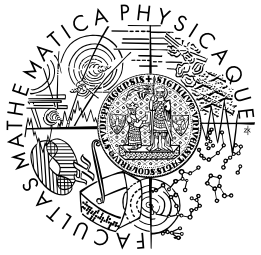

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Karolína Burešová

Text Simplification in Czech

Institute of Formal and Applied Linguistics

Supervisor of the master thesis:	RNDr. Pavel Pecina, Ph.D.
Study programme:	Computer Science
Study branch:	Mathematical Linguistics

Prague 2017

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ..... signature of the author

Title: Text Simplification in Czech

Author: Karolína Burešová

Institute: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Pavel Pecina, Ph.D., Institue of Formal and Applied Linguistics

Abstract: This thesis deals with text simplification in Czech, in particular with lexical simplification. Several strategies of complex word identification, substitution generation and substitution ranking are implemented and evaluated. Substitution generation is attempted both in a dictionary-based manner and in an embedding-based manner. Some experiments involving people are also presented, the experiments aim at gaining an insight into perceived simplicity/complexity and its factors. The experiments conducted and evaluated include sentence pair comparison and manual text simplification. Both the evaluation results of various strategies and the outcomes of experiments involving humans are described and some future work is suggested.

Keywords: text simplification, lexical simplification, readability, understandability

Dedication

There are many people I would love to thank to, actually much more than would fit here. Still, there are some who really, really deserve their place here :)

First of all, I must thank all my annotators for their time and effort. You have done a great work!

As the very next, I would like to thank my supervisor. Thank you for providing me with all the support and information. Thank you for encouraging me to try various ideas as well as for holding me back when I was tempted to plan way too much work. I could have definitely tried more myself but I would not ask more of you.

Mami, díky za všechny rady a porady, kterých se za to čtvrt století nastřádala pořádná hromádka. Jak to bývá, s věkem přibývá třecích ploch, ale jsem ráda, že se nám je stále daří překonávat, a věřím, že nikdy nepřestanu být vděčná, že mám za mámu právě Tebe. (A jakkoliv se strojový překlad zlepšuje, různá poděkování Ti dál budu raději psát česky.)

Daddy, the song which covered a photo-show I had made for you contains, among others, the following verses: God gave me your eyes / but it was you who taught me / how to see. It would be so unfair to many people to claim this the absolute truth but your willingness to try new things with me and to introduce me to all the stuff you had already known has really been a great aid in learning to see.

Marťas, thank you for all the games and adventures we have enjoyed together, be it pretending to be trainers in the Pokémon world back in the time when we had not even heard of role-playing, exploring unknown cities or just trying to solve some problems. I will put it in this way: there were far more days when I wished to kill you, preferrably in a very creative and cruel way, at least once than days when I did not wish to at all, but you are the most important reason for me to wish for my future children to have a sibling of similar age.

Martin, I still remember that before your family moved out and I moved in, there had been a magnet on the fridge which had read "I had a dream and YOU came true". Martin, I could not say it better. Thank you for your support, for our jokes, for your smiles. And since you are so undreamy sometimes, I hope there is a lower risk of waking up one day :)

Jirka, Oggy, Jethro, Medvěd and all KSP guys, thank you for being there. Thank you for your enthusing about Matfyz when I was back at highschool, I probably would not even attempt to study there if it were not for you. Thank you for all your support, trust and consultations – as I must not forget consultations, all the times when we, more or less successfully, sought knowledge together. It was really pushing.

Last but not least, thanks to everyone who makes up Salvátor parish, especially to all my fellow no longer catechumens and confirmees. Thank you for sharing all our questions, answers and experience, thank you also just for the time we had together. The path we had taken has been an indivisible part of my study.

Contents

1. Introduction

This thesis researches text simplification, focusing on Czech, a Slavic language, offering various approaches to some simplification subproblems (albeit the simplification problem is solved neither thoroughly nor as a whole), thus shedding some light on a problem of non-negligible importance for several target groups of notable sizes.

In other words: This thesis deals with text simplification. It works with Czech (a Slavic language). It doesn't solve simplification completely but it tries to solve some of simplification tasks. Text simplification can be important for many different people.

Text simplification, as indicated above, means altering (or completely rephrasing) the text so that it is easier to read and understand; its meaning should be preserved. This definition is far from scientifically exact and unfortunatelly, I will not be able to give a more exact one, but I will describe the task in more detail later in the introduction.

I will also give some examples of what makes the task difficult, suggesting some reasons why the problem remains unsolved despite the growing research. Before speaking about the task and its difficulty, I would like to stress out the role of text simplification.

Even though we do not realize it sometimes, a lot of people work with a kind of simplification and simplified texts on a daily basis. Rules and laws are often reformulated for common citizens, simplified medical information is sometimes provided to patients (and if it is not, patients often tend to seek it in some other way), parents intuitively use a simpler language when communicating with little children.

Some groups of people generally benefit from simplified texts, regardless of the topic. A very common example of such group are second language learners (L2 learners), either learning a language outside of any of the countries where the language is spoken or learning the language of the country they are currently in. Another common example are people suffering form aphasia or dyslexia. I would also like to mention hearing-impaired people, who could be treated as a special case of L2 learners.

Some people, on the other hand, would mostly benefit from simplification of domain-specific texts. Those people generally know the language well, can understand various syntactic structures, but do not have the expert knowledge needed to understand expert texts. A common example, as indicated before, are patients seeking some medical information. It is also worth noting that the understandability of information on various diseases is often studied and reported. However, domain-specific simplification (or at least its need) is not limited to medical information. It is sought regarding regulations (though summarization and explanation might be preferred in this case), it is often sought when grasping the basics knowledge about a new topic.

The demands for simplified texts and efforts to provide them are demonstrated by some existing projects. The best known is probably Simple English Wikipedia[1] which has also served as data source for some studies. As of 10 May 2017, it

[1] https://simple.wikipedia.org/wiki/

reports 124,672 content pages. Those pages are created manually by volunteers. Another project is News in Levels[2] which provides news at three different reading levels. The project is aimed at L2 learners. The authors publish two short articles each working day. There are also associated projects like Videos in Levels[3] or Jokes in Levels[4] to further assist the learners.

While some texts can be expected to be read by many people with different abilities to understand them, and thus written at several levels right away[5], it is reasonable to only simplify the others on demand. Those others might be for example a story which catches an interest of an L2 learner or a medical information on an uncommon illness.

Simplifying different texts for people with different language knowledge is a difficult and time-consuming problem. For this reason, it is often not done and people are left with original texts. Solving the problem of automated text simplification would allow to provide people with texts which would be both on topics important or interesting for those people and at their level of language knowledge.

The task

As stated before, text simplification is the task of altering original text in a way that preserves the meaning and reduces the complexity of the text.

Some further parts of the definition are perceived as implicitly present in it, even though they are not mentioned explicitly in the definition.

First, the original text is assumed to be grammatic, even though the grammatical structures used in the text may be strange, uncommon, difficult to parse. This assumption is generally correct, as errors that occur due to rethinking the sentence while writing are mostly corrected when the author reads his/her text once again or when the text is proof-read.

Sometimes the original text is also assumed to be free of spelling errors. This is often not true as spelling errors sometimes occur in real-life texts despite re-reading and proof-reading. However, they can be solved before performing the simplification itself. There are no reasons to believe that solving them during the simplification process would bring better results.

Second, the resulting text is also expected to be grammatically correct. While there is no reason to introduce a spelling error, ungrammatical structures could be easily created.

Third, it is assumed that reducing complexity (increasing simplicity) helps the text to better serve its purpose. In general, the purpose of a text is to communicate some information, though the purpose could also be for example to help with second language learning.

There are several problematic points in the definition. I would like to point them out and explain their intricacies in more detail.

[2]https://www.newsinlevels.com/

[3]http://www.videosinlevels.com/

[4]http://www.jokesinlevels.com/

[5]Making them only easy is likely to result in some information loss or decreased readability for people more familiar with the topic.

The first point is meaning preservation. While it is easy to understand why meaning should be preserved when simplifying text, it is difficult to quantify the change in meaning and thus to set an acceptable treshold to this change.

Some linguists might argue that no matter what change one does to the original text, the information contained in the text will not be the same. Extremely uncommon words or hard-to-parse structures could be used for a reson (but they could be in the text unintentionally as well), the author might use some words to express his/her knowledge, conform to a social role, provoke a specific sentiment, ...

Specifying a piece of information in more detail could be just giving unnecessary details as well as mentioning something important. Three sentences 'Tom did well', 'James did quite well' and 'Andrew did very well' could probably all be simplified to '(Name) did well' and we would perceive it as preserving the meaning. However, we would be in trouble if we wanted to deduce that Andrew did much better than James then.

