapter, a version of simplified Swedish for the purposes of this thesis was defined as well as an introduction to machine translation for simplification purposes. The formalism for defining transfer rules in the Convertus Syllabus Translator was also presented. This chapter consists of the method and data that make up the process of developing the transformation rules and selecting and extracting the relevant data in preparation for development, testing and evaluation.

3.1 Data Selection

The ongoing discussions on "Klarspråk" with the previously mentioned project "Begriplig text" is focusing largely on text produced by governments and authorities. In an attempt to further the readability efforts on government and authority text, the data selection for this thesis was done by accessing governmental organization’s websites and using their public data. The data that was acquired through this process consists of a section of frequently asked questions on tax return from Skatteverket (Vanliga frågor om deklaration 2018), information on regulations for purchasing items online from Tullverket (Tullverket 2018). A third governmental organization, Bolagsverket, was accessed to further provide data for the thesis work. The data collected from Bolagsverket consist of a general About us section from their website (Om oss 2018). In addition to these governmental organizations, another data set was collected from Uppsala University, containing information about the working environment (Arbetsmiljö och lika villkor 2018) and the general study life of a student (Utbildning 2018). This was done with the presumption that Uppsala University has not consulted the same writing guides as the mentioned government organizations, acting as a stilistic outlier. Further information on the data sets are provided in table 3.1. The information in table 3.1 corresponds to the amount of segments for each data set. The segments are collections of sentences of varying length from the original sources and the amount of sentences range from 1 and 4 for each segment. In conjunction to these manually collected data sets, handwritten sentences were produced and are presented in table 3.2. These hand written sentences were created to assist the development of the

<table>
<thead>
<tr>
<th>Data sets:</th>
<th>Amount of segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uppsala University data set</td>
<td>52</td>
</tr>
<tr>
<td>Bolagsverket data set</td>
<td>17</td>
</tr>
<tr>
<td>Skatteverket data set</td>
<td>111</td>
</tr>
<tr>
<td>Tullverket data set</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 3.1: Amount of text segments for the manually collected data sets.
transformation rules and thus consisting of the constructions intended to be handled by the produced transformation rules. This data set was used as a training set for development of the rules while the data sets from government organisations and Uppsala University were only used during evaluation of the developed rule set.

3.2 Method Details

The transfer and generation rules that are applied in the Convertus Syllabus Translator build on the previously mentioned formalism described by Beskow (1993). The formalism has since evolved considerably and through years of service the rule set of the translation system has increased greatly. The majority of these rules have been developed to handle grammatical constructions from source languages to foreign languages with the purpose of increasing translation quality. Successful attempts have been made to develop pre-editing simplification rules with the intention of improving translation quality further. These rules have also been applied in attempts to increase readability through translation from source language which is presumed to be difficult into simple target language where in this case, both the source and target language is Swedish. One of those rules is presented in this thesis in table 3.3 and acts as an example for all rules developed in this thesis.

The rule in table 3.3 is one that handles word order and reorders a source structure with an initial adverbial to a structure where the adverbial is placed last in the sentence. This rule makes use of a predefined generation rule handling declarative clauses and another rule handling declarative clauses with an initial adverbial. The generation rules in table 3.3 are the foundation of the transfer rule which makes use of the two to transfer the generated text to the target representation. As can be seen in the table 3.3 the formalism differs somewhat between the transfer and generation rules. One of the key differences between the transfer and generation rules is that the transfer rules only expresses unification of feature-values where the resulting structure does not handle word order. The generation rules however determine the order between the individual constituents. Like the previously mentioned formalism by Beskow (1993), both the set of generation rules and the transfer rules use feature-value structures. For example, in both generation rules, the feature `<x1 phr.cat>` refers to the input phrase category with the value of `cl` determining that it is a `clause`. The transfer rule in turn makes use of the feature-value structure defined in the generation rule `<x1 adv.in.fund>` to transfer the generated rule to the target representation.
Generation rules:

Label \textit{cl.decl} \\
x1 \rightarrow x2 :
\begin{align*}
&\langle x1 \text{ phr.cat} \rangle = \text{cl} \\
&\langle x1 \text{ cl.type} \rangle = \text{main} \\
&\langle x1 \text{ mode} \rangle = \text{decl} \\
&\langle x1 \text{ pred} \rangle = x2
\end{align*}

