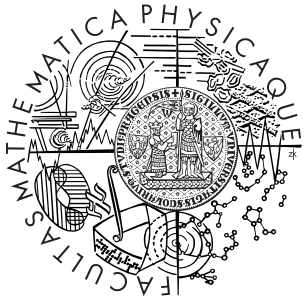
Charles University, Prague
Faculty of Mathematics and Physics

Master Thesis



Magdalena Prokopová

Automatické zjednodušování textů pro překlad

Automatic Simplification of Texts for Translation

Institute of Formal and Applied Linguistics

Supervisor: RNDr. Vladislav Kuboň, Ph.D.

Study program: Informatics, Computational Linguistics

First of all, I would like to thank my advisor Vladislav Kuboň for his constructive suggestions, observations, and help with writing. Also to my family and friends for their support in my life and studies, and to my boyfriend for his valuable corrections.

I certify that this master thesis is all my own work, and that I used only cited literature. I agree with making this thesis publicly available.

In Prague on August 13, 2007

Magdalena Prokopová

Contents

Název práce: Automatické zjednodušování textů pro překlad
Autor: Magdalena Prokopová
Katedra (ústav): Ústav formální a aplikované lingvistiky
Vedoucí diplomové práce: RNDr. Vladislav Kuboň, Ph.D.
e-mail vedoucího: Vladislav.Kubon@mff.cuni.cz

Abstrakt: Tato práce se zabývá využitím automatického zjednodušování (simplifiakce) textů pro účely automatického překladu. Práce srovnává automatické zjednodušování a kotrolované psaní, co mají společného a jaké jsou jejich rozdíly. Dále se zaměřuje na zjednodušování v souvislosti s automatickým překladem. Je popsáno jaké problémy může zjednodušování vyřešit a pro část z nich je řešení navrženo. V rámci práce byl implementován systém ASOFT, který provádí vybrané transformace na větách. Výsledky systému ASOFT ve spojení se systémem automatického překladu PC Transaltor byly vyhodnoceny pomocí několika metrik, zvolili jsem automatické vyhodnocení pomocí BLEU a NIST a vyhodnocení provedené lidskými anotátory. V závěru nastiňujeme jakými dalšími způsoby by se dané téma mohlo rozvíjet.
Klíčová slova: automatické zjednodušování, automatický překlad, kontrolované psaní

Title: Automatic Simplification of Texts for Translation
Author: Magdalena Prokopová
Department: Institute of Formal and Applied Linguistics
Supervisor: RNDr. Vladislav Kuboň, Ph.D.
Supervisor's e-mail address: Vladislav.Kubon@mff.cuni.cz

Abstract: This thesis describes one of the areas where automatic simplification can be used: simplification of texts for machine translation. We start by comparing methods of automatic simplification and controlled language, describing their similarities and differences. Further on we focus only on automatic simplification used as a preprocessing step for machine translation. We describe what issues can be solved and address some of them using our own system ASOFT. A text preprocessed by ASOFT is intended to be translated by a machine translation system PC Translator. We evaluate the output of the PC Translator using two automatic methods, BLEU and NIST scores, and one method of human evaluation. In the end we propose other issues that can be addressed by means of automatic simplification.
Keywords: automatic simplification, machine translation, controlled language

Chapter 1

Introduction

1.1 Motivation

The following quotation can be seen as one of the historical motivations that generated interest in **Machine Translation (MT)** by proposing this task is feasible. In 1947, Warren Weaver of the Rockefeller Foundation proposed to a friend: *"I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text."* Although this statement did not prove to be 100% true, because several different correct translations can exist for a sentence as opposed cryptography where only one can exist, machine translation became a research topic of growing interest.

In recent years more and more information has become available online and thus accessible by wide variety of people all over the world. When resources become available the challenge is then to make this information more understandable for a wider range of audiences. This is where machine translation can play a key role.

There can be several goals for MT, ranging from the idea of getting the basic meaning of the text through translating technical documents all the way to the translation of books and novels. Czech, as a language with free word order but a large scale of word forms, can become an issue for MT systems. Since Czech translation systems are less advanced than MT systems for more widely spoken languages, simplifications in the input grammar can bring important improvements to the results.

This thesis aims to introduce and summarize state-of-the-art approaches

to text preprocessing for MT, namely by means of **controlled language (CL)** and **automatic text simplification (ATS)**. Later in the text we focus on the main task of this thesis, text simplification of Czech language. We point out specific areas where preprocessing can make the translation of a Czech text more understandable and we address some of these issues. In the end it is shown how simplification of text can improve the quality of Czech to English MT by applying standard metrics.

In the next sections we are going to cover basic approaches to MT and we are going to mention how MT can be evaluated. In the end of this chapter, the state-of-the-art systems for Czech translation will be described and briefly evaluated.

1.2 Statistical and Rule-based Machine Translation

The most basic approach to MT is simple word for word translation. Although this approach does not give impressive results it can give the reader a basic idea of what the foreign text is about. The field where word for word translation gives the most satisfying results is translation between languages with close or similar syntactic structure. System Česílko [Hajič et al., 2000], translating between Czech and Slovak is one of many examples of this approach to translation.

Leaving out word for word translation, most of current systems can be divided into two major groups according to the approach they are taking: statistical, or rule-based. **Statistical machine translation (SMT)** systems have two main components: language model, and translation model. The language model uses a monolingual corpus to learn probabilities of words or groups of words. The translation model is built from a bilingual corpus, aligned by words. This alignment can be a part of the system if such resources are not available. Several measures can be counted from word aligned corpus: probability of translation for words and groups of words, fertility of word (to how many words it usually translates) and distortion (whether the word changes position in the sentence e.g. adjectives being before or after noun.) All this information can be included in SMT systems to achieve the best possible results.

The state-of-the-art systems usually use models based on phrases rather than single words, since it can give more accurate results. Probabilities of

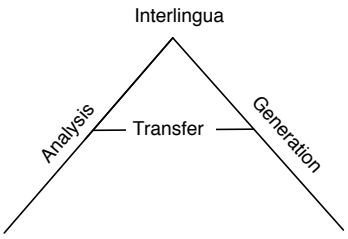


Figure 1.1: Schema of a transfer based translation process used by RBMT systems

all possible translations for a phrase are compared and the most likely one is chosen as a result. Since statistical approach needs a bilingual corpus for creating the model it can be used only for languages with such resources available. There exists several large bilingual corpora for the English-French language pair, but such extensive bilingual corpora are rare for less widely used languages.

For more information about SMT translation and phrase-base models please refer to the latest SMT systems: Pharaoh [Koehn, 2004], and its replacement Moses [Philipp Koehn, 2007].

On the other hand, the idea behind **Rule based machine translation (RBMT)** is to find a language independent representation of text which can be later on generated into sentences of a target language. Such a representation is referred to as *interlingua*. This would be the ideal situation, but it seems unreachable so far, thus the translation is usually divided into 3 parts as shown on Figure 1.1.

During the *analysis* process the source text is transformed as closely as possible to an interlingual representation; morphological and syntactical analysis is performed, and, where needed, text is often disambiguated. Text, along with all the information obtained in the previous step, is *transfered* into the target language and the process of *generation* creates text in a readable form.

RBMT systems relies on grammar and lexical rules used for the transfer faze, but the hand-crafting these rules is tedious and error-prone work. Therefore the issues of RBMT is obvious, development of appropriate large scale grammatical and lexical resources needs an enormous effort from human annotators and translators. More information about RBMT and other appro-

aches to MT can be found e.g. in [Hutchins, 1986].

Comparing these two approaches, SMT is more universal and thus, once the system is developed, it can be used without significant changes for any language pair where enough training resources can be found. In most cases the bigger the training set that is available for the language pair, the better the results that are obtained. On the other hand when using the RBMT approach, it is essential to specify the language pair and the possibility of reusing the same rules for another language pair is limited. Improvements can be solely achieved by adding new rules or making existing rules more exact. Since the system is already devoted to one language pair it can focus on addressing issues appearing only for this language pair. Usually the pool of rules is rather large and therefore adding new rules that do not interfere with existing ones can often be a feasible task for only one person or a small development team.

Several experiments have been made that combine both approaches to further improve results but without major success. Adding rules to SMT can even lower the quality of translated text.

A promising **context-based approach (CBMT)** was presented by Meaningful Machines, LLC group in [Carbonell et al., 2006]. This approach needs neither parallel text nor an enormous set of rules. Target language corpora is used to create a language model that stores probabilities of target language n-grams. The translation process starts with dividing each sentence into n-grams, and for each n-gram all the possible translations are generated based on a bilingual dictionary. Every n-gram can have hundreds of possible translations since there can be several different translation possibilities even for each word. For each of these newly generated n-grams, probabilities are retrieved from the language model and the most likely translation is chosen.

1.3 Metrics

After several different MT systems for the same language pair had been created it was necessary to introduce a way how to evaluate the translation systems' outputs. However only if a computer knew all the rules of the target language it would be able to perfectly distinguish whether the translation is correct or not. Therefore we can say that the problem of evaluation is as difficult as the problem of translation. Nevertheless several different techniques for evaluating translations were introduced all based around a common principle; that the closer a machine translation is to a professional

human translation, the better it is. Therefore most of the automatic evaluation techniques use one or more human translated texts as a reference translation to compare the system output to.

Human Evaluation

By giving the translated text to a human evaluator we can see immediately if the translation is correct and where any mistakes were made. Humans are able to take into account synonyms and thus evaluation is more exact. Last but not least, humans can distinguish between fluency and adequacy and therefore it is possible to see which areas need more of the future research.

However human evaluation has its shortcomings. The most evident one is the cost. Moreover human evaluations show surprisingly low levels of correlation between different evaluators and in several cases not even humans can agree which translation is better. Therefore several computer based evaluations were proposed to lower the cost of evaluating translation results and to make the evaluation more objective.

For more information about human evaluation refer to the Framework for Machine Translation Evaluation in ISLE (FEMTI) [King M., 2003]. FEMTI is an attempt to organize and categorize the various methods that are used to evaluate MT systems. This framework consists of two parts: the first part categorizes the purpose of the evaluation, and the second classifies the evaluation methods hierarchically, with detailed description of how each characteristic of the system should be evaluated.

Precision

Precision is based on counting the words that appear in both a generated sentence s and a reference translation r (or one of the reference translations) and then dividing this number by the number of words in the generated sentence. Since computer generated sentences tend to over-generate words, precision metrics can give better results when instead of counting the simple intersection of words, just one match per word is allowed.

$$Precision(S|R) = \frac{S \cap R}{|R|}$$

Position Independent Error Rate

PER (Position independent error rate) is based on word error rate metrics. Word error rate evaluate a translation by counting the edit distance, that being the minimal number of operations (deletions, insertions and substitutions) to obtain the reference text from the generated translation. PER ignores the word order by treating the sentence as a collection of words and using just insertion and deletion.

$$PER(S,R) = \frac{diff(S,R) + diff(R,S)}{|R|}$$

BLEU

BLEU (Bilingual Evaluation Understudy) To check how close a candidate translation is to a reference translation, BLEU score uses n-gram comparison between both translations. It measures n-gram co-occurrence in the translated text and the set of reference translations. The regular precision measure is used, but modified and instead of words the precision is counted on n-grams. This covers both fluency, by checking the right word order, and adequacy by checking correct translations of words.

Moreover the *brevity penalty (BP)* is introduced. The translated sentence should also match the reference sentence in the length (Sentences that are too long are already penalized by n-grams,but there is need to also introduce a penalty for those that are too short) and therefore if the reference text is longer than the translated text a brevity penalty is counted over all corpus to diminish the resulting BLEU score. For more details see [Papineni et al., 2001]

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} \frac{log(p_n)}{N}\right)$$

NIST BLEU

An improved implementation of the BLEU score was introduced by NIST (the US National Institute of Standards and Technology). It is showing a high correlation with human evaluations than simple BLEU. This measure has become one of the most commonly used, and a de-facto standard. Instead of counting all n-grams equally, the information gain from each n-gram is

taken into account. Those n-grams that appear less often in the text are predicted to contain more information and hence have higher importance.

$$NIST = \overline{BP} \cdot exp\left(\sum_{n=1}^{N} w_n\right)$$

In this formula w_n combines the precision and importance of each n-gram. For more details on NIST scoring see [Doddington, 2002]

Although NIST BLEU is considered a standard for evaluation it can not take into account grammatical and semantic synonymies. The more reference translations that are available, the more accurate the scoring is, but unfortunately often just one human translated copy is available.