Sometimes the information is definitely less specific but possibly still sufficient. In Czech, the word "zábradlí" (*railing*) is far more frequent than the word "balustráda" (*balustrade*), at least in general corpora. While not every railing is a balustrade, substituting 'balustrade' with 'railing' could be treated as meaning preserving while making the text simpler, even though part of the information is lost.

The second, and probably much more important point is simplicity itself. There is no exact definition of text simplicity, and even though some formulas to measure readability and understandability do exist (and I describe them in more detail later in the thesis), the reduction in complexity is subjective and dependent on both the target reader and the purpose of the text in question.

Shorter words are generally considered simpler, the same is true for more common words (it is worth noting that those things seem to correlate, more frequent words are often also shorter).

A Czech speaker, especially one suffering from dyslexia or some other disability, would probably appreciate substituting the word 'neutuchající' (*unflagging*) or the word "nehynoucí" (*imperishable*) with the word "věčný" (*eternal, unending, unfailing*), which is both shorter and more frequent.

A Czech learner whose first language is different, on the other hand, would maybe prefer the word "internacionální" to the word "mezinárodní" (*international*) or the word "koordináta" to "souřadnice" (*coordinate*). The possibly preferred words are both longer than their counterparts and very rare in Czech texts, but they resemble their English translations as well as their equivalents in many different languages.

Substituting a word with a simpler one (in any sense) is also problematic when it comes to technical terms. Speaking about 'procedure' or 'steps' instead of 'algorithm' could make the text easier to understand for someone who has not heard of algorithms before, but it could confuse anyone who has some knowledge of computer science. Such a substitution also makes it much more difficult (virtually impossible) to further search for the term, either in books, via Internet search engines, ..., in case the reader wishes to learn more.

The implicit parts of the definition should not be ignored, and unfortunately, some problems are associated with them too.

While grammaticality and general correctness of the output is definitively desirable in general, it could contradict simplicity under special circumstances. I will give an example using Czech. There is a special possesive pronoun in Czech, "svůj", which expresses belonging to the subject, no matter its person.

While the sentence "Dívám se na svůj test" means *I am looking at my test*, the sentence "Díváš se na svůj test" means *You are looking at your test*. Using pronouns "můj" (*my*) or "tvůj" (*your*) in such sentences would be incorrect, but possibly easier to understand.

The purpose of the text also should not be ignored. If the text is used as a learning aid, if reading it is supposed to help with second language acquisition, then grammaticity plays a great role and the text should not be made too simple (so that there is still something to practice and/or learn). Especially for the "svůj" example, the correct pronoun should be used.

If, on the other hand, the only purpose is to communicate some information, understandability might be preferred, even to the detriment of grammaticity.

The priorities would be yet different if the purpose of the text were communicating some information, but the user had the text machine-translated after the simplification itself. (It is expectable that this will not be needed at all once machine translation is pretty good for any language pair but until that, improving the translation by simplifying source text could possibly be attempted.)

This all should demonstrate that the task is complex and that there are many issues with both simplification itself and with its definiton and evaluation. As usual, such a complex task is not solved as a whole but it is divided into many subproblems which are solved separately (I do the same in this thesis). However, those subproblems are generally still complex, complicated and hard to solve.

Task restriction in the thesis

Since the task is very complex, I limit it to a less complex one. This limitation of course affects the quality of the output, a simplification as complex as the one at the very beginning of the introduction is not even attempted, but it allows to focus on the reduced task (and perhaps get better results on it).

Most of all, I do not consider simplification that is aware of large context; the simplification is performed at sentence level only. The perceived complexity of a sentence seems to depend on the context (and the effect cannot be generalized, the surrounding sentences can cause the sentence in question to seem either more or less complex than it would seem in isolation). The context could also alter what a *good* simplification is. However, when treating sentences as isolated pieces of text, it is well possible to alter them and evaluate the effect of the alteration.

I only consider lexical substitution, i. e. substituting words with other ones. More exactly, I also share some ideas for syntactic and pragmatic simplification (that is, for simplifying the sentence structure and for making the point of the text more clear), but I only implement and evaluate methods for lexical substitution. I will give more details on lexical substitution in chapter 5. Still I will say now that I do not solve substitution selection explicitly but I experiment with all the other common steps (complex word identification, substitution generation, substitution ranking).

I only experiment with simplifying Czech texts. This is partially because of my knowledge of the language, which makes it easier to think about various simplification strategies and simplifies on-the-fly evaluation (although any reported results are obtained from more annotators). To the extent of my knowledge, text simplification has not yet been studied for Czech. This decision has also slightly complicated the process as there no ready sources of complex/simplified texts.

1.1 Thesis organization

First of all, this thesis includes an introduction to the task of text simplification and its incentives, including task restrictions in this work. Following this, some basic terminology is defined so that there is less ambiguity in later descriptions. Related works are described as the last part of the introduction, in section 1.2.

Since this English text discusses Czech, a different language, some notes on Czech are given in the chapter 2. Those notes should give the reader an idea of some grammatical particularities of Czech and stress out their implications on text simplification. This chapter can be safely skipped by a skilled Czech speaker, though realizing the implications could be interesting even for such a speaker.

Chapter 3 gives an extensive overview of text simplification. It recalls the task and describes its various aspects. Various target groups are described in section 3.1 with some emphasis on their special needs and most crucial simplification subproblems for them. The concepts of readability and understandability are discussed in section 3.2, including some common formulas for readability measurement.

The next three sections describe three aspects/levels of text simplification. Lexical simplification is introduced in section 3.3, though it is described in more detail in chapter 5. Section 3.4 describes syntactic simplification, including some suggestions on how to syntactically simplify Czech texts and warnings about the tricky parts of those simplifications. And finally, section 3.5 describes the idea of pragmatic simplification.

Experiments I have conducted and evaluated with humans are described in chapter 4. Each section describes one of the three experiments conducted to gain some insight into what makes people perceive text as difficult and what they try to change to make the text simpler.

Lexical simplification, introduced in section 3.3, is described in more detail in chapter 5. Three steps are described, complex word identification (section 5.2), substitution generation (section 5.3) and substitution ranking (section 5.4). Specific methods for each of the steps are described in the corresponding sections.

Chapter 6 gives the results of evaluations of various methods described in chapter 5 on lexical simplification. The results are given for each of the individual steps (sections 6.1, 6.2, 6.3).

As a lot of programming has been part of the thesis, some notes on implementation are given in chapter 7. The main intention of this chapter is to help the reader set up all necessary prerequisities and run the scripts to see their output and the effects of any potential changes on the output, but it could also be useful just to get a better insight into how the data had been processed.

Finally, the conclusion gives an overview of the results and suggests some future work which could lead to better results.

Terminology

Since natural languages are very vague and often ambigious, discussions on them could get very topsy-turvy easily. Since this thesis discusses one language using another one, it is also necessary to give translations which might require further comments or adaptations. For this reason, I wish to systematize translations giving and define some terms which I will further use. I will describe the translations first.

All Czech snippets, be it single words or whole clauses, are given in quotes, like this: "ukázkové české spojení". English translations are given in slanted text, usually in round brackets following directly the Czech snippet: "Toto je ukázka" (*This is a demonstration*).

Sometimes a note is given in square brackets. If in slanted text, such a note is there to make the English sentence grammatical ("utíkáš" (*[You] run*), to give eliminated context, to express topic focus ("ty půjdeš" (*[It is] you [who] go*) or to specify/restrict word meaning. Otherwise, it explains cultural context or knowledge, e. g. "Jedu do Prahy" (*I am going to Prague* [Czech capital]).

A note in upright font could also be used to explain some grammatical particularities, for example to comment on number or case if needed.

If it is necessary to translate the sentence word by word to demonstrate some phenomenon or simplification strategy effect, the translation is preceded with the word *verbatim*: "dárek dostal každý" (verbatim *a present*/acusative *got everybody*/nominative).

The word *literally* preceding a translation indicates that Czech text was translated into grammatical English text but the real meaning is not preserved, usually because the text in question is a phraseme; in such cases, a more exact translation follows: "teče mi do bot" (literally *It is leaking into my shoes*, *I am in dire straits*). I also use it when translating into near-grammatical English containing a specific grammatical error to demonstrate special phenomena (particularly multiple negative).

Now I would like to specify some terms I will be using throughout the thesis.

- **Sentence** refers to a sentence as produced by MorphoDiTa,[6] which I used for tokenization. In general, a sentence begins with a capital letter and ends with a dot, it might contain commas, colons, semicolons, parentheses and other punctuation as well as several verbs. Common cases of dots which do not end a sentence are dates ("29. 1.", (*29 January*)), ordinal numbers ("1. místo", (1st *place*)) and web addresses.