Label \textit{cl.decl-adv.in.fund} \\
x1 \rightarrow x2 x3 :
\begin{align*}
&\langle x1 \text{ phr.cat} \rangle = \text{cl} \\
&\langle x1 \text{ cl.type} \rangle = \text{decl} \\
&\langle x1 \text{ mode} \rangle = \text{main} \\
&\langle x1 \text{ adv21} \rangle = \langle x3 \rangle \\
&\langle x1 \text{ pred} \rangle = \langle x2 \rangle
\end{align*}

Transfer rule:

Label \textit{adv.in.fund} \\
Source \langle* \text{adv.in.fund} \rangle = ?x1 \\
Target \langle* \text{adv.in.fund} \rangle = ?x2 \\
Transfer ?x1 \leftrightarrow ?x2

Table 3.3: Transfer rule, adverb initial to final from Swedish to English.
(1) **Subject duplication:**
Source: Han dansar och hoppar.
Target: Han dansar och han hoppar.

(2) **Adverbial reordering:**
Source: När du hämtar medicin ska du ange personnummer.
Target: Du ska ange personnummer när du hämtar medicin.

(3) **Subjunction insertion (and Adverbial reordering):**
Source: Har du en livshotande sjukdom måste du söka läkarvård.
Target: Du måste söka läkare om du har en livshotande sjukdom.

(4) **Explicit infinite verb:**
Source: Alla kommer få tillträde.
Target: Alla kommer att få tillträde.

*Table 3.4: Examples of developed rules, source and target sentences.*

The notion to make use of previously defined rules to create new rules as well as defining partial rules that together create a more complex, complete rule was done through the development of the rules in this thesis. By using previously defined transfer and generation rules as guidelines along with documentation on the formalism and making use of the manually created data set of test sentences, four rules were developed in this thesis. The rules are provided in table 3.4 and their definitions are provided below:

**Subject duplication:** Where two main coordinate declarative clauses are present in a source sentence as in the example in table 3.4, the rule inserts the implicated (missing) subject from the initial clause into the latter. The produced sentence structure of the applied rule in table 3.4, where an increase in readability is achieved through explicitivity.

**Adverbial reordering:** For cases where the source structure contains an initial adverbial, the rule reorders the source structure and in the target presentation places it last in the sentence. In the example source sentence provided in table 3.4 the sentence has an initial adverbial which is transformed (reordered) in the target representation.

**Subjunction insertion:** In addition to the adverbial reordering rule, if the source sentence structure is a declarative main clause and lacks a subordinate clause marker, the third rule developed inserts one. The subjunction insertion is handled with the Swedish word *om* (English *if*) which retains the semantic structure of the source sentence meaning. The objective of this rule is to work in conjunction with the adverbial reordering rule which can be seen in the example presented in table 3.4.

**Explicit infinite verb:** As previously discussed in chapter 2 of this thesis, explicitivity is a good way of improving readability. To achieve this, a rule for handling infinite verbs that are missing an infinitive marker in the source representation was developed. An example of a source sentence missing an infinitive marker is described in table 3.4, along with the target representation. Another rule similar to the one depicted here was also developed in order to handle constructions with adverbs and negations present in an otherwise equal source structure.
The combination of the defined simplified Swedish in chapter 2 and the examples of readability increase through text conversion proposed by Österlund (2011) and Språkrådet (2014) is what motivated the four rules developed in this thesis. The simplified Swedish consists of simplifications agreed on by all sources used to define it, which limits the amount of potentially applicable simplification rules. The writing guides provide examples of text with bad readability that should be avoided and instructions on how to increase readability during the writing process. The instructions on readability increase used to develop the rule set are instructions that overlap with the defined simplified Swedish as well as instructions that are considered conservative enough as to not alter the meaning and semantic values of the source texts.
4 Results

In the previous chapter, the method for creating the simplification rules was described as well as the method for collecting data sets for testing and evaluation. In this chapter, the process for evaluating the readability of source and target data will be described as well as results from the manual and automatic readability evaluation processes.