F-Measure

F-measure, introduced in [Turian et al., 2003], uses the idea of "maximum matching" from graph theory and takes advantage of metrics that are widely used to evaluate NLP systems: precision and recall. The word sequences appearing in both the reference translation and the text translated by the MT system are counted. Starting from the longest one to the shortest, sequences are added in such a way that no token is ever counted twice. Although calculating Maximum Match Size (MMS) is NP hard problem, greedy algorithms exist that find the true maximum in 80% to 99% cases.

Recall and Precision are defined on the maximum match.

$$Precision(cand|ref) = \frac{MMS(cand, ref)}{|cand|}$$

$$Recall(cand|ref) = \frac{MMS(cand, ref)}{|ref|}$$

To reward correct word order, it is necessary to reward sequences more than linearly in proportion to their length. Here the reward for longer matches is defined as the square root of the squares of individual lengths. The final F-measure is the harmonic mean of both precision and recall.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

[Turian et al., 2003] is claiming to have higher correlation with human evaluation than either BLEU or NIST has.

Sentence Level Evaluation

All the above mentioned metrics fail to achieve satisfactory results on evaluating single sentences and therefore a novel method was introduced by [Kulesza and Shieber, 2004]. This method classifies sentences as either human translated or machine generated, and gives a certainty value for this classification.

This method is based on a machine learning technique used for classification, Support Vector Machines (SVM). Moreover this technique does not need a reference translation to evaluate a sentence. When SVM is trained on one reference translation text and one machine generated translation text it produces a separator (a hyper-plane that divides examples into human translated half and machine generated half) that is later used to separate translated sentences. We can count which half of the example space our sentence falls into. By counting also the distance from the separator (by removing the sign operator in the expression $sign(< w, x > *b)$) we can obtain a score of how good the translation is.

1.4 State-of-the-art Systems for Translations Between Czech and English

All state-of-the-art systems have something in common, they can be priceless for people who do not know the target language at all, but on the other hand the translated text cannot usually be used without further post-editing.

1.4.1 PC Translator

PC Translator is a commercial product developed by the company LangSoft. This system is a standalone Windows application that provides the user with the possibility to translate text to and from English. Its translation dictionary contains more than 800 000 words and phrases and nowadays it is one of the most widely used systems for people speaking Czech but not English.

We compared the quality of the two translation directions and this comparison showed that the quality of results for English to Czech translations is much higher, but still there is a lot to be improved. The Czech to English translation process does not perform deep linguistic analysis of the source sentence and with exceptions of predefined phrases it is close to simple word

Direction	Data	BLEU	NIST
Cz–En	DevSet	0.2355	7.2739
	EvalSet	0.2057	6.7499
En–Cz	DevSet	0.2226	5.3397
	EvalSet	0.2014	4.9422

Table 1.1: BLEU and NIST score for PC Translator

for word translation. The resulting text does not even follow some of the basic rules of English grammar.

We evaluated translation results on Prague Czech Dependency Treebank [Čmejrek et al., 2004] using 5 reference translations for Czech to English and one reference translation for English to Czech. PCEDT contains two data sets for training and evaluation purposes, and we will refer to these shortly as *evalSet* for the evaluation set and *devSet* for the development set. Although it has been shown that neither BLEU nor NIST scores can provide an unbiased comparison of systems, it can give us at least some idea of system quality.

The evaluation results on the development and evaluation data sets are summarized in Table 1.1. However since only one reference translation was available for translations from English to Czech, the results presented in this table do not support the statement that this translation direction is better. On the other hand this shows that comparing BLEU and NIST scores for different translation directions or even different systems does not necessarily prove which system is better.

Further on in this thesis we will use the latest version of the system – PC Translator 2007; since it was shown that this system gives similar or better results than other commercial and noncommercial systems.

1.4.2 Eurotran / Wordmaster

The Company Microton offers two commercial translation solutions based on the same technology. Eurotran is usually run as a part of a web browser and offers interactive translation of English web pages. Wordmaster is a standalone application which offers translation in both directions, English to Czech and Czech to English, though it is intended to assist with a translation rather than complete one. It offers list of possible translation for each word

Direction	Data	BLEU	NIST
Cz–En	DevSet	0.1925	6.0996
	EvalSet	0.1361	5.2837
En–Cz	DevSet	0.2214	5.3051
	EvalSet	0.1937	4.9197

Table 1.2: BLEU and NIST score for Wordmaster

or phrase and user should select the most suitable one, otherwise it just chooses the first available one.

Both Eurotran and Wordmaster do not translate text on a local machine, but instead connect to a translation server. This server contains approximately 500 000 dictionary entries. Although the dictionary size is smaller than the one of PC Translator the translation from English to Czech is slightly better and the grammar rules for subject-verb agreement are more accurate.

We performed the same evaluation test on PCEDT data as for PC Translator. The results of the evaluation are summarized in Table 1.2.

1.4.3 Pharaoh

As an example of non-commercial system we chose the statistical system Pharaoh. Pharaoh, trained for translating from Czech to English was described in [Cuřín, 2006]. This system was trained on three parallel corpora: The Prague Czech Dependency Treebank (PCEDT), the Readers Digest Corpus and the corpus of IBM operating system manuals. Afterwards the systems was trained on data that were enriched with morphological information and preprocessed. The evaluation of retraining showed improvements in translation.

1.4.4 Comparison

In general the task of comparing different MT systems is challenging and is not within the scope of this thesis. Machine evaluation techniques are more suitable for evaluating different versions of one system than for comparing different systems, but not even human evaluation is suitable for the task and can not ensure maintaining objectivity and include all characteristics of the system in one evaluation.

	Eurotran	PC Translator	Skik
Understandability	2.8	2.9	3.2
Lexical quality	1.9	2.0	2.4
Grammatical quality	2.2	2.4	2.6

Table 1.3: Human evaluation of systems

Although each translation system has its shortcomings, evaluation of most common translation systems showed that the quality of systems is comparable.

Human Evaluation

[Cintl, 2000] describes in his thesis the comparison of the following translation systems: PC Translator 2000, Eurotran 98 and Skik 5.0. Short texts were given to 12 evaluators, who were asked to perform two tasks on the text. Firstly they needed to evaluate the resulting texts' understandability without knowing what the source was. Secondly, having both source and result text, they had to evaluate the quality of the translation by looking at both lexical and grammatical aspects.

Each evaluation was assigned a score from 1 to 5, where 1 denoted best and 5 worst. The results are shown in Table 1.3

As we can see the results of the systems are similar with exception of Skik that seems to be worse than other two.

Automatic Evaluation

Since the published comparisons of MT systems does not apply to most the recent versions of these systems we performed a short evaluation using the BLEU score. We focused only on translation from Czech to English and the comparison of systems mentioned earlier in this chapter is shown in Table 1.4.

	DevSet	EvalSet
PC Translator	0.2355	0.2057
WordMaster	0.1925	0.1361
Pharaoh	0.3858	0.3650

Table 1.4: BLEU score of Czech to English translation

Chapter 2

Simplification versus Controlled Language

Long and complicated sentences pose various problems to many state-of-the-art natural language processing technologies. In this chapter we are going to focus on ways to make the process of Machine translation (MT) easier and hence help MT systems to improve the quality of resulting text. In the following paragraphs we explore both how the user can help when writing text using methods of **Controlled Language (CL)**, and how text can be processed without any help from a user by means of **Automatic Text Simplification (ATS)**. Although both CL and ATS are trying to address similar issues their approach shows significant difference.

By explaining what problems can be addressed in general we will move to the problems that are more specific for the Czech language and translations between Czech and English.

2.1 Controlled Language

When addressing MT problems by means of CL we usually speak about authoring systems. These systems check for grammar errors and ambiguities at the moment when a text is written. Such system can notify a user of potential ambiguities and ask the author to rewrite the sentence. This mechanism helps the user to write simple and non-ambiguous text that will not contain phrases and structures that can cause problems to MT systems. CL has two main objectives. One is authoring an input for MT systems. The other is to assist author in publication of controlled text which is not

meant to be translated, but rather intended to be easy to read (e.g. for non native speakers). In this case the focus is on grammar, concise sentences, clarity of sentences and consistency of vocabulary used, which helps readers to understand the document.

The controlled text should not have more than one interpretation (in a given domain). All ambiguities should be resolved when the text is written. Disambiguation can be addressed in two different areas, grammar ambiguities and vocabulary ambiguities.

2.1.1 Controlled Grammar

By controlling the grammar the results of MT can be improved, as well as the text can be standardized and its readability improved. Possible grammatical ambiguities can be identified in two different ways: Either we can use a finite set of negative patterns, or implement complete grammar in the system. Both of these approaches were found to be restricting and therefore a combination of pattern matching and analysis by a parser is said to be the best way to identify problematic sentences.

This approach was addressed especially in project KANT [Mitamura, 1999].

Issues Addressed by Controlled Grammar

As a first source of ambiguities we will look at *coordinations*. Both arguments and modifiers of verbs in coordination can be potentially ambiguous and thus when used can make the text unclear. The following is an example of two possible meanings for one sentence.

(Push) and (pull the rod) ***X*** *Push and pull (the rod)*

And similar example in Czech follows [1]

(Pøujčoval si a četl) knihy ***X*** *(Pøujčoval si) a (četl knihy)*

(He-borrowed <Refl>and he-read) books **X** *(He-borrowed <Refl>) and (he-read books)*

[1] When necessary Czech examples are translated in a way that every Czech word is translated into exactly one English word.

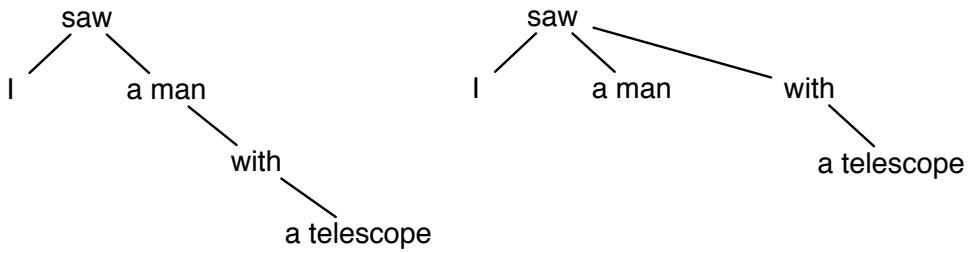


Figure 2.1: Two possible parse trees for a sentence "I saw a man with a telescope."

Prepositional phrases can potentially modify any word in the sentence and are considered to be the most ambiguous construction.

The following sentence is a well known English example, where two interpretations are possible. Both interpretations are shown on Figure 2.1.

I saw a man ***with a telescope.***

We can find similar sentences also in Czech, two possible interpretations of the following sentence are shown on Figure 2.2.

Zaútočil na postavu ***s kopím****.*

He-attacked at a-man with a-spear.

Further on we should mention *phrasal verbs*. Prepositions used with phrasal verbs are often ambiguous and therefore when rules of CL are applied verb synonyms that are not phrasal should be used instead.

2.1.2 Controlled Vocabulary

The task of controlled vocabulary is to limit the number of possible meanings per word. Although the simplest way to avoid any vocabulary ambiguities is allowing only one meaning per word, such a restriction is too difficult to follow. Authoring systems can usually provide the possibility of selecting the correct meaning for each word in the sentence. When only one meaning per word is specified, MT systems can easily choose the appropriate translation.