- **Phrase** refers to either a part of sentence or a phraseme. When referring to a part of sentence, it usually refers to a clause simplex.

- **Simplification strategy** or only **strategy** (if the context is clear) refers to a strategy/operation used to simplify text. A simplification strategy could be for example sentence splitting or lexical substitution.

- **Simplification** refers to the process of simplifying which generally employs some simplification strategies.

[6] http://ufal.mff.cuni.cz/morphodita
More information on MorphoDiTa is given in section 7.1.

- **Complex sentence** is a sentence which is considered difficult and needs to be simplified. Simplification is performed on complex sentences.

- **Complex word** is a word which is considered difficult and needs to be replaced with something simpler. Presence of a complex word in a sentence makes such sentence a complex sentence (though a complex sentence need not contain a complex word).

- **Resulting sentence / resulting word** is the result of simplification performed on a complex sentence / complex word.

- **Form** is a word as it appers in the text. It could be for example "Větší" (*bigger*, part of a name or the first word in a sentence because of the capital), "stál" (*[he] stood*, third person singular, past tense) or "neotřesitelnou" (*ingrained*, acusative or locative, singular).

- **Lemma** is a conventional base string common to all forms of given word. The lemma is as produced by MorphoDiTa and can contain some additional information (hints on form derivations or word sense markings are common in lemmata recognized by MorphoDiTa). Lemmata for the previously given forms would be "velký", "stát-3_^(někdo/něco_stojí,_např._na_nohou)", "neotřesitelný_^(*13otřást)".

- **Stripped lemma** is a substring (not necessarily proper) of the lemma which does not contain the additional information. Information on word sense is also omitted. Most dictionaries use stripped lemmata to give meanings or synonyms. Stripped versions of the previously given lemmata would be "velký", "stát", "neotřesitelný".

- **Key** is either a string derived from the complex word used when searching for substitutions or a string derived from the substitution used to rank the substitution. Key can be either a form, a stripped lemma or a lemma.

- **Content words** are nouns, adjectives, verbs and adverbs.

1.2 Related works

As the research of text simplification is growing, there are many works which deal with various subproblems of text simplification and focus on different languages. Nice overviews of such works are given by Siddharthan (2014) and Shardlow (2014). This section is partially based on those overviews.

Shardlow (2014) reports a checking tool for writers of simplified English (Hoard et al. (1992)) to be the first work towards automated simplification. This checker was designed for the writers of Boeing aircraft manuals and should ensure compliace with ASD STE-100[7], a standard for simplified English used in aerospace manuals.

Angrosh et al. (2014) attempted to automatically extract transformation rules from a complex-simple parallel corpus, using the aligned corpus of English and

[7]http://asd-ste100.org

Simple English described in Woodsend and Lapata (2011). This work (Angrosh et al. (2014)) is interesting also because it was evaluated using non-native speakers' results in comprehension tests.

As for lexical simplification, several ideas arose in competing solutions to the 2012 SemEval task, as described by Specia et al. (2012). This task was limited to ranking pre-selected substitutions, other steps of lexical simplification were omitted. However, only one of nine participating solutions managed to score better than a baseline solution (ranking based on frequency only).

Paetzold and Specia (2015) present a simple software framework for lexical simplification. Even though the title mentions only the framework, novel strategies for some lexical simplification steps are also presented in the paper, namely using word embeddings produced by Word2Vec[8] to select substitutions and treating ranking as a binary classification task.

Some of the other works on lexical simplification include Devlin and Unthank (2006) (targeted at aphasic people), Elhadad and Sutaria (2007) (making use of corpus alignment), Yatskar et al. (2010) (learning paraphrases from Wikipedia edit history) or Deléger and Zweigenbaum (2009) (dealing with medical domain). Some works also explicitly employ word sense diambiguation, for example De Belder et al. (2010) or Biran et al. (2011). Thomas and Anderson (2012) attempted to make use of hypernymy captured in WordNet for lexical simplification. As for languages other than English, languages which have already been researched include Japanese (Inui et al. (2003)), Brazilian Portuguese (Watanabe et al. (2009)) and Swedish (Keskisärkkä (2012)).

Syntactic simplification mostly focuses on making sentences shorter, works on this task include Carroll et al. (1998) (targeted at aphasic readers), Kandula et al. (2010) (medical domain, sentence splitting after explanation generation) or Klerke and Søgaard (2013). The technical report by Siddharthan (2006) is notable because of discource maintanance. Paetzold and Specia (2013) attempted to automatically learn simplificatons as operations on trees. Several languages have been researched in terms of syntactic simplification: Dutch (Daelemans et al. (2004)), French (Seretan (2012), semi-automatic rule acquisition), Vietnamese (Hung et al. (2012), intended to improve machine-translation), Basque (Aranzabe et al. (2013)), Italian (Barlacchi and Tonelli (2013), targeted at simplifying children's stories), Korean (Chung et al. (2013), targeted at deaf people, dealing with web documents) and Spanish (Štajner et al. (2013)).

Some works, for example Chen et al. (2012) or Stymne et al. (2013) do not research text simplification because of the simplification itself but they use it as an aid for machine translation.

Specia (2010), Wubben et al. (2012), Coster and Kauchak (2011a) and others, on the other hand, treat the simplification itself as a problem of machine translation. They consider a complex language and a simplified language and attempt to translate from the complex language to the simplified language using existing machine translation systems.

Besides works on automated text simplification, I would also like to mention works which are rather on psychology or creative writing but are very useful for this task as they try to give answers to the question of simplicity measurement. Some works give an easy-to-use formula which, based on some features of the

[8] http://code.google.com/p/word2vec/

text, produces a single number. Those works include Flesch (1948), Gunning (1952), Fry (1968) and McLaughlin (1969) and will be described in more detail in subsection 3.2.2.

Some works also offer a computer science approach to classifying texts into various levels reflecting the simplicity or complexity of the text. An example of this is Petersen and Ostendorf (2009). In this work, an SVM classifier is trained to label texts from Weekly Readerwith corresponding reading level. Being able to train a classifier could be very useful in adjusting the classification for different target groups.

Smith and Taffler (1992) focused on comparing some readability and/or understandability measurement techniques. They found out that the outputs of those techniques do not correlate sometimes.

Napoles et al. (2011) deal with sentence compression[9] instead of text simplification, however some of their conclusions are interesting also for the simplification task. The paper reviews evaluation of sentece compressions, in particular it discusses some problematic aspects. Most importantly, it shows that compression rate is tightly connected with human rating of compression quality. It also shows that grammaticality of the compressed sentence is rated better when the compressed sentence is presented in isolation, in contrast to presenting both original and compressed sentence. A hypothesis could be that similar effects would be present in text simplification evaluation.

I would also like to mention two studies dealing with data sources. Having a good source is crucial for both automatic rule acuisition and machine translation training. The ParallelSEW corpus, a parallel corpus of English and Simple English Wikipedia by Coster and Kauchak (2011b) is often used, though the below mentioned studies suggest it could have serious issues.

Amancio and Specia (2014) attempted to automatically annotate the simplification strategies (called transformation operations by them; strategy annotation is its categorization using categories such as sentence splitting or drop of information) used on resulting sentences in Simple English part of ParallelSEW corpus. In the process, they manually annotated the strategies and found out that cca 40 % are paraphrases (which are difficult to automate) and more than 7 % of parallel sentences are not really parallel (and therefore do not serve the training well, actually they could rather confuse the system being trained).

The others to explore the ParallelSEW corpus were Xu et al. (2015). They manually inspected 200 randomly sampled sentences and found out that only half of the sentences are real simplifications. The authors suggest using a new corpus for learning in text simplification.

[9]Sentence compression is the task of reducing sentence length, potentially at the cost of losing some information.

2. Notes on Czech

This thesis deals with texts written in Czech so I find it useful to make some notes on Czech, especially on Czech morphology. The morphology is rich and there are many things to keep in mind when altering Czech sentences.

Some of those things can be easily treated with a morphological analyzer and a morphological generator, some can be treated with a parser, some are even a bit more complicated.

Inflection

Nouns, adjectives, pronouns and numerals are inflected in Czech. The inflection expresses

- case, there are seven cases: nominative, genitive, dative, acusative, vocative, locative, instrumental; and
- number: either singular or plural.

All Czech inflective words have one of four grammatical genders: masculine animate, masculine inanimite, feminime and neuter. This gender reflects the natural gender when natural gender exists ("žena" (*woman*) is feminime, "chlapec" (*boy*) is maculine animate, "kuchařka" (*female cook*) is feminime) and is conventional otherwise ("stůl" (*table*) is masculine inaminite, "kniha" (*book*) is feminime, "auto" (*car*) is neuter). Grammatical gender affects the inflection.