4.1 Test Setup & Results

In order to compare the readability scores for the target translation, all data sets first had to be evaluated to produce an initial score. The LIX and OVIX values presented here for both the source and target data sets were calculated using an online tool (LIX räknare 2018) which is a free service used to calculate these scores along with several other readability metrics not included in this thesis. In order to calculate NR the source and target data sets needed to be part of speech tagged. As this is a natural order in the modular translation process in the Convertus Syllabus Translator, the tagger used in the translation was accessed independently to collect the part of speech tags for the source and target data sets. Using the collection of part of speech tags, NR was calculated as previously discussed in this thesis. The data from these readability metrics is presented in table 4.1. The data sets represented in this table are three different data sets for each of the government organizations Bolagsverket, Skatteverket & Tullverket, a complete data set which involves all government data sets along with the Uppsala University data set. In the table, a label depicting whether the produced values correspond to either the Source or Target data set and the respective values produced from the evaluation methods.

<table>
<thead>
<tr>
<th>Data set</th>
<th>LIX</th>
<th>OVIX</th>
<th>NR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uppsala University Source</td>
<td>47</td>
<td>63.39</td>
<td>1.0606</td>
</tr>
<tr>
<td>Uppsala University Target</td>
<td>47</td>
<td>63.19</td>
<td>1.4942</td>
</tr>
<tr>
<td>Bolagsverket Source</td>
<td>47</td>
<td>56.81</td>
<td>1.0059</td>
</tr>
<tr>
<td>Bolagsverket Target</td>
<td>47</td>
<td>56.98</td>
<td>1.0871</td>
</tr>
<tr>
<td>Skatteverket Source</td>
<td>40</td>
<td>49.9</td>
<td>1.4819</td>
</tr>
<tr>
<td>Skatteverket Target</td>
<td>40</td>
<td>48.97</td>
<td>1.0081</td>
</tr>
<tr>
<td>Tullverket Source</td>
<td>43</td>
<td>48.09</td>
<td>1.1644</td>
</tr>
<tr>
<td>Tullverket Target</td>
<td>44</td>
<td>48</td>
<td>1.1644</td>
</tr>
<tr>
<td>Complete data set Source</td>
<td>41</td>
<td>53.68</td>
<td>1.0516</td>
</tr>
<tr>
<td>Complete data set Target</td>
<td>42</td>
<td>53.66</td>
<td>1.0565</td>
</tr>
</tbody>
</table>

Table 4.1: LIX, OVIX & NR for source and target data sets.
In order to do a qualitative, manual evaluation, the produced translations were compared to the original data sets in order to filter out sentences that had been simplified through the translation process. Table 4.2 depicts the amount of total segments from the source data sets (complete data set and Uppsala University data set) and data representing the amount of segments that differ from the source representation. Since the translation and generation rules are additions to an already existing grammatical setting previously developed at Convertus, these numbers include both the simplification rules developed in this thesis as well as previously developed rules. Some examples of such rules are hyphenated words, date and time representation, capitalization and spacing of special characters (parentheses, percent sign, quotation marks etc.). These erroneous translations are listed in table 4.2 as well.

With the information in table 4.2, the next step in the manual evaluation was scoring the translation not previously classified as erroneous. From the eleven translated segments, 11 sentences that were translated were given a score of 1 or 0 where 1 represents a simplified, more readable version of the text in comparison to the source sentence and 0 depicts that the translation has worse readability compared to the source sentence. The first 6 sentences are presented in table 4.3 with the scores from the manual evaluation and information on which rules have been applied during the translation. The last 5 sentences are presented in the same way in table 4.4.
<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Results</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>I tabellen nedan ser du när du får ditt slutskattebesked och din skatteåterbäring.</td>
<td>Du ser när du får ditt slutskattebesked i tabellen nedan och din skatteåterbäring.</td>
<td>Score: 0</td>
<td>Adverbial reordering</td>
</tr>
<tr>
<td>När du betalar ska du ange bankgiro 5050-1055 och betalningsmottagare Skatteverket.</td>
<td>Du ska ange bankgiro 5050-1055 när du betalar och betalningsmottagare Skatteverket.</td>
<td>Score: 0</td>
<td>Adverbial reordering</td>
</tr>
<tr>
<td>Stämmer uppgifterna behöver du inte göra någonting.</td>
<td>Du behöver inte göra någonting om uppgifterna stämmer.</td>
<td>Score: 1</td>
<td>Adverbial reordering and Subjunction insertion</td>
</tr>
<tr>
<td>När du beställer en vara från ett land utanför EU ska du alltid betala moms.</td>
<td>Du ska alltid betala moms när du beställer en vara från ett land utanför EU.</td>
<td>Score: 1</td>
<td>Adverbial reordering</td>
</tr>
<tr>
<td>Om du är osäker på om din produkt är ett kosttillskott eller ett läkemedel kan du kontakta Läkemedelsverket innan du beställer varan.</td>
<td>Du kan kontakta Läkemedelsverket innan du beställer varan om du är osäker på om din produkt är ett kosttillskott eller ett läkemedel.</td>
<td>Score: 1</td>
<td>Adverbial reordering</td>
</tr>
</tbody>
</table>