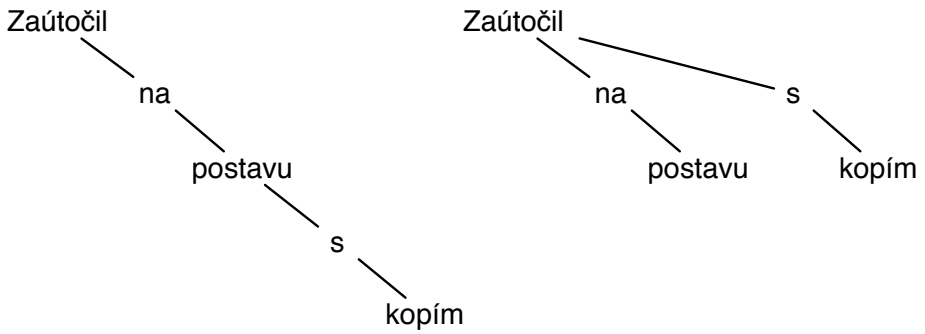


Figure 2.2: Two possible parse trees for a sentence "Zaútočil na postavu s kopím."

However this feature must be supported by both authoring system and MT system.

Although nouns are usually considered the biggest source of ambiguities, when techniques of controlled vocabulary are applied, all words that are possibly ambiguous need to be clearly marked. This includes for example modal verbs. Moreover acronyms and abbreviations in a technical controlled text should be well designed to avoid ambiguities with other words of the language. If this is not possible, again, the correct sense needs to be clearly marked.

There are practical limits to a restricted vocabulary. In a system where vocabulary is extremely limited, the author needs to write long, convoluted sentences to express complicated meanings and thus the text becomes less readable. Therefore the challenge is to make the vocabulary only as broad as necessary but not any broader.

2.1.3 Controlling on Discourse Level

So far we have focused only on controlling text on a vocabulary and syntax level, but CL can be even extended to the discourse level. New techniques have been introduced to help create texts that are easy to read, even for non native speakers, and easy to translate. [Bernth, 2006] addresses the issue of style and proposes that the text should have a *good rhythm*. This means sentences should be of the similar length and the selected rhythm

of alternating declarative, interrogative and imperative sentences and noun phrases should be maintained through out the whole text.

[Bernth, 2006] addresses the issue of a bulleted list. A Bulleted list poses problems to MT systems since items of the list are not complete sentences, but form the whole sentence when combined with the lead in sentence. Thus CL can be implemented in such a way that it controls whether the bulleted list is formed correctly.

The last issue addressed on the discourse level is topics. A well formed paragraph should contain exactly one topic. Moreover the topic of the paragraph should be introduced by a *topic sentence*. Expanding CL from sentence level to the discourse level is certainly an interesting question and introduces a lot of new subjects to be addressed.

2.1.4 Controlled Language Rules

An example of guidelines for writing controlled text is presented in [MUEGGE].

- Write sentences that are shorter than 25 words.
- Write sentences that express only one idea.
- Write the same sentence if you want to express the same content.
- Write sentences that are grammatically complete.
- Write sentences that have a simple grammatical structure.
- Write sentences in the active form.
- Write sentences that repeat the noun instead of using a pronoun.
- Write sentences that use articles to identify nouns.
- Write sentences that use words from a general dictionary.
- Write sentences that use only words with correct spelling.

2.1.5 Example System

A Spanish to English CL system was developed by [Sammer et al., 2006] and it was integrated with a SMT system. This system was created to help email-users not speaking the same language to exchange information. While writing a message a user is offered multiple meanings for each word and is prompted to choose a single one. Additionally he has the possibility to mark the word as a proper noun and thus not to be translated.

As a part of the project a word sense lexicon was generated from a bilingual dictionary and a machine readable dictionary for the target language. (WordNet was used as a machine readable dictionary and extra rules for pronoun disambiguation were created). The system focuses on controlling a user's input when he writes an e-mail. Email was chosen because it is impossible to distinguish in advance the domain of a message, and often terms from several domains can be used together. Therefore it is essential to disambiguate word by word and not to try to place the whole message in one domain.

Results of the experiment showed improvement in the translation. Overall improvement in the BLEU score was 0.9% (from a baseline 21.7% to 22.6% using CL). For the BLEU and PER metrics the biggest impact resulted from indicating proper nouns, on the other hand the most interesting part of the task, word sense disambiguation by itself, improved the BLEU score just by 0.2%. Position independent word error rate lowered from 45.0% to 44.3%. The shortcoming of CL is evident, placing the burden to mark words by their meaning on a user.

2.2 Text Simplification

Although the word "simplification" can be slightly misleading, text simplification is not summarization. Summarization aims to reduce an information content and generally increases syntactic complexity. On the other, hand text simplification attempts to make the text easier to read and process using **Natural Language Processing (NLP)** applications.

According to the most simple definition, automatic text simplification (ATS) is a set of transformations on a given text that aims to reduce syntactic and lexical complexity and at the same time preserve the meaning and the information content. The only difference between ATS and Controlled Language (CL) is that for ATS transformations are automatic, i.e. done by a

computer without any (or only a little) human supervision. Similarly to CL, ATS can be used as a preprocessing step for MT, and moreover a text can be simplified to increase its readability, e.g. for people with reading difficulties.

Further on in this thesis we describe in more detail how ATS can be used as a preprocessor for MT. The general idea is to focus on internal rewriting rules, specific for the source and target language, and use them to create simple and standardized text. The text preprocessed for MT does not necessarily need to be in a human readable form since it is not expected that anybody would read it before translation, and therefore an input sentence can be transformed so that it follows some of the grammar rules of the target language.

Moreover text simplification can be viewed as semantic simplification, words can be disambiguated according to the domain of the text or the paragraph. However this interesting issue is not addressed within the scope of this thesis.

2.2.1 Automatic Induction of Rules

The task of ATS is often addressed on the sentence level, because long and complicated sentences may cause difficulties to a MT system. Therefore simplifying these sentences is supposed to improve system results.

Since the hand-crafting of rules for text simplification is usually a time consuming and tedious task, [Chandrasekar and Srinivas, 1997] proposes an algorithm to obtain simplification rules automatically. By applying these automatic rules to long and complicated sentences with relative clauses (what, which, etc.), text composed of short and simple sentences can be created. Although only the issue of relative clauses is addressed , other syntactic structures can be handled in a similar way.

The solution, [Chandrasekar and Srinivas, 1997] proposes, is based on an annotated aligned corpus of complex and simple text, where each complex sentence is linked to a corresponding manually simplified sentence. From these, resources rules are automatically induced. During the training process both complex and simplified texts are syntactically analyzed to the form of dependancy trees. Further on, articulation points are identified, i.e. points where sentences may be split for simplification. Subsequently, the algorithm computes tree-to-tree transformation rules to convert source sentences into simplified sentences. Words are replaced by variables leaving only the morphological and syntactical information gathered during the analysis, this

way rules are generalized to match more varieties of sentences. In addition to finding articulation points, gap filling routines are specified to complete new short sentences where the head noun is missing. Although this can lead to awkward sounding text, inserting pronouns would need complex knowledge of the discourse.

Even in this narrow domain, an automatic approach to simplifying relative clauses still faces some problems. As an example, the relative order of simplified sentences needs to be specified, and, in case the order of newly generated sentences is not the same as the order in the source sentence, it is essential to ensure that a pronoun does not precede the expression it refers to. Moreover selecting the right tense when creating sentences is not a straightforward task.

Simplification Example:

Talwinder Singh, who masterminded the 1984 Kanishka crash, was killed in a fierce two-hour encounter.

becomes: *Talwinder Singh was killed in a fierce two-hour encounter. Talwinder Singh masterminded the 1984 Kanishka crash.*

2.2.2 Hand-crafted Rules

When rules are hand-crafted more control over the simplification is achieved and issues can be addressed gradually be creating more detailed rules as the simplification system is being developed. [Siddharthan, 2003] proposes a 3 stage approach to simplifying text and splitting sentences into more simple ones: analysis, transformation and regeneration. A text is analyzed and tagged with morphological information and syntactical functions, then sentences that can be separated into several shorter ones are simplified. The task of the regeneration stage is mainly to find the right sentence order based on constraints created during the analysis and simplification process.

The following example shows the input and resulting output of the simplification

Mr. Anthony, who runs an employment agency, decries program trading, but he isn't sure it should be strictly regulated.

Mr. Anthony runs an employment agency. Mr. Anthony decries program trading. He isn't sure it should be strictly regulated.

This example shows the importance of sentence ordering. In the case the first two sentences would be incorrectly ordered and hence switched, we would make an incorrect connection between "strictly regulated" and employment agency.

One of the issues worth mentioning in this section is the selection of a connecting word. When sentences connected by certain conjunctions like *because* are simplified, newly generated sentences cannot be simply placed one after another, instead we need to preserve the relationship between them. Several experiments were conducted and showed that simple cue words like *so* and *but*, and usage of punctuation, result in faster reading times. Therefore the new simplified text was connected by only simple cue words, *so* used for the result relation, *but* for the concession relation, etc.

Syntactic transformations can change the grammatical function of noun phrases and alter the order in which they are introduced into the discourse. To preserve anaphoric structure after simplification, pronouns are checked to ensure they refer to the same noun phrase as before. Hence a model of references for both original and simplified text is maintained. When the reference is not the same in the original and simplified text, a problem is identified and solved by inserting the original noun phrase. Only one level of references is maintained and therefore when "him" refers to "he" no further resolving is performed to control coherence.

2.3 Changing the Sentence Structure

In this thesis we are proposing a method that does not split a sentences into multiple short ones, but rather performs syntactical transformations on the sentence level. This method is mainly targeted to improve the MT of languages with a free word order. When translating between Czech and English the translation output suffers from many problems and changing the sentence structure is one of the ways to improve the translation. The original technique of ATS makes text more simple, but little difference in the quality of translation and moreover there is no standard technique to evaluate the results.

For languages where word order is fixed or with only a low level of flexibility, the translation system can learn simple rules that e.g. "a noun in front of a verb is a subject" or "an adjective should be moved from the position before the noun to the position after the noun". In the case of a free word order these rules are not so easy to learn and ideally deep syntactic analysis

would be included in all MT systems. Therefore we can separate the task of preprocessing from the task of MT. Preprocessed text can be further used by several different MT systems, possibly based on different technologies.

Our task is to focus on preprocessing for the MT and thus, as mentioned before, the preprocessed text does not need to follow the grammar rules of the source language, it is only meant to improve the translation. To obtain the best possible results from statistical MT systems, text simplification can be also included, but the system has to be trained on already simplified data.

One more advantage of performing the simplification within a sentence is evaluation. It gives us possibility to evaluate the resulting text via standard metrics like BLEU and NIST using the PCEDT evaluation sets. No automatic evaluation methods are available currently for translations where the number of sentences in source and output text does not match.

In the remainder of this section we are going to point out some issues that can be addressed by means of automatic text simplification (ATS) for translations between Czech and English. However when ATS is used for MT preprocessing it is important to identify syntactic and semantic features that cause the highest number of difficulties to a given MT system and focus on these.

Further on in the text we are going to use some of the standard terms in natural language processing (NLP). If you are not familiar with different levels of annotation, morphological or syntactic analysis, please refer for example to the documentation of [Hajič et al., 2001].

2.3.1 Word Reordering

Although the word order of most Czech sentences is similar to the word order in English, it does not always have to be true. Czech is a free word order language and therefore MT systems should not rely on the word order, however, conflictingly this is impossible to ask from MT systems translating on word for word basis. In the following paragraphs we are going to focus on simplifications that change the order of words in the sentence.

The word order of most English sentences is (with some exceptions of course):

Subject	Verb(s)	Indirect Object	Direct Object	Place	Time
I	will tell	you	the story	at school	tomorrow.

Subject Position

One of the basic English grammar rules defines the position of a subject in the sentence. The subject must always precede a verb it belongs to, with the exception of interrogative and imperative sentences. In Czech texts a subject is often identified by morphological case, not necessarily by the position in the sentence, therefore when text is translated on a word for word basis the resulting sentence might not follow the English rules.

Reordering Czech sentences so that subject precedes the verb can significantly improve understandability of the translated text. The following example shows how the translation can be improved.