Suffixes are used in inflection. Most nouns and adjectives are regular and follow a standard pattern: 14 noun patterns and 4 adjective patterns are taught in Czech schools (but some irregularities to those patterns are still regular and solvable by using more patterns).

When substituting a word, the substitution should be correctly inflected. A sentence containing a wrongly iflected word could be understandable under some but not all circumstances and in some special cases, the meaning of the sentence can change due to incorrect inflection.

Conjugation

Verbs are conjugated. The conjugation expresses

- person: first, second or third;
- number: singular or plural;
- tense: there are only three tenses: past, present, future;
- voice: either active or passive; and
- mode: indicative, imperative, present conditional or past conditional.

Verb aspect is perceived sometimes as affected by conjugation, sometimes as unchangeable property of the given verb. Nevertheless, there are two aspects in Czech, perfective and imperfective.

While some category combinations result in only one word (e. g. "píšu" (*I write*), first person, singular, present tense, active voice, imperfective aspect, indicative), some result in multiple words (e. g. "přišel bys" (*You would come*), second person, singular, active voice, perfective aspect, conditional; created as active participle + respective conditional form of verb *to be*; or "zpívali jste" (*You sang"*), second person, plural, past tense, active voice, imperfective, indicative; created as past participle + respective present form of the verb *to be*).

Free word order

Czech has very relaxed rules on word order. Some patterns prevail both in spoken and written language but many others are also correct.

Word order is not completely free. Not only are some patterns considered errorneous, word order is an important aspect in topic-focus articulation.

The 'verbal factorial of three' started rather as a mathematical joke but it should serve the purpose also here. The verbal factorial is six sentences, giving the words "Nemám rád faktoriál" (*I do not like factorial*) in all six possible permutations – and they are all valid Czech sentences meaning the same thing.

However, while "Nemám rád faktoriál" would most likely be really translated as *I do not like factorial*, the variant "Faktoriál nemám rád" is rather *As for factorial, I do not like it* and the variant "Rád nemám faktoriál" is rather *It is factorial which I do not like.*

The most frequent word order is subject – verb – object (SVO). However, object – verb – subject (OVS) is also common. The role of subject/object is determined by the grammatical case (and/or by the context, and/or by the subject-predicate agreement).

Some inflected forms are homonymous which can, because of free word order, cause subject/object ambiguity. The sentence "Růži dostala máma" (verbatim *Rose*/acusative *received mum* nominative) is clear because should the meaning be vice versa, the forms would be different, resulting in the sentence "Mámu dostala růže".

If mum did not receive a rose but rather heard a song, the sentence could be "Máma slyšela píseň". Both nominative and acusative of the word "píseň" (*song*) have the same form ("píseň"), but since the nominative and acusative is different for the word "máma", the meaning can still be unambigously derived.

The words "myš" (*mouse*) and "sýr" (*cheese*) both have the same form in nominative and acusative. Based on noun forms only, the sentence "Sýr viděla myš" could be interpreted both as *a mouse saw a chesse* and *a cheese saw a mouse.* In fact, only the first variant is correct because of noun-gender agreement. The verb "viděla" is feminime, so is the word "myš", but the word "cheese" is masculine inaminite.

However, if a children saw a school report, the sentence would be "Dítě vidělo vysvědčení" (or "Vysvědčení vidělo dítě"). Both "dítě" (*child*) and "vysvědčení" (*school report*) and neuter and both have the same form in nominative and acusative. The only way to determine the subject and object is therefore the context (and word semantics).

Determining subject and object can therefore be quite tricky, especially if the SVO order is not followed. There is probably no good answer to the question

whether subject/object ambiguity is a more severe problem than using an uncommon word as the answer depends on the reader, context and other factors. The risk of creating such ambiguities should however be kept in mind.

Subject-predicate agreement

In Czech, the subject and the predicate have to agree in gender (plus in person and number, which follows from conjugation). As explained before, four genders are distinguished, masculine animate, masculine inaminate, feminime and neuter.

The agreement manifests itself in indicative past tense active voice, in the passive voice (regardless of tense) and in conditional mode. To give some examples, "dívky zpívaly" and "děvčata zpívala" could both be translated as *girls sang* but since the gender of "dívky" (feminime) and "děvčeta" (neuter) is different, the form of the verb is also different.

Similarly, verbs are different in "chlapci byli pochváleni" (*boys were praised*) and "dívky byly pochváleny" (*girls were praised*) because the gender of "chlapci" (masculine animate) differes from the gender of "dívky" (feminime).

This should not present a problem for verb substitution in case a morphological analyzer and generator are employed. The analyzer would analyze the verb including its noun gender and the generator would generate a correct form for the substitution.

This can, however, present a problem when substituting the subject. In case someone wanted to substitute the less common word "děvčata" with the more common word "dívky" but left the verb in place, this would result in the sentence "dívky zpívala" which is not grammatically correct (and could actually be perceived as a spelling error of either the intended "dívky zpívaly" or unintended "dívka zpívala" (*a girl sang*)).

I should note that this becomes even trickier when the subject is coordinated as there are rules which gender determines the agreement when more are present in the subject.

Adjective-noun agreement

All adjectives have to agree with their head noun in case, number and gender.

This can be illustrated for example by the sentences "Mladá dívka přichází" (*a young girl is coming*, nominative), "Mladý chlapec přichází" (*a young boy is coming*, nominative), "Vidím mladou dívku" (*I see a young girl*, acusative) and "Vidím mladého chlapce" (*I see a young boy*, acusative).

Similarly to subject-predicate agreement, substituting the adjective for another one should not cause any troubles as the analyzer and generator would take care of generating a correct form. If the girl was not young but happy, a simple substitution would be enough to produce the correct sentence "Šťastná dívka přichází" (*a happy girl is coming*).

However, there is a similar problem: substituting the noun could result in a grammatically incorrect sentence. If for some reason the word "dívka" were substituted with its synonym "děvče", which has neuter gender instead of feminime, a simple substitution would produce the sentence "Mladá děvče přichází", which is incorrect (it should be "Mladé děvče přichází").

Using an analyzer to get the gender of the new noun and then generating a new form of the adjective with a generator would solve this issue (given that the analyzer and generator work flawlessly) but parsing would have to be employed to determine which adjectives depend on the noun and thus should be modified.

It is likely that any adjectives preceding the noun with no other word in between are headed by this noun but it does not have to be true: "na přední stranu napište jméno, na zadní odpověď" (verbatim *on front side write [the/your] name, on back answer*, *Write the name on the front side, the answer on the back side*) would be a natural counter-example. The word "zadní" is an adjective preceding the noun "odpověď" though it is headed by the noun "stranu".

There is one more issue with adjective-noun agreement, thought this is mostly an issue of morphological analysis. There is great homonymy among adjectives and an adjective could get analysed wrongly, which might later cause an incorrect form to be generated.

To illustrate, though an analyser is likely to cope with such a simple example, singular nominative of the adjective "zimní" (*winter*) is the same for feminime and masculine so there is "zimní bunda" (*winter jacket*, feminime) and "zimní kabát" (*winter coat*, masculine inaminite). It is however different for the adjective "teplý" (*warm*), there is "teplá bunda" but "teplý kabát".

Multiple negative

Multiple negative is used when expressing negation in Czech. When using a negative verb, relevant pronouns and adverbs are also negative. For example, there is the affirmative sentence "Půjdu tam" (*I will go there*) and its negation "Nepůjdu tam" (*I will not go there*). A more strict variant could be "Nikdy tam nepůjdu" (literally *I will not never go there*, *I will never go there*).

This could be an issue when substituting either the relevant pronouns/adverbs or, more likely, the verbs themselves. For example, it could seem easier to say *I have always lost* than *I have never won.* While the original sentence would be "Nikdy jsem nevyhrál" (literally *I have not never won*), the resulting sentence would be "Vždycky jsem prohrál" (*I have always lost*), not only has *not-win* changed to *lose*, *never* has also changed to *always.*

Verb aspect

As mentioned before, verbs have one of two aspects in Czech, either perfective or imperfective. Informally, perfective verbs represent an execution of someting, imperfective verbs represent the process of something. For example, "zpívat" (*to sing*) is imperfective. "Zpívám" means *I am singing*, "Zpívám ve sboru" means *I sing in a choir.* Contrarily, "zazpívat" is perfective. "Zazpívám" means *I will sing [a song].* When a perfective verb is formally in present tense, it actually represents future action.

Aspects should not present an issue for automatic simplification as verb pairs differing in aspect are generally considered different words by dictionaries. They could however confuse some readers and add to the overall complexity.