**Table 4.3:** Translated sentences 1-6 from the Uppsala University and complete data set.
<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Results</th>
</tr>
</thead>
</table>

Table 4.4: Translated sentences 7-11 from the Uppsala University and complete data set.
5 Discussion

In this chapter, discussions are provided covering the data and method used in this thesis. This chapter also includes a discussion on the overall performance of the implemented rules, the applied evaluation metrics and their reliability to accurately determine text readability and the fulfillment of the purpose of this thesis. Each of the mentioned discussion topics are divided into individual Sections.

5.1 Data

The motivation for manually collecting and developing data sets for the work conducted in this thesis was heavily influenced by the initial decision of implementing a manual evaluation step. Even though the manual evaluation metric used in this thesis was binary, distinguishing whether a translation was more or less readable through manual evaluation is time consuming. By limiting the data sets in terms of size and amount of segments, this would presumably reduce the amount of translations that needed to be manually evaluated. The smaller data sets further proved this reasoning as seen in the results in table 4.2 where only 7 translations were produced of the 52 existing segments in the Uppsala University data set and 18 out of the 178 segments in the complete data set. These numbers were reduced further during inspection, showing that only 10 out of 18 translations for the complete data set were actual translations and 1 out of the 7 translations of the Uppsala University data set were actual translations. As a means of counteracting these low amounts of actual translations, implementing larger data sets size might help give an increase to these numbers.

Another point on the data set is that the three data sets that were used in the complete data set were all collected from government organization websites. This means that the writers of the initial material should presumably have access to and consulted material on how to produce readable text (Språkrådet, 2014). This is not to say that the writers of the material in the Uppsala University data set are not following any form of regulation for producing readable material. The results in table 4.2 indicate that source material that is presumed to be readable can still be subject to further increase in readability but should not inhibit further experiments in other domains.

5.2 Results

While observing the results of the automatic redability scores in table 4.1 and comparing the scores with the numbers in table 4.2 the variation between source and target languages is not surprising. The lack of change in LIX values for the Uppsala University data set is due to the lack of change of the text from the
translation step. The differentiating values in OVIX and NR (however slight) are due to the previously mentioned erroneous translations in table 4.2 and the fact that only one sentence was actually simplified. Concerning the government organization data sets, the only data set seeing any difference in LIX value is the Tullverket data set. This is due to a combination of two things. Firstly the amount of erroneous and actual translations that have occurred. Secondly, the size of the data set is the smallest of all data sets used which means any changes to the source data affect the values more. The slight changes in OVIX and NR over all data sets are due to erroneous translations and actual translations. The LIX and OVIX evaluator distinguishes hyphens and word with hyphen structures as separate words which is the cause for the results provided here in terms of erroneous translations. As the rule for explicit infinite verbs is the only rule that changes the amount of words. The rule handling explicit infinite verbs is the second reason for the affected scores in table 4.1 being the only rule that alters the amount of words and sentence lengths.

When looking at the results provided from the manual evaluation and the results of the automatic evaluation metrics there are two things that are important to keep in mind. Firstly, using values such as LIX, OVIX and NR all favour shorter sentences and words over more complex and longer constructions. This means that using a conservative simplification model will most likely score higher (worse) than an aggressive one since a conservative system often only reorders and favours explicitivity over reduction of content as a means of increasing readability. Secondly, to build on the idea that the automatic evaluation metrics favour shorter sentences and words, this might not be true when applying manual evaluation. A problem with manual evaluation however, is reference bias, which might inhibit higher or lower scores depending on who is doing the evaluation. For the purpose of this thesis, all manual evaluation was carried out as a step in the experiment which gives a huge bias towards the target sentence structure. A way to get around this issue is to outsource the manual evaluation to actual translators. The problem still remains as different translators might have different ideas of what a more or less readable version of text is. This is one of the biggest problems facing the ongoing work with Klarspråk in Sweden, which have decided to use their actual target audience (persons with reading disabilities) to help better understand what makes a text more or less readable. To use this form of manual evaluation of the produced sentences from the experiments in this thesis could prove useful in the discussions on Klarspråk. Further experiments might indicate how accurately the simplified Swedish defined in this thesis is in terms of readability when involving an actual target audience.