Vyhořel celý døum. → *Burned down the whole house.*

As opposed to

Celý døum vyhořel. → *The whole house burned down.*

Objects

Objects can cause similar problems as subjects. In Czech sentences both direct object and indirect object are identified by morphological case (accusative and dative morphological case respectively). On the other hand English distinguishes between direct object and indirect object mainly by preposition or place in the sentence.

This can be solved similarly to the subject reordering case. Moreover we have not addressed yet the ambiguity of Czech word forms. Some words have the same form for dative and accusative case and therefore only semantic analysis can distinguish what should become the direct object and what should become indirect object.

The following example shows the limits of sentence interpretation and translation if the translation only follows the word order of the Czech sentence.

Podal opici lžíci. → *He gave a monkey to the spoon.*

Podal lžíci opici. → *He gave a spoon to the monkey.*

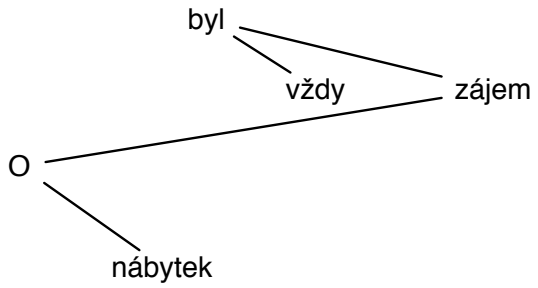


Figure 2.3: The parse trees for a sentence "O nábytek byl vždy zájem."

Place and Time

Some of the information that might need to be placed on different positions in Czech and English sentence have semantic character rather than syntactic. Expressions to identify place and time should be placed at the very beginning or at the end of an English sentence as opposed to almost any position in a Czech sentence. Nevertheless identifying these expressions is a challenging problem that we are not going to address in this thesis.

Non-projective Dependencies

Czech free word order provides a possibility to formulate sentences containing non-projective dependancies, for explanation of non-projectivity refer to [Petkevič, 2001].

When syntactic parser is allowed to include non-projective dependencies, the number of possible parse trees, it generates, becomes enormous and therefore parsers are often restricted and exclude these trees from the set of possible solutions. Therefore if we are able to identify such dependencies we gain possibility to transform or reformulate them into strictly projective parse trees. This transformation would make the sentence easier to process and the result of the translation is likely to be better.

The following sentence is an example of non-projectivity, Figure 2.3 shows the appropriate parse tree.

O nábytek byl vždy zájem.

In furniture there-was always interest.

2.3.2 Reformulating

Some language constructions are typical for one language, some for the other and when these constructions are translated without reformulating them the resulting text might be even correct, but sounding awkward or completely incorrect. Therefore where controlled language (CL) would restrict such constructions simplification tries to reformulate them.

Present Participle

Present participle is an example of verb form that is rare in some languages but more frequent in others. In Czech this form is rarely used and therefore improvements in MT can be achieved by reformulating present participle forms into use of simple tenses. The following sentence shows the difference in the translation when a present participle is used versus a simple present tense. Although the first translation is not incorrect a text containing too many such formulations does not sound natural.

Needing money to pay my rent, I forced myself to beg my parents. → *Potřebujíc...*

As opposed to:

I needed money to pay my rent. I forced myself to beg my parents. → *Potřeboval jsem...*

Czech Reflective Pronoun SE

The Czech language is characteristic for usage of reflective pronouns (*se* and *si*). However these pronouns can be used in more than one way. A reflective pronoun can be used to simply reflect back to the subject (*vzbudit se*), but also it can be used in a situation when we do not want or cannot explicitly indicate the subject (*udělá se to*). The semantics of this case is similar to using a passive form, but the Czech language prefers the use of a reflective pronoun to the use of a passive verb form.

The following is an example of using a reflective pronoun and both the correct translation using passive and the incorrect translation by omitting the pronoun.

Poplatky se platí.

Can be translated as:

Charges are paid. ***or*** *Charges pay.*

Unfortunately the use of a passive tense is not a general rule. A semantic analysis would be necessary to identify constructions where passive should be used and where omitting pronoun is sufficient. We certainly do not want sentence

Petr se postavil.

to be translated as

Peter was stood up.

Inserting Pronouns

In the Czech language, when a subject is expressed by a pronoun, it can be (and usually is) left out of the sentence. The appropriate pronoun can be deduced from the verb form. Languages with this characteristic are also called *prodrop* (pronoun - drop). Leaving out the subject poses a challenge to MT systems when text is translated to the language that requires to use a pronoun as a part of verb form. The following sentence is an example how the translation can become inappropriate if an MT system fails to insert a pronoun.

Studuji každý den.

Study every day. ***X*** *I study every day.*

Simplification can improve the translation by adding pronouns into the Czech sentence. Moreover it can be used even for English sentences. In cases where a pronoun is not repeated for the verbs in a coordination it can be additionally inserted. Although Czech text does not have to include pronouns by themselves, the verb has to be produced in a correct form and to generate this form the information from the pronoun is needed. Even though deep syntactic analysis of an English sentence would produce this information, when pronoun is added during the simplification, the analysis performed during the translation process does not need to be as thorough.

The following sentence is an example where inserting pronouns into an English sentence can improve the translation.

He said goodbye and went home. → *Rozloučil se a jít domøu.*

Verb Tenses

English use of different verb tenses is more elaborated and the verb tense usually contains more information than can be expressed in Czech. To form different tenses English uses auxiliary verbs *to be* and *to have* and their combinations, when a correct analysis of the source text is not done the resulting translation might incorrectly include also translation of these words. The following sentences are examples of incorrect translations where auxiliary verbs have not been left out.

Doctor who has been working at a hospital... → Doktor kdo má pracoval v nemocnici...

Police are currently working... → Policie jsou aktuálně pracuje...

For translations from English to Czech several different verb tenses would be translated as one and therefore the simplification process can address this issue by identifying these tenses and transforming them into more simple ones.

Sentence Wide Reformulations

Language constructions used in English are often rare in Czech and therefore an MT system usually face problems when translating them. For a perfect translation complete reformulating of the sentence is needed, on the other hand such reformulations does not need to be part of the translation process and can be done by means of text simplification. The following sentence is an example where the structure of the sentence needs to be reformulated to obtain a reasonable translation.

The man is the sixth person to be arrested. → The man is the sixth arrested person.

Indirect Speech

AN indirect speech is used in different ways in Czech and English. In English, everything is relative to the time of the speech as opposed to Czech, where we keep the same verb tenses as in the original speech. The following example shows the difference.

Řekl, že přijde. (He-said, that he-will-come.)

but: *He said he would come.*

Text simplification for MT can therefore address these differences and change the tense of the verb to be adequate for the target language. Thus for long and complicated indirect speeches this capability will not have to be included in the MT system.

2.3.3 Crutches, Helpers

When simplification is tight with the process of MT, a text can be enriched with so called *helper words*. These words are suited for an MT system and can be translated, or can be marked as not translatable if the MT system provides this possibility. Such words can include prepositions that the MT system fails to insert in the text. Other possibility is that an ambiguous preposition is replaced by its disambiguated translation.

Preposition OF

MT systems, working on word for word bases or simple matching of phrases, often fail to include the preposition *of*. The fairly easy task for a preprocessing highly improves the readability of the translated text. A Czech text can be preprocessed by inserting *of* for every two nouns in a relationship formed by genitive morphological case. The following sentence is an example of preprocessing.

$$\textit{Hrnek kafe} \rightarrow \textit{Cup coffee}$$

$$\textit{Hrnek } \textbf{\textit{OF}} \textit{ kafe} \rightarrow \textit{Cup } \textbf{\textit{OF}} \textit{ coffee}$$

This issue was also addressed by [Cuřín, 2006] for SMT systems.

Preposition BY

Adding the preposition *by* into a Czech text is not as straight forward as the preposition *of*. This preposition would be characteristic for nouns in instrumental morphological case. However not all instrumental cases should be replaced by the preposition *by*. The following examples show different cases where two nouns have the relationship formed by an instrumental case, but in each example different preposition should be used.

Říznutí nožem (by) Cut by-a-knife

Procházka lesem (through) Walk through-a-forest

Obchod čajem (with) Shop with-tea

Therefore using *by* is not so straightforward, but with deeper analysis of instrumental case usage, simplification would be able to address this issue.

Czech Preposition ZA

MT systems often fail to choose the correct translation of the Czech preposition *za* which is used either with an accusative morphological case in the sense of *for* or with an instrumental case in the sense of*behind*. Therefore sometimes awkward translations are produced. Replacing the Czech proposition by its correct translation already in the source text would make the translation task easier. The following sentence shows two different possibilities how the preposition *za* is used.

*Boj za (**for**) nazávislost.* ***X*** *Šel za (**behind**) mnou.*

Fight for independence. **X** He-went behind me.

TO to Indicate Infinitive

Another issue for MT systems can be including *to* to indicate an infinitive form of a verb. However not all Czech verbs in infinitive form should be translated into a verb accompanied by *to*. Analyzing this problem in greater detail and possibly creating a list of exceptions where *to* should not be included can increase the understandability of the generated text. The following examples show two different situations where *to* should and should not be used.

Rozhodl se to udělat. → *He decided* ***to do*** *it.*

Musí to udělat. → *He must* ***do*** *it.*

Articles

As opposed to English, Czech nouns are not accompanied by articles and therefore inserting these should be a part of the translation system. When translating from Czech to English some systems fail to include articles in the resulting text. Thus this is another domain where preprocessing can be useful. This issue was also addressed by [Cuřín, 2006].

Chapter 3

ASOFT

ASOFT (**A**utomatic**S**implification**OFT**exts) is a perl tool for preprocessing of Czech sentences. Perl was chosen as a programming language because of its platform independence. This tool was developed to support machine translation (MT) from Czech to English provided by PC Translator. ASOFT identifies some of the problematic sentence constructions and preprocess these sentences to improve the quality of the translation provided by PC Translator.

ASOFT can be used as a command line tool to process sentences that are already annotated on analytical layer. Moreover ASOFT provides a graphical user interface for sentence transformations with possibility to choose relevant transformations and review the transformed sentence.

Morphological and syntactical analysis are not part of the ASOFT tool and thus the input text must be provided in suitable format. In the case ASOFT is provided with data analyzed by automatic annotation tools and containing some annotation errors, the quality of the simplification output is in proportion to the quality of provided data.

3.1 Translation Direction

We let the PC Translator translate large amount of texts from different domains to assess the quality of the translation. We analyzed the output of the PC Translator to identify the most severe issues of the translation that can be addressed by means of automatic text simplification (ATS). As a result the ASOFT tool was developed to preprocess Czech texts that are meant to be translated into English. However the other translation direction

can be improved by preprocessing also. This section describes what are the main problems of each translation direction.

3.1.1 Czech to English

The translation from Czech to English provided by PC Translator is done mostly on word for word basis. Therefore the translation suffers mainly from the two issues: an incorrect word order and a selection of an appropriate word in target language. The problem of a wrong word ordering is mainly caused by translating on word for word basis and not making deep analysis of the source text. The system has no means to influence the resulting word order and thus it fails to identify and correctly handle parts of sentence where the word order needs to be changed. On the other hand, this task can easily be addressed by preprocessing the source text. The word order can be changed so that it follows the rules of the target language grammar.

The following sentence and its translations are examples of a situation where PC Translator fails to change the word order. The second English sentence is the appropriate translation with applied preprocessing.

Výzkumy na tomto poli provádí G. Charpak.

Researches in this field does G. Charpak. ***X*** *G. Charpak does researchs in this field.*

Moreover the text translated by PC Translator is lacking pronouns in the case the subject was not specified in the Czech text. Enriching the Czech text with pronouns would highly increase the readability of the resulting translation.

Lack of better morphological analysis also causes that some word sequences are missing correct number when disambiguation of Czech tags makes mistake in assigning morphological information.

Pro ***tyto situace*** *je lepší si pořídit fax.* $\rightarrow$ *To* ***these situation*** *is better come off fax.*

In the previous example the noun *situace* is marked as singular (genitive morphological case) and thus the resulting translation is incorrect.

Another big area where translation needs improvement is word sense disambiguation and associating the right English word with the Czech one.