Sequence of tenses

Czech uses the natural sequence of tenses, tenses are not shifted in indirect speech and similar places. Some examples demonstrating this could be "Říkal, že píše dopis" (literally *He said he is writing a letter*, *He said he was writing a letter*) or "Říkal, že přijde" (literally *He said he will come*, *He said he would come*).

The absence of sequence of tenses should not be an issue for automatic lexical simplification as the tense of the verb is kept anyway. It should also not be an issue if, for a reason, someone wanted to automatically convert between direct and indirect speech. It could however become tricky if direct or indirect speech were transformed into informative sentences.

3. Text simplification

I have described the task of text simplification itself in the introduction and I would like to give more details in this chapter. I will discuss the target groups of this task and their special needs as the priorities are different for different target groups.

After that, I will describe the various layers of text simplification and some of approaches employed or attempted at those layers. I describe also lexical simplification, though some of its subproblems are described in more detail in chapter 5.

3.1 Target groups

I have already said that there are several target groups who can benefit from simplified texts. Here I would like to describe those groups in more details, especially to make some notes about what their complexity/simplicity criteria are.

It it worth noting that those groups need not be disjoint. A person with dyslexia can also be an L2 learner, a hearing-impaired person can also be a non-expert, etc. While the characteristics of all groups a person belongs to hold for the person, simplicity should be achieved and evaluated with respect to the main role the person has when accessing the text in question.

If a dyslexic student is an L2 learner and he/she asks for a text to practice the second language on, the guidelines for L2 learners should take precedence over guidelines for dyslexic people.

If, on the other hand, a dyslexic person reads a text in a foreign language because the information is available in this language and the person is by chance a learner of this language, the guidelines for dyslexia should be dominant when simplifying.

L2 learners

L2 learners are a common example of a target group for text simplification. Those people can speak a language very well but for some reason, they want to learn another. They might be learning the language somewhere it is spoken or outside of any such place. This condition could slightly affect how much they are accustomed to written text (instead of spoken language) and how capable they are of recognizing various pragmatic references in the text, be them geographical, cultural, ...

Gregg and Krashen (1986) suggests that L2 learners should not be provided with texts so simple that they would not present any challenge to them. The texts should be easy enough to be understood but hard enough to teach. This is true for both syntax and vocabulary (and, in a way, especially for vocabulary).

If the meaning of an unknown word can be easily, unambigously guessed from the context, leaving it in the place is a good way for the learner to learn the word. The number of unknown words must, however, be kept low in the text, their cotext is also important.

There are studies which have proven simplified texts to improve L2 learners text comprehension, Long and Ross (1993) and Gardner and Hansen (2007) are examples of that. Some authors, for example Long and Ross (1993), suggest to solve the conflict between understandability improvement and L2 acquisition by elaborating texts.

In any case, grammaticity is very important in texts for L2 learners. While they would be able to understand the meaning of a text containing some grammatical errors, they could easily retain those ungrammatical structures. Grammaticity should definitely be preferred, even if it comes at the cost of lowered understandability and possibly the need of searching for some words/structures in textbooks or consulting an instructor.

Dyslexic people

Dyslexia is probably the most well known disorder of what we could call dysfunction family. Sometimes it is also called a reading disorder, which is self-descriptive of the most notable symptom. While the difficulties caused by dyslexia are not limited to reading (most importantly, they also cover sound imagination and time management, see Zelinková and Čedík (2013) or Zelinková (2015)), dyslexia *does* complicate reading.

Dyslexic people have trouble making up words from individual characters. While this becomes more evident when reading aloud, it happens also when reading without speaking. Shorter words (which need less combining of smaller units) and more frequent words (which are often present in texts and thus can be more easily learnt as a unit themselves) are therefore easier for dyslexic people to read and understand. This is also evidenced by Rello et al. (2013). Substituting uncommon words with their more common synonyms is a great help for dyslexic readers, no matter how easy it is to understand the word just by context.

Dyslexic readers are also predisposed to interchanging some letters or words, mostly those which are very similar to each other. It can be hard for them to distinguish for example the 'd' and 'b' characters. Avoiding any unnecessary similarity and making sure that similar words do not occur close to each other are also ways to simplify reading for dyslexic people.

Zelinková and Čedík (2013) point out that dyslexia is also connected with difficulties regarding orientation in the text, especially in long paragraphs. It is difficult to find the desired information (even shortly after reading it) as well as not to get lost in the text while still reading. Cutting long parathers into several shorter ones can help a lot.

Aphasic people

Aphasia is a language disorder which affects the ability to produce and understand speech and/or written text, it results from brain damage.

Syntactic complexity is a great issue for people with aphasia. Passive voice, coordinated and relative clauses seem to greatly reduce their ability to understand the text (Caplan (1992)). For this reason, text simplification for aphasic people should focus on syntactic simplification.

Even though syntactic complexity is the worst, uncommon words and long

sentences are also hard to understand for aphasic people (Shewan and Canter (1971)). Sentence splitting can help, especially if the subject is still clear.

Autistic people

Based on my consultations with NAUTIS[1], a Czech organization offering varied service to people suffering from autism spectrum disorders, the most severe issue with text comprehension in autistic people is pragmatics. Any pragmatic simplification would be of great help for autistic people.

Visual arrangement can also be an issue. While this might seem to be mostly a formatting issue, it need not be the case. Splitting the text into smaller units could improve readability a lot, especially paragraph splitting. Enhancing the text with simple questions inciting to rethink the read text and realize its meaning could improve understandability.

Hearing-impaired people

Hearing-impaired people often have great difficulties understanding written texts and could greatly benefit from their simplification. It is however worth mentioning that some of hearing-impaired people's issues with written language can be attributed to the methods of teaching the language to them (Komorná (2008)).

Hearing-impaired people usually have little knowledge of grammar. This could be an even more serious issue in languages with rich morphology. The knowledge of combining smaller text units into larger ones also tends to be low. Short, simple sentences written just one after another are most easily understood by hearing-impaired people, even though they can be intrusively (even irritatingly) primitive for a skilled speaker.

Vocabulary is often also an issue. The vocabulary of a hearing-impaired person is usually smaller than the vocabulary of a hearing person. A hearing-impaired person is likely to know few synonyms for words, if any. It can also happen (and it does happen in Czech, as reported by Komorná (2008)) that a word is recognized only in some of its forms (this is usually a consequence of the morphology, though it could happen because of recognizing not the word, but rather its context). Choosing the known synonym is of great help (if it is, by any means, possible to guess which synonym will be known by the person), preferring some forms to others can also increase the readability and understandability of the text.

All kinds of figurative expressions seem to be harder to understand by hearing-impaired people. Those include, but are not limited to metaphors, metonymy, proverbs and sayings. Explaining those increases understandability a lot.

Last but not least, hearing-impaired people, as a consequence of several interacting problems, tend to have lower general knowledge. Even texts meant for laymen are therefore too difficult for them. The only way to overcome this (beside eliminating the problems causing the lack of knowledge) is to make the texts even simpler and explain even the basics.

[1] http://www.praha.apla.cz/

Visually-impaired people

Based on my e-mail consultations with Okamžik[2] and Tyfloservis[3], Czech organizations assisting visually-impaired people, the issues visually-impaired people might experience are more often caused by formatting than content itself.

Supposing visually-impaired people have a screen reader, refreshable braille display or other device to present them with the text, there is usually no need to simplify text content for them just because of the visual impairment.

Text structure can however present an issue as those devices do not distinguish headings from surrounding text, they sometimes omit numbers in numbered lists, treat dates and abbreviations in a specific (often counter-intuitive) way. Special punctuation can also cause some trouble. Presenting tables or nested lists is problematic, not necesserily because the devices were bad but because the structure is hard to linearize.

A specific task is describing visual information, most importantly images or graphs. Those description should always begin with the image/graph in its entirety and only then proceed to describing subareas and details.

However, the above suggest that tailoring a text to visually-impaired people is more a task of structure alteration than text simplification.

Children

This subsection is largely based on my intuition and consultations in Sun publishing house[4], which specializes in child books. Unfortunately, I was not able to find a good source on child text *writing*. A lot of studies concerning *talking* to children exist, though. Hayes and Ahrens (1988) or Brodsky and Waterfall (2007) are examples of that. Studies dealing with simplification for children also exist, for example De Belder and Moens (2010), though I was not able to find any discussion about specific child needs.

Depending on their age, young children can have great difficulties understanding written text that is not adapted for them. Their difficulties usually relate to attention span and abstract imagination. It is however worth mentioning that those difficulties fade out as children get older. I have consulted the specifics of child texts with an editor specializing in child books, and this editor believes that at the age of about 9 years, a child can read an ordinary text given the topic is interesting for the child.