5.3 Rules

The subjunction insertion rule, along with erroneous translations were the only translations affecting the automatic evaluation scores. In table 4.3 and 4.4, 1 sentence was produced with the subjunction insertion rule and 11 with the adverbial reordering rule. All of the target sentences produced in this thesis except 4 were given a score of 1. These results were considered more readable by consulting the simplified Swedish defined in this thesis. Sentences 1, 2 and 5 in
Table 4.3 and sentence 11 in Table 4.4 were all given a score of 0. Sentences 1, 2 and 5 display the same type of unwanted translation result when the object of the source sentences contain conjunctival noun phrases. The rule for adverbial reordering does not handle this kind of conjunctival noun structure as an object but does however apply and only partly reorders the source sentence. This is most likely caused by a lack of specificity in the rule but could also be due to the parsing process. As the parsing is set up, the rules apply even though a full parse might not be achieved as previously described in this thesis. The results from this becomes apparent in the target sentence structures in Table 4.3. For sentence 11 in Table 4.4, a more severe grammatical error has occurred due to the adverbial reordering rule. In this case, it is not due to a conjunctival noun identification error but rather an erroneous identification of the constituents of the source sentence structure. A correct translation of the sentence should be: "Du får arbeta praktiskt med det du lärt dig på laborationerna under föreläsningar och lektioner". One of the reasons for this might again be an error during the parsing process. A more probable reason is that the prepositional phrase of the source sentence is not recognized properly during the translation and application of the adverbial reordering rule. To determine the cause for these mentioned incorrect translation, further experiments need to be carried out to cancel out whether the rules or the parsing is reason.

Further analysis of the results in 4.3 and 4.4 indicate that only 2 of the developed rules had any effect on the source data sets, the adverbial reordering and subjunction insertion rules. The other rules, subject duplication and explicit explicit infinite verb worked as intended on the manually created test sentences but did not apply during any of the translations in the experiments. As previously mentioned for the rules that did apply, however sometimes incorrectly, the issue again might be caused by incorrect parsing. Although a more probable reason for the rules not applying during the translation is the lack of matching source sentence structures. If the source data does not contain any structures that equal to the structures defined in the rules, a translation of that source sentence will never occur. As the collected data sets were not manually evaluated before nor after the experiments conducted in this thesis, the data sets could in theory contain such structures however not very likely. There are two measures that can be applied to fully expand on this hypothesis further, evaluation of the manually collected data sets used in this thesis or redoing the same experiments on different data set altogether.
6 Conclusion

The purpose of this thesis was to develop and test transformation rules in the Convertus Syllabus Translator to increase readability through monolingual machine translation. Although all rules described in this thesis produce target sentence structures as intended during development, the lack of source sentence structures inhibit some rules from transforming sentences. The results from the automatic and manual evaluation from the tests in this thesis show that the majority of target sentence structures produced sentences with higher readability. Due to the fact that all developed rules did not apply during evaluation, with further development of new rules, improvement to the rule set developed in this thesis and using different data sets, the results of manual evaluation has the potential of increasing further. However, achieving better scores using the automatic evaluation methods described in this thesis might not be possible with attempts at increasing readability using the conservative simplification model developed in this thesis.
Abrahamsson, Emil, Timothy Forni, Maria Skeppstedt, and Maria Kvist (2014). “Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language”. In: 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR).


Feblowitz, Dan and David Kauchak (2013). “Sentence Simplification as Tree Transduction”. In: 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations.


Glavaš, Goran and Sanja Štajner (2015). “Simplifying Lexical Simplification: Do We Need Simplified Corpora?” In: 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers).

Graves, Alex (2012). Supervised Sequence Labelling with Recurrent Neural Networks. Springer.


Jurafsky, Dan and James H Martin (2014). Speech and Language Processing. Pearson Education.