Since PC Translator has built in dictionary and during the translation process it chooses the first translation that matches the phrase or the word, it is almost impossible to influence the result. Nevertheless some improvement can be achieved by replacing the word with its synonym that translates better, however a new version of the PC Translator system can change the dictionary and these replacement would lose sense.

Comparing the translation results of PC Translator for both directions, we can see that the quality of the Czech to English translation is worse, therefore preprocessing of a text can highly improve the translation results.

3.1.2 English to Czech

The results of English to Czech translation show remarkably better results, although preprocessing can be helpful in this case also. The most serious is not word order any more since translating from fixed word order to free word order languages do not bring to many issues, but another issue arises and that is selecting the right word form. Moreover we cannot forget about the issue all translation systems face, a selection of the right word sense. By closely examining the translation results we can identify several places where improvement can be achieved by means of ATS.

Although the translation results show that some analysis was performed and some of the words are generated in an appropriate form, a large number of words is not in the right form. This is often caused by a phrase inserted into the group of words that should bear same number and/or gender.

***You**, my queen, **are** fairest of all.* → ***Vy**, má královna, **jsou** nejférovější všech.*

In this case it is easy to see what separated the word group, although other examples are not so straightforward and addressing this issue would need deeper analysis, but definitely some improvement can be achieved. The next sentence is another example of wrong generation of word forms.

An eighth person has been arrested in Australia. → *Osmá **osoba byl** zadržený v Austrálii.*

English language features high ambiguity of word forms. The same word can be noun, adjective or verb, therefore in more complicated sentences PC Translator fails to choose the correct form and then the translation becomes incorrect. An example is provided bellow.

A second doctor working in Australia... → Druhý ***doktor pracování*** *v Austrálii...*

English follows a simple rule for negative sentences, use always only one negation, whereas Czech sentence can contain more. PC Translator usually fails to satisfy this rule. This problem can be solved by adding other negation into English sentence, although the problem can arise when deciding where exactly and how many negations should be added into the sentence. The following sentence is an example of an incorrect translation by PC Translator.

Nobody came. → Nikdo přišel.

Moreover English usage of verbs differs from their usage in Czech and therefore some issues arise. As mentioned in the section 2.1.1 phrasal verbs are typical for English, but they can cause lot of ambiguities. This can be addressed both by means of controlled language and by text simplification. Translation can be improved when the preprocessing system is able to identify these verbs and replace them with non phrasal synonym.

Another area where translation faces problems concerns verb tenses. Usage of tenses is different for every language and PC Translator sometimes fails to correctly translate an English verb in tense that contains auxiliary verbs. Moreover additional problems, that PC Translator fails to solve, arise when an indirect speech is used, where English uses past tense for an indirect speech in the past, the Czech sentence prefers present tense.

3.2 Simplification Process

This section gives a brief description of the simplification process performed by ASOFT. The simplification was divided into several modules, each module addresses only one issue and these modules are described in detail later in this chapter. Steps 2 through 4 are performed for each simplification module.

1. **Load a sentence** The sentence in CSTS format (Czech Sentence Tree Structure), or PML format (Prague Markup Language) is loaded. The explanation of data formats is given in [Hajič et al., 2001]. Sentences are loaded together with their parsing as provided within these file formats.

2. **Identify problems** A sentence is processed from the beginning to the end and possible problems are identified.

3. **Check the list of exceptions** For each problem where a list of exceptions is available the instance of the problem is checked against this list before making any changes to the sentence. In the case an exception is identified no further preprocessing is performed.

4. **Modify** The simplification includes two different areas:

   Either new words are inserted into the sentence. These insertions are done in such a way that all dependencies are preserved and the new word is inserted alongside with morphological lemma and tag (either real or artificial). The new word is inserted as a part of the original sentence dependency tree (e.g. OF is inserted in the tree between words to represent their relation).

   In other cases the word order of the sentence is modified.

   By keeping the sentence dependency tree valid at all times, all modules of the simplification are independent and thus can be run in any order without obtaining different results.

Using Helper Words

Some modules of the simplification insert so called *helper words* in the sentence. These helper words are words in English inserted in the Czech text which are not meant to be translated. PC Translator provides a possibility to indicate words that should not be translated, if the preferences are set correctly capitalized words are not translated. Therefore it is essential to allow this option in PC Translator preferences pane for correct integration of simplification into the translation.

3.3 Addressed Issues

In this section we are looking in greater detail at some of the issues previously mentioned that are addressed by ASOFT. We describe what are the challenges and provide examples.

Examples of the translation quality are provided in the appendix of this thesis. A sample text from the PDT was used to compare the translation quality without preprocessing and with preprocessing.

All examples in this chapter are translations from the PDT by PC Translator.

3.3.1 Subject Position

Correct identification of a subject in a sentence is one of the key requirements for understanding the meaning of the sentence and therefore it is the first issue we address. Let us focus only on declarative sentences. In an English sentence a subject is identified mostly by its position. An English subject always precedes the verb as opposed to Czech, where a subject is identified by the morphological case, i.e. the word form used.

As mentioned earlier PC Translator translates texts on word for word basis and cannot change the word order of the resulting sentence. Thus it is necessary to place the subject in front of the verb already in the source sentence to obtain the correct English translation The following sentences are examples of a situations where, in Czech, a subject does not precede the verb.

Státní příspěvek tentokrát získalo 12 knih z 9 nakladatelství. → State benefit this time got 12 books from 9 publishing house.

V tehdejším Německu nebyli demokraté. → In then Germany weren't Democrats.

Whereas more appropriate translation that can be achieved by preprocessing the source text would be:

tentokrát 12 knih z 9 nakladatelství získalo Státní příspěvek. → this time 12 books from 9 publishing house got state benefit.

demokraté nebyli V tehdejším Německu. → Democrats weren't in then Germany.

ASOFT module for changing subject position identifies a noun phrase (for the sake of simplicity we are going to refer to the part of the sentence that is dependent on a subject as a noun phrase) associated with the subject and moves it in front of the verb if it is not already there. Since a subject can be represented by several different parts of speech, correct identification of the subject of the sentence is a rather complex task. This task is usually

Percentage of sentences in PDT	31.85%
Percentage of sentences in PCEDT	24.04%

Table 3.1: Percentage of sentences transformed by subject transformations.

addressed as a part of surface syntactic analysis and thus it is not included in the ASOFT program. Therefore our preprocessing strongly relies on the correct identification of a *subject* for each *verb*. When a sentence is incorrectly analyzed, preprocessing might make the translation even worse.

The following sentence provides an example where the subject of a sentence is not a noun or pronoun as usual, but a number.

V ústavech sociální péče je asi 100 dětí. → *In institutions welfare be about 100 puppy fat.*

Table 3.1 shows the percentage of sentences where subject reordering was needed.

Pitfalls

Most of the problems that can arise when trying to correct the subject position are caused by not reliable morphological and surface syntactic analysis. In some cases noun is incorrectly marked as being in a nominative morphological case during morphological analysis and afterwards as a subject, which can lead in situations where one sentence contain several subjects for one verb, in such cases ASOFT does not perform any simplification.

A more difficult situation appears for sentences where a subject was originally missing and incorrect analysis assigned a subject function to one of the words of the sentence. These sentences are hard to identify and simplification is applied although it should not be.

The following sentences are examples from Prague Czech-English Dependency Treebank (PCEDT) where automatic analysis marked two words as subjects:

Hoare Govett jedná jako investiční bankéř konsorcia.

To je zázrak.

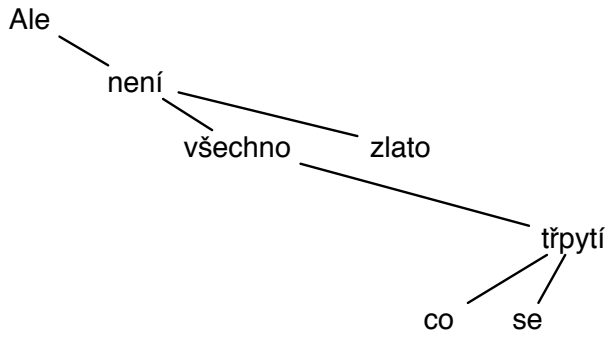


Figure 3.1: The parse trees for a sentence "Ale není všechno zlato, co se třpytí."

ASOFT also suffers from a wrong assignment of dependencies between words, since it needs to identify the whole noun phrase belonging to subject to be moved. When analysis fails to assign dependencies within the subject noun phrase, ASOFT fails to move all words that should be moved.

A special case is when the sentence contains *non-projective dependencies* and a word not dependent on the subject is in between borders of the noun phrase. Such cases are excluded from the simplification process because it is not straightforward to distinguish what words should be moved where.

The following sentence is an example of non-projectivity. The subject *všechno* is identified but not moved, because the word *zlato* causes non-projectivity. Figure 3.1 shows the parse tree for the sentence.

Ale není všechno zlato, co se třpytí. → *But isn't everything gold, what glitter.*

3.3.2 Object Position

Czech grammar rules do not specify, as strictly as in English, positions of an object and other verb modifiers in a sentence. Therefore these parts of the sentence can end up at incorrect place in the English translation of the sentence if an MT system does not perform word reordering as a part of the translation process.

Percentage of sentences in PDT	39.22%
Percentage of sentences in PCEDT	37.09%

Table 3.2: Percentage of sentences transformed by object transformations.

This problem is similar to the problem of subject reordering. In Czech texts, analytical functions of words in a sentences are defined mainly by a morphological case used as opposed to English where key identification features are position in the sentence and preposition used. Therefore ASOFT identifies verb modifiers in the Czech sentence and places them on the position that would be appropriate in an English sentence. All sentences are simplified for the translation by the following the simple rule that all modifiers are placed after the verb, ASOFT finds words and noun phrases that are placed in front of the verb and moves them after the verb.

When moving sentence parts we focus only on noun phrases and prepositional phrases to avoid changing the meaning of the sentence. By moving around all words in the sentence the meaning can be affected, some of the "dangerous" words can be e.g only, just, nearly, barely.

The following sentence provides an example where both position of the object and the subject are changed. The original sentence was:

Vědci propadají nervozitě, když se ***do jejich záhad pletou nezasvěcenci.*** → *Scientists get through nervousness, when to the their mysteries babble outsider.*

and after preprocessing it was simplified to:

Vědci propadají nervozitě, když se ***nezasvěcenci pletou do jejich záhad.*** → *Scientists get through nervousness, when outsider babble to the their mysteries .*

Table 3.2 shows the percentage of sentence where object reordering was needed.

Pitfalls

Similarly to the change of the subject position, changing the object position in a sentence is strongly dependent on correct surface syntactic analysis.

ASOFT checks the position of *objects* and *adverbials* identified by the analysis and if necessary moves them. However mere relying on analytical functions assigned is not sufficient, some pronouns (e.g. který, which) are marked as object but should not be moved to keep the fluency of the sentence.

Když na ledě vypukne rvačka, ***při které*** *teče krev.* → *When on ice will break out fight, at which flows blood.*

would be incorrectly transformed into

Když rvačka, krev teče ***při které*** *vypukne na ledě.* → *When fight, blood flows at which will break out on ice.*

Apparently this would decrease the readability of the text and is not likely to improve the translation. Thus ASOFT handles several exceptions and only nouns and noun phrases are moved as a part of this preprocessing.

In several cases an object transformation can change the sentence in a great extent from which punctuation might suffer. However we argue that punctuation is a not key feature to understand a text and thus we are not addressing this issue. The following sentence shows how the preprocessing might destroy the punctuation usability.

Zásadou je, se slušnými dlužníky jednat vždy slušně, zdůraznil Obuszák.

The-principle is, with polite debtors to-deal always fair, emphasized Obuszak.

, jednat se slušnými dlužníky vždy slušně je Zásadou, Obuszák zdůraznil.

, to-deal with polite debtors always fair is the-principle, Obuszak emphasized.

One more issue from which our preprocessing suffers is the length of phrases. For a long sentence where the object phrase is relatively long the preprocessing makes the new sentence confusing and a reader might not be able to identify what are the connections between words. The following sentence is an example that can be improved only by limiting the length of phrases to be reordered.