When children are old enough to learn to read, they usually already know most of the grammar specifics of their native language and can understand various syntactic structures well (though they might not be able to produce such structures correctly). They are also familiar with many phrasemes and sayings, though they might have difficulties understanding those not heard before.

Abstract vocabulary, on the other hand, can be hard to understand and tangible stories and characters are easier to follow than abstract ones.

Text simplification should therefore focus mostly on splitting (because of the attention span) and abstract word substitution or explanation.

[2]http://www.okamzik.cz/

[3]http://www.tyfloservis.cz/

[4]http://www.sun-knihy.cz/

Laymen

In general, the only difficulty laymen have to face when dealing with expert text is technical vocabulary (though this vocabulary could also include some phrasemes or new senses or valencies of otherwise known words).

The unfamiliarity with the field can however make sentence parsing or pronoun resolution much more difficult since the intuitive use of context and other knowledge to disambiguate the text is not possible.

Still the simplification should focus on lexical simplification and possibly explanation generation. The potencial issues are substituting specific (though complex) terms with more general words which are not searchable further (and thus preventing the reader from finding out more using other sources) and introducing ambiguities either by substituting several occurences of one term with multiple different substitutitions or substituting several different terms with the same substitution.

People with low literacy skill

There are several reasons for people to have low literacy skills, however they usually all result in slow reading and thus difficulties with understanding longer sentences, complex references and long words. Unwillingness to read further could also be an issue.

The simplification should focus on substantial reduction of text length. If such reduction is not possible, sentences should be made as short as possible and short paragraphs should be ensured.

Using patterns common in spoken language could also help when altering syntactic structure of the text.

Foreigners

Foreigners (and, to some extent, immigrants) are a special target group. These people are not actually attempting to learn the language in question, they generally know little about its grammar as well as they know only little vocabulary but they need to obtain some information provided in the language. In fact, the existence of this target group is a consequence of the imperfectness of machine translation.

Understandability should be preferred over any other criteria. Grammaticity of resulting text is not important. International words (or possibly words similar to their counterparts in the foreigners' mother tongue) should be preferred in lexical substitution, even if they are longer or less common in the language.

3.2 Measuring simplicity

In the task of text simplification, being able to decide which text is simple and to measure simplicity of any given text seems to be a crucial subtask. However, there is no universal definition of text simplicity and therefore no universal tool of measuring it.

Nonetheless, some methods to measure text simplicity exist. There are various readability formulas which take some features of the text as their input and produce some values. In most cases, the resulting values give an expected minimum grade of education completed by a reader who would enjoy and understand the text.

Before describing some of those formulas, I would like to discuss readability and understandability more as they are the key concepts.

3.2.1 Readability and Understandability

Recent works on text simplification distinguish readability from understandability. Readability is reading ease, the better readability, the easier it is to read the text, to get through the text, to perceive the text as easy to follow.

Understandability, on the other hand, expresses the ease of obtaining information from the text. The better understandability, the more will the reader learn from the text (or maybe more exactly, the less information will remain concealed to the reader – the fact that a skilled computer scientist does not learn anything new from a text on basic algorithms does not imply that the text itself had low understandability).

Even though readable texts tend to be understandable and understandable texts tend to be readable, this need not be the case.

A common case of text with high readability and low understandability is a well-written expert text read by a layman. Such a text may contain well-structured sentences, a reasonable proportion of pronouns and other qualities contributing to a great readability. However, if some previous knowledge is assumed, the understandability for someone who does not have the presumed knowledge is very low.

On contrary, a poorly-written introduction text should still be well understandable for an expert in that field. Even if the sentence structure is complex, the text contains ambigious references, some sentences are very long and contain many nested phrases or there are other contributions to bad readability, the expert is likely to understand the text well.

The examples however do not have to be so extreme. A report on an event might be badly understandable because time settings and causalities are not clearly stated, but since each phase of the event is told in its own paragraph, the reader does not realize the unclarity. Clear subjects and easily dereferrable pronouns can then add to overly good readability.

The combination of good readability and low understandability indicates some deceptiveness of the text. Such texts are likely to be read by their readers and seemingly understood by them. The information learnt by the user may however be very different from the information really present in the text.

Readability and understandability is generally considered a quality of the text but it could rather be a quality of the text-reader pair. This is not to say that a text which is nearly unreadable for one person would be delightfully readable for another person (thought it could probably happen), rather to stress out the role of target reader.

There are several more things that should be kept in mind when dealing with text readability, though they should not be used to give excuse for bad scoring.

First, there is a kind of temptation to keep reading, which some texts provoke in their readers. This is more common in fiction though it could also be the case of science books. Such temptation can partially compensate for lower readability as the reader is motivated not only by the process of reading itself but also by the desire to learn the rest of the text. This temptation could be partially predictable using features similar to those used by Louis and Nenkova (2013).

Second, readers have various motivation for reading a text and this motivation could also surmount bad readability. While the temptation to keep reading is provoked mainly by the text itself, this motivation is crated by the circumstances. A student is likely to keep reading an assigned book even if its readability is not good, a programmer is likely to keep reading explanation on his programming language behaviour even if its understandability is low, etc.

3.2.2 Formulas

Many different readability formulas have been proposed. Most of them have been designed for English, some of those have been adapted for other languages as well, some of the other formulas have been designed directly for another language.

Prchalová (2013) gives three formulas for Czech and four formulas for English and I stick to describing those six formulas.[5]

It is worth mentioning that most of the formulas are based on some surface measures and have been shown by their authors to predict the simplicity, usually measured by reader's success rate in a comprehension test. However, some doubts about the indicative power of such features exist. Pitler and Nenkova (2008) showed that features like average number of characters per word or average number of words per sentence do not correlate with human understandability complexity ratings for Penn Discours Treebank (for details on the treebank, see Prasad et al. (2008)).

Flesch reading ease

Flesch reading ease is one of the oldest readability formulas. It was developed by Rudolf Flesch in 1948 and it was highly inspired by the formula he had developed three years earlier.

The formula measures reading ease, it is often called a readability formula (exactly as I did in the previous paragraph) but under the distinction between readability and understandability, the formula rather measures understandability.

Flesh based his formula on an analysis of texts used in standard comprehension tests by McCall and Crabbs (1926) and children's answers to comprehension questions regarding those texts. The grade of a child who could answer three quarters of the comprehension questions correctly was used as a feature of the texts.

The result of the original formula was the average grade of a child who could answer the given amount (three quarters) of comprehension questions, but the formula presented in 1948 results in a number ranging from 0 to 100. The higher the result, the better the readability.

[5]They are six and not seven because one of them, the FOG index, is reported to be used for both languages.

The reading ease formula is often quoted and sometimes used but it is worth mentioning that Flesch proposed two formulas in his paper, which should be both used when measuring readability. The first is the reading ease formula, the second one is a human interest formula, which measures an effect similar to the temptation to keep reading described before.

The formula for reading ease is

$$\text{R.E.} = 206.835 - 84.6wl - 1.015sl,$$

where wl refers to average word length in syllables and sl refers to average sentence length in words. Whole text can be used to compute the values but only a few samples can be taken as well.

It is worth noting that the original formula utilized number of affixes instead of syllables. Flesch suggests that syllables are easier to count (mostly because it is mechanical routine) and the results should be similar.

I will include the formula for human interest for completeness:

$$\text{H.I.} = 3.637pw + 0.314ps$$

Here pw refers to the number of personal words per 100 words. Personal words are defined as all nouns with natural gender and all pronouns except neuter genders (and words 'people' and 'folks').

The value ps refers to the number of personal sentences per 100 sentences. This definition is more vague. In short, personal sentences are all sentences addressed directly to the reader (questions, commands, ...), spoken sentences and grammatically incomplete sentences.

Even though both formulas are designed for the evaluation of longer texts, the reading ease formula can be used also for the evaluation of one sentence only. Both features, average word length and average sentence length, can be easily computed automatically.[6]

Some works, for instance De Belder and Moens (2010) or Yakovets and Agrawal (2013), use Flesch-Kincaid score when evaluating text simplification (either alone or as one of evaluation methods). Flesch-Kincaid formula uses the same features with different weights and results directly in the target grade.

Given how the reading ease formula was derived, the outputs of Smith and Taffler (1992) are remarkable. They took accounting narratives of failed and surviving companies, turned them into CLOZE tests (see Taylor (1953)) and had both accounting undergraduate students and accounting practitioners complete the tests. They found out that while practitioners' CLOZE results correlated with Flesch reading ease score, undergraduate students' CLOZE results did not.

According to Prchalová (2013), Flesch reading ease is not used for Czech. Based on a few experiments, it seems that it cannot be easily reused for Czech as the results ranged from negative values to values over 200 and, by manual inspection, did not seem to reflect the sentence complexity well.