Úvahy o tom, že Navrátilová má šanci porazit Connorse, protože bude handicapován ztrátou jednoho servisu a polem posunutým o šedesát centimetrů do šířky na každé straně, ***považuji za nesmyslné****."*

→ *Thinking about it, that the Navratilova has chance knock down Connorse, because will handicapped waste one's set and field heterochronic about sixty centimetres abroad on each side,* ***consider nonsensical.*** *"*

Becomes after preprocessing:

považuji *Úvahy o tom, že Navrátilová má šanci porazit Connorse, protože bude handicapován ztrátou jednoho servisu a polem posunutým o šedesát centimetrů do šířky na každé straně,* ***za nesmyslné****."*

→ ***Consider*** *thinking about it, that the Navratilova has chance knock down Connorse, because will handicapped waste of one's set and field heterochronic about sixty centimetres abroad on each side, behind* ***nonsensical.*** *"*

Last but not least as a part of verb attachments we change also the position of some word forms of Czech pronouns like *vám* (you or to you). When such pronoun is placed after the verb it cannot be confused any more with the subject of the sentence. However we do not posses tools to be able distinguish in which cases vám should become *to you* and when only *you*.

The first sentence is an example where *to* is not needed and therefore simplification is successful.

Prodávající vám musí dát kromě návodu k obsluze... → Selling you have to give except instruction to attendance...

Prodávající musí dát vám kromě návodu k obsluze... → Selling have to give you except instruction to attendance...

In the second sentence simplification increases the understandability, but is not perfect

Vynaložené peníze se vám brzo vrátí. → Expended on money you soon will return.

Vynaložené peníze se brzo vrátí vám. → Expended on money soon will return you.

3.3.3 Inserting Pronouns

In the Czech language a verb form often implies what the subject of the sentence is without explicitly mentioning it or substituting it for a pronoun which would be necessary in English. PC Translator often fails to add these pronouns and therefore adding pronouns into the Czech sentence makes the resulting English sentence easier to understand. Thus ASOFT addresses this issue by resolving the form of the verb and inserting an appropriate pronoun in the Czech sentence, as a result, PC Translator, by itself, does not have to add pronouns into the translation.

The following sentence is an example of a pronoun insertion (also some word order changes are applied):

Přiznám se, že jsem z toho trošku deprimován. → *Confess, that the am from that bittock heartsick.*

já *Přiznám se, že* ***já*** *jsem trošku deprimován z toho.* → ***I*** *confess, that the* ***I*** *am bittock heartsick from that.*

ASOFT also addresses a more difficult situation that is caused by phrases with coordinations. To ensure that the resulting text will be fluent both coordinated subjects and verbs should be treated correctly. When verbs are in a coordination relation the analysis assigns the dependancy between the subject and the conjunction forming coordination, therefore only the second verb in the coordination needs an additional pronoun.

Not every verb should be assigned a pronoun and thus ASOFT carefully handles special cases, e.g. past tense. A Czech verb group in past tense *chtěl jsem udělat* is assign exactly one pronoun, although it contains three verbs. Moreover past tense brings another problem, deciding which pronoun is appropriate. ASOFT determines the form of a pronoun from an auxiliary verb *jsem* (*to be*), and the new subject is marked as being dependent on *chtěl* the verb with a meaning, to be consistent with the rules of syntactic analysis.

Vstoupil jsem do Shalomu, protože ***jsem chtěl být*** *s lidmi, kteří...* → *he came in am to the Shalomu, because* ***am wanted be*** *abreast of people that...*

já Vstoupil jsem do Shalomu, protože ***já jsem chtěl být*** *s lidmi , kteří...* → *I he came in am to the Shalomu, for* ***I am wanted be*** *abreast of people that...*

Percentage of sentences in PDT	23.26%
Percentage of sentences in PCEDT	26.24%

Table 3.3: Percentage of sentences transformed by pronouns transformations.

At the same time we can see the limits of simple inserting pronouns. Where auxiliary verb *jsem* was already used in Czech it is translated into English as *am* although it should not be there any more.

Table 3.3 shows the percentage of sentences where a pronoun was inserted when the PDT and the PCEDT was simplified.

Pitfalls

Although in general inserting additional pronouns improves the translation there are still issues to be solved. On of them is distinguishing whether to add *he, she* or *it*. Even if we are able to guess which one was dropped from the Czech sentence, we need to face the problem that usage of these pronouns is different in Czech and English (he and she is used only for people in English, but also for things in Czech). Therefore only resolving the original subject can allow us to use the correct English pronoun. Finding the proper subject is not in the scope of this thesis, but can be interesting problem to solve, therefore for the third person of singular ASOFT uses the pronoun *it*.

The following example shows wrong insertion of a pronoun.

Je naprosto jiný než Bob Johnson. → *is absolutely other than Bob Johnson.*

Anyone can see that the sentence is missing pronoun *he*, but to realize that we need to reason that Bob is a person. It is difficult issue to include such reasoning in the simplification system and thus ASOFT incorrectly inserts the default *it*.

ono Je naprosto jiný než Bob Johnson. → *it is absolutely other than Bob Johnson.*

The last issue of combining insertion of pronouns with PC Translator is a pronoun duplication. In some cases PC Translator contains pronoun as a part of the dictionary entry and in such situations inserting an additional pronoun cause duplicating it in the translation. The example is provided below.

Percentage of sentences in PDT	47.54%
Percentage of sentences in PCEDT	57.81%

Table 3.4: Percentage of sentences transformed by OF transformations.

Rozhodl jsem se, že uteču. → *I have decided, that the run off.*

já Rozhodl jsem se, že já uteču. → *I I have decided, that the I run off.*

3.3.4 Missing Preposition OF

Even short look at the text translated by PC Translator reveals that it is missing most of the prepositions *of* that should be there. Although this issue is easy to address it can highly improve the readability of the translated text. Therefore ASOFT addresses this problem by adding capitalized *OF* into the Czech sentence for every two words that are connected by genitive morphological case. Capitalizing the preposition is a clue for PC Translator that this preposition is not supposed to be translated.

Although enriching the Czech text by adding English prepositions is not a universal way to improve the translation, if the translation system fails to insert these prepositions it can notably improve the translation quality. Moreover we assume that the preprocessed text will be handled by the translation system without other human interaction and thus creating text, that does not follow Czech grammar rules any more, is not an issue.

The following short sentence is an example of the translation improvement:

Jaká je funkce spánku? → *What is function sleep?*

Jaká je funkce ***OF*** *spánku?* → *What is function* ***OF*** *sleep?*

English has two ways to express that something belongs to something: either by using preposition *of*, or by appending *'s* to the end of the word. The approach we are taking focuses only on adding *of* although a typical English text contains both, more research can be done to distinguish these two situations and suggest preprocessing method focusing on this issue in greater detail.

The percentage of preprocessed sentences from training data sets is given in Table 3.4.

Pitfalls

PC Translator translates the text mostly on word for word basis, with an exception that it tries to match phrases from dictionary first. Therefore when the translation dictionary already contains the phrase *snížení daní* changing this phrase by inserting an additional *of* makes it impossible to match the dictionary phrase. The example of correct translation with a dictionary match and incorrect with preprocessing follows.

snížení daní →tax abatement

snížení OF daní →decrease OF given

This issue can be covered by creating list of exceptions for dictionary phrases connected by genitive morphological case. However for addressing this issue by means automatic simplification detailed knowledge of PC Translator phrase-matching system would be necessary and therefore ASOFT does not contain such list of exceptions.

On the other hand even more simple list of exceptions can still improve the quality of the translation. Several dictionary entries are single word matches that contain *of* at the beginning or end of their english equivalent (e.g. *a bit of* for the Czech word *kus*). ASOFT checks for such expressions to avoid double appearance of *of* in the translated sentence. The following sentence is an example of double appearance.

Kvalita tisku každého uvedeného způsobu je... → Quality printing of each of mentioned way is...

*Kvalita **OF** tisku **OF** každého uvedeného způsobu je... → Quality **of** printing **of of** each of mentioned way is...*

3.3.5 Add TO for Infinitive

Last issue of PC Translator addressed by ASOFT is usage of *to* to indicate verb infinitives. PC Translator inserts *to* into the translation only if it is part of the appropriate dictionary entry, it does not recognize infinitives. ASOFT addresses this issue by inserting capitalized *TO* in front of the verb, this word is not going to be translated, but stays before verb as it is.

The following sentence shows an example of *TO* being inserted.

Percentage of sentences in PDT	15.03%
Percentage of sentences in PCEDT	15.49%

Table 3.5: Percentage of sentences transformed by TO transformations.

Obec postavila nový vodovod a chce stavět čistírnu odpadních vod. → *Municipality built new water supply and wants build sewage works.*

Obec postavila nový vodovod a chce ***TO*** *stavět čistírnu odpadních vod.* → *Municipality built new water supply and wants* ***TO*** *build sewage works.*

Table 3.5 shows the percentage of sentences where inserting additional TO was needed.

Pitfalls

The usage of infinitive form in Czech does not exactly match the English usage of *to* for verbs, therefore we created a list of exceptions for verbs that are used without *to*. Some of the verbs where the usage is different are *may*, *can*, etc. However one of the most common examples where verb infinitives appear in Czech is future tense with auxiliary verb *být*, since English does not use *to* for future tense future forms of *být* became also part of the list of exceptions.

The following sentence is one of the examples that is covered by the list of exceptions.

Fotbal nemůže existovat, jestliže... → *Football cannot exist, if ...*

Fotbal nemůže TO existovat, jestliže... → *Football cannot TO exist, if ...*

Some of the phrases in the PC Translator dictionary already contain *to* as a part of the dictionary entry, we need to exclude these entries from our simplification process to avoid *to* appearing twice in the translation. This part of list of exceptions is similar to exceptions for the preposition *of*. The following sentence shows over-generation of *to*.

Chceme podnikat. → *Want to carry business.*

my Chceme TO podnikat. → *we want to TO carry business.*

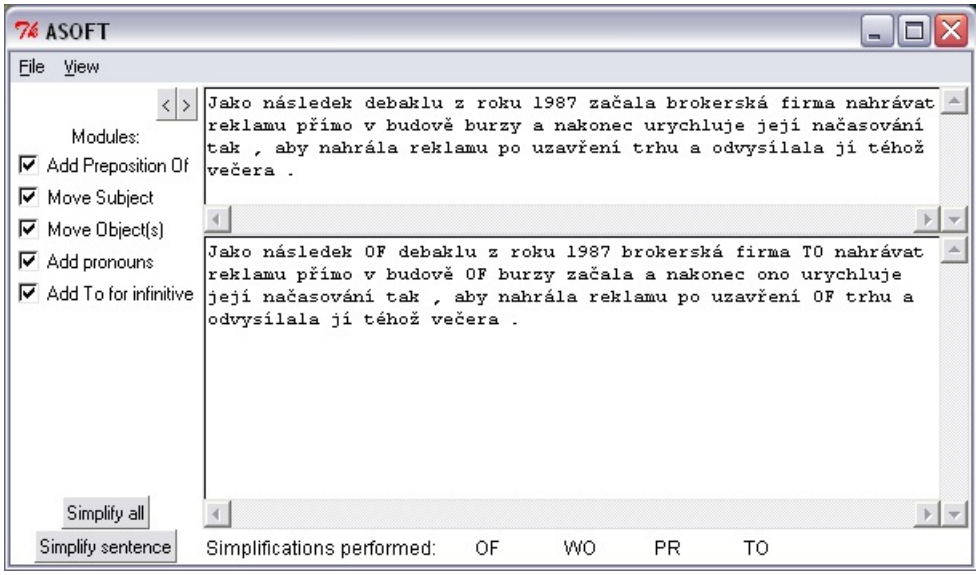


Figure 3.2: Graphical interface of ASOFT

3.4 User Guide

3.4.1 Graphical User Interface

The ASOFT provides a Graphical User Interface (GUI) for users not familiar with using the command line. The ASOFT GUI is shown on Figure 3.2. This interface provides a user with the same possibilities as the command line tool and in addition sentences can be browsed one by one with the possibility to select different simplification modules for each sentence.