[6]Or at least it can be estimated automatically. For example, Flesh instructs to compute the sentence length based on thoughts rather than on punctuation which is hard to do artificially.

FOG index

FOG index was proposed by Gunning (1952). It combines average number of words in the sentence with complex word ratio. The formula is

$$0.4 \left[\left(\frac{\text{words}}{\text{sentences}} \right) + 100 \cdot \frac{\text{complex words}}{\text{words}} \right].$$

The result of the formula is the number of years of education a reader must complete before he/she is able to understand the text.

According to Prchalová (2013), a Czech adaptation exists. This adaptation uses a sample of exactly 100 words (the number of sentences is rounded for the calculations). The adapted formula is

$$\text{FOG index}_{\text{CDV}} = 5 + 0.4 * (A + L),$$

where A refers to average number of words per sentence and L refers to the number of words of at least three syllables.[7]

The result of the adapted version is a number which can be mapped to one of four levels of difficulty, with results in the range 20-24 representing a very easy text and results exceeding 40 representing a hard to read text.

SMOG

SMOG grading was proposed by McLaughlin (1969). It offers a very simple procedure to obtain the grade needed for the reader to fully understand the text.

To obtain the SMOG grade, one should take three groups of ten sentences (one at the beginning, one in the middle, one at the end of the text) and count all words in those sentences which have three or more syllables, then estimate the square root of their count and add three.

McLaughlin argues that a sample of 30 sentences is large enough to be representative of the text and small enough to still add valuable information and be bearable to compute by hand.

The formula, $3 + \sqrt{p}$, where p refers to the number of pollysylabic words (words with three or more syllables) in the sample, is actually simplified so that it is easier to compute manually. When computing automatically, a more exact version could be used, which is $3.1291 + 1.0430\sqrt{p}$.

While Flesch designed his formula to predict the grade of a child who would be able to answer three quarters of questions correctly, McLaughlin predicts the grade of a child who would be able to answer everything. He argues that complete comprehension is a more reasonable requirement.

SMOG grading is designed for whole texts and probably would not give good results if applied to a single sentence only.

Fry graph

Fry readability was proposed by Fry (1968). Fry formula is not exactly a formula. In instructs the user to count two features of the text in question, those two

[7] Similarly to Prchalová, I found the constant dubious.

features then function as coordinates to a designated graph. Readability can be read from this graph.

The two features used are average number of syllables and average number of sentences; both values are taken from a sample of 100 words.

Mistrík formula

According to Prchalová (2013), Mistrík formula was proposed by Jozef Mistrík in 1968. It was designed for Slovak, which is a language very similar to Czech.

The formula assumes a sample of the text and takes four features into consideration: average number of syllables per word (P), avarage number of words per sentence (S), the number of distinct words in the sample (L) and the total number of words in the sample (N).

The readability is computed as follows:

$$R = 50 - S \cdot V \cdot \frac{L}{N}$$

Prchalová (2013) states that Mistrík formula is used for Czech without any changes. The formula si designed for measuring readability of a sample larger than one sentence, its variables however permit to use it also when only one sentence is available too. For this reason, I have attempted to use the formula in the substitution ranking step.

Complex measurement of text complexity

According to Prchalová (2013), complex measurement of text complexity[8] is based on the work of German psychologist Käte Nestler.[9] According to Janoušková (2008), the formula proposed by Nestler was then adapted for Czech by Jan Průcha and further improved by Miroslav Pluskal. This measure focuses on educational texts, especially textbooks.

I will not give the exact formula here but I will say that it sums two complexities, syntactic and semantic. The syntactic complexity is computed from average number of words per sentence and per phrases. The semantic complexity is computed from several counts: total word count in the sample, counts of concepts, common words, technical terms, factual terms, numerical tokens and repeated words.

Becuase the formula employs features like the number of technical terms in the sample or the number of factual terms in the text, it cannot be simply measured automatically unless dictionaries to recognize such terms are created.

3.3 Lexical simplification

Lexical simplification deals with substituting complex words with simpler ones. The meaning should be preserved, which would suggest substituting with synonyms and this is indeed the most common case. It is however also possible

[8] "Komplexní míra obtížnosti textu in Czech"

[9] Unfortunately, I have not been able to find any references more specific than Käte Nestler – Textkompliziertheit.

to use words with similar meaning, hypernyms (more general words, e. g. *animal* is a hypernym of *dog*) or hyponyms (more specific words, *dog* is a hyponym of *animal*).

As in all subtasks of text simplification, the resulting sentence should be grammatical. This might require generating a coorect form of the substituting word and/or altering the rest of the sentence to match the new word.

One of the issues in lexical simplification is word sense ambiguity. If a word comes in many senses, it might have different sets (though not necessarily disjoint) of substitutions for each of the senses. Obviously, a substitution for one sense should not be used for another sense.

Lexical simplification is usually divided into four steps:

- Complex word identification: the name is probably self-explanatory, complex words are identified as only complex words should be substituted
- Substitution generation: for a complex word, all possible substitutions are generated at this step; For exapmle, for the complex word "zápal" (*inflammation, enthusiasm*), we might want to obtain substitutions "nadšení" (*enthusiasm*), "zaujetí" (*keen interest*), "zánět" (*inflammation*) and "zanícení" (*ardour, inflammation*).
- Substitution selection: from a set of complex word substitutions, only those matching the complex word meaning in the text are kept; the others are eliminated at this step (this is the step to employ word sense disambiguation if it is done explicitly); If the word "zápal" occured in the context of "zápal plic" (*pneumonia*), we would want to eliminate "nadšení" and "zaujetí" at this step and keep only the substitutions "zánět" and "zanícení".[10]
- Substitution ranking: all remaining substitutions are ranked at this step to reflect their suitability for the case

In general, lexical substitution substitutes one word with another. There is however the ever-present question of what a word is. To refresh some problematic examples, treating 'Czech Republic' as one word could be reasonable, treating 'computer science' as one word could be reasonable too (thought it could also be reasonable to treat it as two separate words), similarly treating 'lock chart' as one word is arguable.

This question is important both when selecting words to be substituted and when generating the substitutions. The international phrase "à la" could be easier when substitued with "podle" (*according to*) or "jako" (*as*), depending on the context. The word "bohoslužba" (*church service*) could be easier than "mše svatá" (*Holy Mass*) for some readers. Similarly, substituting "vlče" (*wolf cub*, a non-compound term) with "vlčí mládě" (verbatim *wolf cub*) could result in a simpler text.

Treating such word combinations as one word would however complexify the process of morphological analysis and generation.

It is also worth pointing out that while uncommon words are often considered complex, substituting such a word when a part of a phrase (like a quotation,

[10]Though we could argue whether the word should be substituted at all given than "zápal plic" is a traditional name of that condition.

a saying or a proverb) could increase the complexity of the text. Some sayings could still be easily understood, some could become very hard to find or guess.

Transforming "Devatero řemesel, desátá bída" (literally *nine trades, tenth destitution* (using a special form of the numeral nine); *A man of many trades begs his bread on Sunday*) to either "devět řemesel, desátá bída"' (using the basic form of the numeral nine) or "devatero řemesel, desátá nouze" (literally *nine trades, tenth poverty*) would probably still allow the reader to understand the meaning.

If, however, the complex word "bačkory" were substituted with the word "přezůvky"[11] (*slippers*, the first word is a hyponym of the second one) in the phrase "Natáhnul bačkory" (literally *He pulled the slippers*, *He kicked the bucket*), it would be very hard to understand (and, because of the substitution, to look up).

3.4 Syntactic simplification

Syntactic simplification aims to decrease complexity by simplifying the structure of sentences. The task is in a sense wider than lexical simplification as there are many kinds of complex structures, each of which is solved in a different way.

Here I would like to present some syntactic simplification strategies which could be potentially employed to simplify Czech texts. I discuss their effects and mostly the dicey parts of their potential implementation.

Sentence splitting

Sentence splitting is a way to reduce the average sentence length, though the overall length of the text could be slightly increased because of the need to insert some words, e. g. to repeat the subject in the newly created sentence.

A very simple strategy would be to split a sentence constituted by phrases coordinated by a coordinating conjunction "a" (*and*). It could be however a little bit tricky to find such conjunctions as the conjuction could be used in coordinated constituents such as objects, or even subjects or predicates.

Still, splitting on conjunctions generally seems to be a good idea. If the conjunctions were left in place, the resulting sentences would probably not be rated as stylistically good but could serve their purpose anyway. It would probably not be reasonable to just delete the conjunctions because they help to perceive the relations between the sentences and thus simplify reading. If sentence splitting is done by humans, they usually move the conjunction in the newly-created sentence to improve the stylistics, however I have not been able to figure out reasonable rules for such shifts.