ASOFT is a tool written in Perl Tk and thus it it is compatible with most of the nowadays systems. If you are using Windows you will need ActivePerl 5.8, on linux and Mac OS systems perl version 5.8 and perl Tk are necessary.

The simplification task is divided in three parts described in the following sections.

Load Data

ASOFT can open files in a CSTS (Czech Sentence Tree Structure) format and a PML (Prague Markup Language) format, moreover gunzipped variants of these files are supported.

Simplify

The simplification can be performed sentence by sentence or all sentences at once, by pressing appropriate button. The list of check boxes on the left sides provides possibility of choosing which simplification modules will be used. Thus if a user does not like simplification result he can exclude some simplification modules and re-simplify the sentence.

The status bar at the bottom of the window shows what simplifications were performed.

Before simplifying, please, make sure that the text encoding is set correctly, if not some words might not be recognized and thus some of the simplification steps may not provide accurate results.

Save Results

Text simplified by ASOFT is intended to be further used as an input to PC Translator and thus the only possible output format is a plain text format. The text is saved in a file one sentence per line.

Keyboard Shortcuts

ASOFT supports usage of keyboard shortcuts to process files efficiently. The following keyboard shortcuts are available:

Ctrl-S	Save file
Ctrl-O	Open file
Ctrl-H	Help
S	Simplify sentence
A	Simplify all
arrows	move to the next/previous sentence

3.4.2 Command Line Tool

When ASOFT is called from a command line, it processes all files in given directory and prints the simplified text on the standard output. It is important to set the correct encoding otherwise ASOFT might not correctly recognize some words and the simplification result will not be accurate. The complete set of command line options is listed below.

```
Usage: ./asoft.pl -ftpsoa [-e encoding]  -CP dir
   -f    add OF into sentences
```

```
-t    add TO into sentences
-p    add pronouns into sentences
-s    correct subject position
-o    correct position of object(s)
-a    all above

-e    set input encoding (utf8, iso-8859-2, cp1250)

-C    CSTS format, specify directory for input files
-P    PML format, specify directory for input files
```

Chapter 4

Evaluation

Text simplification is a new research topic and thus there is no commonly used metric to evaluate simplification results. Moreover there is no general rule that defines what is the correct result of simplification. Therefore we choose to evaluate the simplification by comparing the difference between the translation of the original sentence and the translation of the simplified one.

As mentioned already in the first chapter of this thesis an evaluation of machine translation (MT) brings a lot of potential problems. A source text can have several possible correct translations and even for incorrect translations it is almost impossible to rate them by mere numbers, because different translations of the same text can suffer from different problems, more or less significant.

Given all these obstacles we chose three techniques to evaluate the impact of the simplification on translation results. We use two techniques of automatic evaluation and one technique of human evaluation.

The data we used for the evaluation are parts of the PCEDT and the PDT. Table 4.1 shows the percentage of sentences that were modified by each of our simplification modules. The overall percentage of modified sentences is also listed. demonstrates how much each simplification module contributes to the simplification output (the module can be counted several times for one sentence if it performed several simplifications).

	PCEDT	PDT
Preposition of	57.81%	47.54%
Move object(s)	37.09%	39.22%
Add pronouns	26.24%	23.26%
Move subject	24.04%	31.85%
To for infinitive	15.49%	15.03%
All	83.99 %	77.49 %

Table 4.1: Percentage of reformulated sentences in data sets

	PCEDT	PDT
Preposition of	43.72 %	35.74 %
Move object(s)	22.61 %	26.90 %
Add pronouns	14.26 %	13.08 %
Move subject	11.58 %	16.24 %
To for infinitive	7.80 %	8.018 %

Table 4.2: Percentage of all simplifications for each simplification variant.

4.1 BLEU and NIST Evaluation

For automatic evaluation we chose the BLEU [Papineni et al., 2001] and NIST [Doddington, 2002] scores. The evaluation was performed on PCEDT [Čmejrek et al., 2004] data, and Tables 4.3 and 4.4 show the results obtained from the translation of the Evaluation and Development data sets. These data sets contain 5 reference translations which makes them suitable for use in automatic evaluation. The numbers in bold show where the simplification brought an improvement to the translation.

We addressed the task of simplification in a slightly different way than is usual. Instead of splitting long sentences into several short ones, we focused on grammatical simplification and reformulating. Although this gives us the possibility to use some of the available evaluation resources their suitability for the task is arguable.

	BLEU	NIST
Base line	0.2355	7.2739
Preposition of	**0.2396**	7.2074
Move object(s)	0.2350	7.2726
Add pronouns	0.2330	7.2484
Move subject	0.2301	7.2561
To for infinitive	**0.2374**	**7.2928**

Table 4.3: Evaluation on PCEDT Development set

	BLEU	NIST
Base line	0.2057	6.7499
Preposition of	0.2043	6.7230
Move object(s)	0.2051	6.7480
Add pronouns	0.2042	**6.7626**
Move subject	0.2030	6.7499
To for infinitive	**0.2066**	**6.7752**

Table 4.4: Evaluation on PCEDT Evaluation set

4.1.1 Subject and Object Reordering

The translation guidelines for creating the Czech part of the PCEDT advise making the translation of the English sentence as exact as possible. Therefore these sentences often keep their word order, and not much reordering is needed.

Moreover automatic analysis sometimes failed to correctly assign syntactical functions on which ASOFT relies. Incorrect analysis switched a subject with an object or even, in several cases, two different words in one phrase were both marked as subjects. As a result of incorrect analysis ASOFT can make incorrect assumptions about sentence parts, and fail in reordering the sentence. This situation is more common for PCEDT data since quite often no simplification is needed.

By observing the results table we can see that reordering has not improved the translation, but we predict that on a different data set significantly better results could be achieved.

4.1.2 Adding TO and OF

The evaluation tables show that adding *to*, to indicate an infinitive form, into sentences brought an improvement in the translation results. This proved the importance of analyzing the translation results of an MT system before deciding what the simplification task should address. The task of inserting *to* also demonstrated that even simple adjustments can bring more significant results than complicated sentence reordering.

Adding *to* into a sentence is not as dependent on analysis results as sentence reordering, since it only depends on correct identification of the infinitive and correct dependancy assignment for only one word. We argue that this is one of the reasons this part of simplification had a positive impact on the BLEU and NIST scores.

A similar situation arises when adding *of*. The only reason we can't see a real improvement in the translation results is that English does not use only the preposition *of*, but also *'s* to indicate a similar relationship between words. Automatic metrics fail to take this grammar synonymy into account. Even constructions where *'s* would be suitable are improved by using *of* compared to not using anything.

	E1	E2	E3	E4	E5	E6	**All**
Simplified	61	21	12	11	18	60	**183**
Original	22	9	8	9	12	30	**90**
No answer	17	0	0	0	0	10	**27**
Count	100	30	20	20	30	100	**300**

Table 4.5: Human evaluation on PDT data

4.1.3 Adding Pronouns

The evaluation of the insertion of pronouns did not bring positive results, but we argue that the results are actually better than shown in the results table. We expect that pronouns increase the readability of text in general and therefore where ASOFT adds the pronoun *it* instead of *he* or *she*, the readability increased, but the score lowered. Another problem of automatic evaluation is coordinations. In several cases a pronoun was not used for the second verb in a coordination, but ASOFT added one. We argue that when PC Translator failed to generate the correct verb form, this additional pronoun assisted in understanding the text.

4.2 Human Evaluation

For the second evaluation we used a technique of human evaluation presented in [Vilar et al., 2007]. As opposed to previously used techniques of human evaluation this one does not try to evaluate fluency and accuracy of a text, but human evaluators are given two translations of one sentence and they are asked to choose which one is better. The advantage of this technique is that no complicated guidelines are needed and therefore the consistency of the evaluation is higher.

For the human evaluation we used data from PDT that is already manually annotated. Each of our bilingual evaluators received randomly generated samples ranging in size from 20 to 100 sentences. Evaluators were asked to choose the better translation from two possible translations of the Czech sentence. The details of this evaluation are shown in Table 4.5 and a chart is presented in Figure 4.1.

Our human bilingual evaluators in 183 case out of a total of 273 answers chose the simplified sentence as a better translation than the original one.

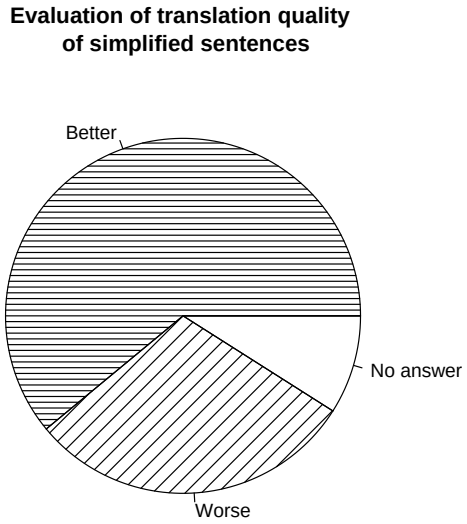


Figure 4.1: Human evaluation results

This means simplification was successful to provide better translation in 67% of sentences.

The human evaluation shows considerably better results than the automatic evaluation. There are several reasons for this. Firstly, in the analysis PDT data were manually annotated and therefore ASOFT cannot make incorrect assumption about the sentence and hence the transformations are more exact.

The second reason is that the chosen technique of human evaluation does not take into account the length of the sentence and thus gives an advantage to short sentences. When a short sentence contains one easy simplification step that improved the translation, it has the same weight as a long and complicated sentence with several possible simplifications. On the other hand it is important that short sentences are being improved since the MT of long sentences does not give impressive results anyway.

The last reason is the evaluation of lexical and grammatical synonymy. Human evaluators are better at taking this synonymy into account and therefore they were able to appreciate changes that are not included in any reference translation and thus not noticed by automatic metrics.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis we described the aims and techniques of Controlled Language (CL) and Automatic Text Simplification (ATS), and explained in detail the difference between the two to avoid confusion.

Afterwards we focused on the problem of simplification used as preprocessing for translation between Czech and English. We took a slightly different approach than other state-of-the-art systems. Rather than the usual approach of splitting long sentences into multiple short ones, we decided to address the problem by sentence reformulating, where grammar constructions can be difficult to translate. We highlighted several issues that can be addressed by means of ATS and chose the most critical ones. A tool ASOFT was developed to address these issues.

At the end of this thesis we evaluated the simplification as a part of a machine translation (MT) task. To make our evaluations as objective as possible we chose one technique of human evaluation and two techniques of automatic evaluation. Reformulating a sentence rather than splitting it gave us possibility to use standard evaluation techniques, since there is no available corpus for evaluating the translation of split sentences.

Our simplification approach is tightly connected with the machine translation system PC Translator, but we argue that it can more generally improve the output of both RBMT and SMT systems. By determining what are the issues of a given translation system, and focusing on a chosen set of them, simplification can bring a considerable improvement to the translation.

It is worth emphasizing that simplification is a rather new technique

and it is not known yet what its limits are, nor the full range of domains it could be useful in. Currently research focuses on simplification for MT preprocessing, preprocessing of texts for non-native speakers and people with reading disorders, and also for summarization.

5.2 Future Work

Being a new research topic simplification can still expand to many more new areas. As one of the areas, the domain of MT offers several possible ways that simplification can be further extended. We addressed some of the possibilities but many more issues can be chosen and addressed for different MT systems.

Instead of creating new rules one by one, new *grammar* can be introduced, and if this grammar is well designed it will make adding new rules much easier, guaranteeing that new added rules will not interfere with existing ones.

One topic tightly connected with creating a grammar is *automatic generation* of simplification rules. This task requires access to simplification corpora, where original sentences are aligned with simplified ones. The task of automatic generation of rules was addressed by [Chandrasekar and Srinivas, 1997]. An interesting question is whether it is possible to apply some of these approaches to the task of word reordering and inserting additional information in a text.

Another area that was not in the scope of this thesis, but is definitely challenging and interesting is *the other direction of the translation*, from English to Czech. Some of the issues simplification can address in this domain were already mentioned, but more can be identified and implementation could show its usefulness.

We mainly focused on simplifying input for PC Translator, but we argue that improvements can be achieved for other systems. Using these techniques for *SMT systems* can bring interesting results. When an SMT system is trained on simplified text its task is easier and thus the translation results can be improved greatly.

Another shortcoming of the ASOFT system is its dependance on the results of *syntactic analysis*. The future work in this area should try to overcome this limitation and create a system that will identify all the problematic constructions by itself or be able to recognize errors in the analysis and handle them in the appropriate way.

Bibliography

A. Bernth. Easyenglishanalyzer: Taking controlled language from sentence to discourse level. In *Proceedings of CLAW 2006, 5th International Workshop on Controlled Language Applications at AMTA 2006*, Cambridge, Massachusetts, USA, 2006.

J. Carbonell, S. Klein, D. Miller, M. Steinbaum, T. Grassiany, and J. Frei. Context-based machine translation. In *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation"*, pages 19–28. Association for Machine Translation in the Americas, August 2006. URL `http://www.mt-archive.info/AMTA-2006-Carbonell.pdf`.

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

V. Cintl. Srovnání současných komerčně či volně dostupných nástrojů počítačové podpory překladu. Master's thesis, Vysoká škola ekonomická, Prague, Czech Republic, 2000.

M. Čmejrek, J. Cuřín, J. Havelka, J. Hajič, and V. Kuboň. Prague Czech-English Dependency Treebank. Syntactically Annotated Resources for Machine Translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, volume V, pages 1597–1600, Lisboa, 2004. European Language Resources Association.

J. Cuřín. *Statistical Methods in Czech-English Machine Translation*. PhD thesis, Charles University, Institute of Formal and Applied Linguistics, Prague, Czech Republic, 2006.

G. Doddington. Automatic evaluation of machine translation quality using

n-gram co-occurrence statistics. In *Proceedings of the HLT-02*, San Diego, California, USA, 2002.

J. Hajič, V. Kuboň, and J. Hric. Česílko - an MT system for closely related languages. In *ACL2000, Tutorial Abstracts and Demonstration Notes*, pages 7—8, Hong Kong, 2000, 2000. ACL, ISBN 1-55860-730-7.

J. Hajič, E. Hajičová, P. Pajas, J. Panevová, P. Sgall, and B. V. Hladká. Prague Dependency Treebank 1.0 (Final Production Label), 2001.

W. J. Hutchins. *Machine translation: past, present, future.* John Wiley & Sons, Inc., New York, NY, USA, 1986. ISBN 0-470-20313-7.

P.-B. A. . H. E. King M. Femti: creating and using a framework for mt evaluation. In *Proceedings of Machine Translation Summit IX*, New Orleans, Louisiana, USA,, 2003. URL `www.issco.unige.ch/femti`.

P. Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *AMTA*, pages 115–124, 2004.

A. Kulesza and S. Shieber. A learning approach to improving sentence-level mt evaluation. In *Proceedings of the 20th Meeting of the North American Association for Computational Linguistics, NAACL-04*, Boston, USA, 2004. URL `citeseer.ist.psu.edu/kulesza04learning.html`.

T. Mitamura. Controlled language for multilingual machine translation. In *Proceedings of MT Summit VII, Asian-Pacific Association for Machine Translation (AAMT)*, Tokyo, Japan, 1999.

MUEGGE. http://www.muegge.cc/controlled-language.htm.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association of Computational Linguistics, ACL-2002*, Philadelphia, USA, 2001. URL `citeseer.ist.psu.edu/papineni02bleu.html`.

V. Petkevič. *Neprojektivní konstrukce v češtině z hlediska automatické morfologické disambiguace českých textů.(Non-projective constructions in Czech from the viewpoint of automatic morphological disambiguation of Czech texts).* Masarykova univerzita, Brno, Czech Republic, 2001.

A. B. Philipp Koehn, Hieu Hoang. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P07/P07-1002`.

M. Sammer, K. Reiter, S. Soderland, K. Kirchhoff, and O. Etzioni. Ambiguity reduction for machine translation: Human-computer collaboration. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas, AMTA-2006*, Cambridge, Massachusetts, USA, 2006.

A. Siddharthan. *Syntactic Simplification and Text Cohesion*. PhD thesis, University of Cambridge, 2003.

J. P. Turian, L. Shen, and I. D. Melamed. Evaluation of machine translation and its evaluation. In *Machine Translation Summit IX*. International Association for Machine Translation, Sept. 2003. URL `citeseer.ist.psu.edu/turian03evaluation.html`.

D. Vilar, G. Leusch, H. Ney, and R. E. Banchs. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-0713`.

Appendix A

Abbreviations

MT	machine translation
SMT	statistical machine translation
CBMT	context based machine translation
RBMT	rule base machine translation
CL	controlled language
ATS	automatic text simplification
POS	part of speech
NLP	natural language processing
PCEDT	Prague Czech English Dependency Treebank
PDT	Prague Dependency Treebank
CSTS	Czech Sentence Tree Structure
PML	Prague Markup Language

Appendix B

Sample Translations

Source: Mario Basler (26), záložník Werderu Brémy, bude zřejmě dalším Němcem, který zkusí štěstí v italské lize.
Velký zájem o něj projevil Juventus Turín a prý je připraven zaplatit Brémám sedm miliónů marek.
Přinejmenším měsíc se bude muset FC Barcelona obejít bez Rumuna Gheorgha Hagiho (33), fotbalista utrpěl při tréninku svalové zranění na levé noze.

Result: Mario Basler (26), reservist Werderu Brémy, will evidently to other Germans that the will try luck in Italian league.
big interest in he disply manifest Juventus Turín and is said is ready pay Bremen seven million mark.
at least month will have to FC Barcelona go round without Rumuna Gheorgha Hagiho (33), footballer suffered at training muscular injury on the left leg.

Simplified: Mario Basler (26), reservist OF Werderu Brémy, will evidently to other Germans that the will try luck in Italian league.
Juventus Turín disply manifest big interest in he and is said it is ready TO pay Bremen seven million OF mark.
FC Barcelona will have to at least month go round without Rumuna Gheorgha Hagiho (33), footballer suffered muscular injury on the left leg at training.

Source: Vybral jsem si Paříž
Romana Kameše jsem našla v malé tiskárně v Belleville, kde kontroloval soutisk barev na plakátu Jiřího Koláře.
Poblíž tiskárny bydlí spisovatel Lubomír Martínek, a tak jsme šli ve třech na pivo.
Nejdříve jsme probrali pařížské drby o Lidových a Literárních novinách.
Ty nám ale nevydržely moc dlouhou dobu.
Pane Kameši, ve výtvarném okruhu Revue K patříte k nejznámějším jménům.
Ve Francii ani jiné zemi však příliš nevystavujete a neprodáváte.
Jaká byla vaše cesta osamělého běžce od začátku emigrace až k Revue K?

Result: chose am Paris
Romana Kamese am find in small printer in Belleville, where checked colour register on poster George wheelwright.
nearby printer lives writer Lubomir Martínek, so we're went in three on beer.
first we're go through Parisian gossip about people's and literary paper.
you to us but untimbered much for a long time.
Mr. Kamesi, in creative circle review to be part of best known names.
in France nor other provincial however too no expose and no sell.
what was your way lonely runner from the start emigration as far as review to?

Simplified: chosen I am Paris
I am finded Romana Kamese in small printer in Belleville, where it checked register OF colours on poster OF Jiriho wheelwright.
writer Lubomir Martínek lives nearby printer, so we're went in three on beer.
first we're go through Parisian gossip about people's and literary paper.
you but untimbered to us much for a long time.
Mr. Kamesi, you belong to in creative circle OF Revue to to best known names.
in France nor other provincial however too you no expose and you no sell.
what was your way OF lonely runner from the start emigration as far as review to?

Source: Dvě učebny a kancelář nakonec gymnáziu pronajalo Střední odborné učiliště stavební.
Trestní oznámení na čtyři vojáky základní služby podal k Vojenské prokuratuře v Plzni velitel vojenského útvaru Radošov u Karlových Varů plukovník Bedřich Švehla.
Minimálně v jednom případě prý šikanovali mladé vojáky.
Dva nováčci základní služby uvedli, že v polovině srpna je v noci, když byl dozorčí posádky pryč, starší vojáci postupně pozvali k sobě na pokoj, kde je nutili říkat, kolik jim zbývá dnů do civilu, a pak je mlátili řemenem přes zadek.
Oba zbité vojáky převezli druhý den na čtrnáctidenní pozorování do karlovarské vojenské nemocnice.

Result: two schoolroom and office in the end secondary school hired middle training college building.
complaint on four soldiers basic services handed up to military prosecution in Plzni master military formation Radosov near Karlovych Varů colonel Bedrich Svehla.
minimally on one occasion is said vexed cub soldiers.
two newcomers basic services introduced, that the midway August is at night, when was supervisory crew away, older soldiers step by step invited on on room, where be compelled say, how much them there is left days to the mufti, and then is threshed strap over back.
both hangdog soldiers ferry second day on semi - monthly sighting to the Carlsbad military hospital.

Simplified: in the end middle training college building hired two schoolroom and office secondary school.
master OF military formation Radosov near Karlovych Varů colonel Bedrich Svehla handed up complaint on four soldiers OF basic services to military prosecution in Plzni.
is said vexed cub soldiers minimally on one occasion.
two newcomers OF basic services introduced, that the, when supervisory crew he was gone, older soldiers step by step invited is at night midway OF August on on room, where they urged is TO say, how much there is left them days to the mufti, and then they threshed is strap over back. they ferry both hangdog soldiers second day on semi - monthly sighting to the Carlsbad military hospital.

Source: Volič může žádat referendum - pch!
milión podpisů pro nás nic neznamená;
může se bránit výstavbě Gabčíkova a dalším ekologickým katastrofám - takové lapálie jeho zástupce nezajímají.
Je - li tomu tak, pak k příštím volbám nemusím chodit, protože ve Federálním shromáždění nebude nikdo, kdo - maje ode mne k tomu pověření - by tam zastupoval mé zájmy.
Jsem přesvědčena, že urážlivá slova o hlásné troubě by si žádný poslanec v "zavedených" demokraciích nemohl dovolit.
Nejsmutnější na všem je, že se v džungli stranických, partikulárních a mafiánských zájmů zcela zapomnělo na to, o co vlastně jde: o vytvoření demokratické společnosti svobodných lidí.

Result: voter is able to request referendum - pch!
million signatures for us nothing no entails;
is able to defend build - up Gabcikova and next ecological disasters - such trifle his vice desinterest.
is - if that so, then to next election I need not go, because in federal gathering won't anybody, who - maje from me hereto commission - would there deputized my politics.
am satisfied, that the offensive words about warning system fathead would no deputy in "established" democracies couldn't allow.
saddest on everything is, that the in the jungle party, cellular and mafia focus quite forgot at it, about what as a matter of fact walks: about creation democratic companies free people.

Simplified: voter is able to request referendum - pch!
million OF signatures means nothing for us;
it is able to defend build - up OF Gabcikova and next ecological disasters - such trifle desinterest his vice.
is that - if so, then I I need not go to next election, because anybody, who - maje hereto commission from me - would there deputized my politics won't in federal gathering.
I am satisfied, he should no deputy couldn't in "established" democracies afford offensive words about warning system fathead.
saddest on everything it is, that the quite it forgot in the jungle OF party, cellular and mafia focus at it, about what as a matter of fact it walks: about creation OF democratic companies OF free people.

Appendix C

Content of Attached CD ROM

This thesis is accompanied by the CD ROM containing source code of the ASOFT implementation and samples of data for simplification. The CD ROM is organized as follows:

`/README.TXT`
: Brief description of the content of the CD ROM.

`/asoft/`
: Source codes for the ASOFT tool.

`/doc/`
: Electronic version of this thesis.

`/sample_data/`
: Samples of PCEDT and PDT data ready for simplification.