Splitting on the pronoun "který" (*which*) also seems promising. Being able to link the pronoun with the referred word would however be necessary. Chances are that the first preceding noun of respective gender would be the one referred to by the pronoun, but this hypothesis needs to be tested first.

[11] Actually, the word "bačkory" is more common according to CNC.

De-passivization

Passive voice is usually considered more difficult than active voice, sometimes the passive voice is not interpreted correctly by the reader and is understood as if active, resulting in subject and object swap.

Any de-passivization would highly depend on some parsing tool, on the other hand, if such tool was employed, de-passivization should be easy in some cases.

The presence of both subject and object in the sentence is crucial for the de-passivization. This does not need to be the case, a passive-voice sentence could be both "Žák byl zkoušen učitelem" (*The student was examined by the teacher*) and "Žák byl zkoušen" (*The student was examined*).

Subject and object transformations should be easy if a morphological generator is available, as the subject is always transformed into nominative and object is transformed into acusative.[12] Subject and object position in the sentence should be swapped, the verb should be transformed into active voice.

Verb transformation could be problematic because of subject-verb agreement, "Žák byl zkoušen učitelem" (*The student was examined by the [male] teacher*) transforms into "Učitel zkoušel žáka", "Žák byl zkoušen učitelkou" (*The student was examined by the [female] teacher*) transforms into "Učitelka zkoušela žáka" and "Žák byl zkoušen učiteli" (*The student was examied by teachers*) transforms into "Učitelé zkoušeli žáka".

However, if the subject is clear from parsing and gets correctly morphologically analyzed, a verb can be generated based on the subject analysis.

Though, another problematic step could be stentence reordering if more valency slots of the verb are used. For instance, "Žákovi byl zabaven mobil učitelem" (*The cellphone was seized from the student by the teacher*) could be transformed both to "Žákovi učitel zabavil mobil" and to "Učitel zabavil mobil žákovi". The decision which sentence is more appropriate would highly depend on the context.

Pronoun resolution

Czech genders allow an extensive use of pronouns instead of referred nouns. Such use may help to reduce word repetitions and shorten the sentences, it can however also complexify the text for a reader not accustomed to wide pronoun use.

For instance, imagine complex sentences "Pes šel do boudy, viděl jsem ho" (*The dog went into the dog-kennel, I saw it* [the dog]) and "Pes šel do boudy, viděl jsem to" (*The dog went into the dog-kennel, I saw it* [the process of going]). Even though the semantics is very similar, if not the same, the syntax is slightly different.

Now if the dog was instead a bitch, the first sentence would turn into "Fena šla do boudy, viděl jsem ji". The pronoun "ji" could however refer either to the bitch or to the dog-kennel, and even though the reference should be clear from the context, it requires some insight and could possibly confuse the reader for a moment.

Another sentence to demonstrate to use of pronouns could be "Když k němu letěl míč, držel v ruce tašku, hned ji pustil, aby ho mohl chytit" (*When the ball*

[12] I believe this article confirms that the subject would always become acusative during de-passivization: http://nase-rec.ujc.cas.cz/archiv.php?art=5209

flew towards him, he was holding a bag in his hand, immediately he dropped it [the bag] *so that he could catch it* [the ball]). The meaning is clear because of gender agreement but requires noticing the agreement.

Last but not least, a complex sentence snippet could be "park s fontánkou, který snadno najdu" (*a park with a fountain which* [the park] *I can find easily*) or "park s fontánkou, kterou snadno najdu" (*a park with a fountain which* [the fountain] *I can find easily*). Substituting the pronoun "který" (*which*) with either "ten park" (*the park*) or "tu fontánku" (*the fountain*) could increase readability (though such altered text would probably sound unnatural to a native speaker).

However, such substitutions would require reliable results of coreference resolution, which is not yet solved.

Pronoun insertion

It is possible to drop the subject in Czech and it is often done if the subject is a personal pronoun. In such case, the subject is (almost) clear from the verb form. For some readers, it could be however simpler if the subject was there.

No further resolution is required, morphological analysis of the verb is enough to derive the correct pronoun (except for gender and/or number ambiguity of some third-person forms).

However, word reordering could be needed to make the sentence sound natural. For example, a sentence with dropped subject could be "Viděl jsem to" (*[I] saw it*) while the sentence with the subject re-inserted would be "Já jsem to viděl".

Inventing some rules for word reordering after pronoun re-insertion would therefore be needed to really employ this strategy.

3.5 Pragmatic simplification

To the extent of my knowledge, pragmatic simplification has not yet been researched. The goal of pragmatic simplification is to make the meaning of the text more clear and easier to understand. This includes for example explaining metaphors, rephrasing polite ways to ask/forbid the reader to do something into explicit orders, stating the context or including some implicit information.

The distinction between lexical simplification and pragmatic simplification can be hazy as there is no hard distinction between substituting a word (or a phrase) with another word of the same meaning and explaining the word.

The distinction between syntactic simplification and pragramtic simplification can also be hazy. For example, rephrasing thank-yous for not doing something into orders not to do that thing can be perceived both as just a syntactic operation and as a pragmatic simplification.

Enriching the text with implicit information or explanations can also simplify the text. As I describe in section 4.3, this strategy is often employed by humans when simplifying a complex text. It however seems to be very difficult to do it automatically as it requires understanding the text in question.

An exception to the need of understanding the text could be explaining complex terms. Such terms could be searched for in a knowledge base like Wikipedia and if found, a very short summary of the respective entry could be used as an explanation.

4. Experiments on people

In the course of working on the thesis, I have come with many ideas for experiments involving people to get better insight into simplicity and the factors affecting it. Most of those ideas have not been realized for various reasons, including doubts about their feasibility or merits. I will describe those which has been taken to a step when some participants really took part in them. There were three such projects.

The first experiment concerned with explicitly rating the readability of sentences with implicit regard of the context – whole articles would be rated sentence after sentence. As it became more and more clear that context-aware simplification would not be feasible in this thesis, I discontinued the experiment just after testing the concept. The experiment seems to be viable though a lot of details would have to be tuned.

The second experiment was about comparing pairs of sentences – selecting which one of the two sentences is simpler. This experiment has been finished and evaluated.

The third experiment was human text simplification. Three annotators were asked to manually simplify a set of complex sentences. The main goal was to get an insight into what people do to make text simpler, which strategies they use and which not. The experiment has been finished and evaluated.

4.1 Sentence readability ranking

The idea of this experiment was to task the participants to rank all sentences in a complete article with respect to their readability. The data should reveal not only what adds to the complexity of sentences in the sentence itself but also what is the role of context and whether specific characteristics of preceding sentences affect the perceived readability of following sentences.

Participants should read an article sentence by sentence and during the course of reading they should also rank each of the sentences. A sentence should be ranked without peeking either backward or forward in the text, though some context is expected to be still kept in memory.

The ranking should reflect the readabililty. The ranking scale is contininous (to the limits of the media used) though some discretization is expected to be used in the evaluation. Participants should be encouraged to rank partially intuitively, mostly because the perceived readability is subjective and any objective criterion could be more easily measured automatically.

I was also thinking about measuring the reading time of every sentence and them comparing it with the rankings, which is also a reason to encourage participants to rank intuitively and, more importantly, quickly. (Some normalization with respect to sentence length would of course have to be performed.)

Even though the original idea was to create a simple web and/or Android application (which would allow both controlling and monitoring the shift between sentences), the experiment has been tested with paper sheets. The sheets were scanned after the experiment and are attached to this thesis via the enclosed CD (see attachment A.1).

I decided to use paper sheets partially because preparing a few of them seemed both easier and quicker than preparing a complete application, partially because it allowed me to better observe the process of ranking by the first (and only) two participants.

As suggested before, using paper as the media limits the possibilities of controlling whether the participant peeks backward/forward or not (though, if really desired, this could be achieved by presenting every sentence on its own paper). Such a peek could occur both as a conflict with the guidelines or just unintentionally. Paper sheets however still allow to continously go through the text and rank the sentence one after another.

It is also worth mentioning that there is probably not a good way to measure the reading time of each sentence when using paper sheet.

I have tested the experiment with two participants. Both of them were males, native speakers of Czech, university students, accustomed to reading causual texts in Czech (and expert texts on their field in English). No written instructions exist, I instructed the participants orally and stayed present during the whole experiment to answer any questions or solve any trouble.

The instructions were however similar to the description above: I asked the participants to go through the texts, read each sentence and mark its readability on the scale. I asked them not to be afraid of ranking intuitively but to rather try to rank honestly. I told them they should not go back in the text but they of course should perceive the sentence in the surrounding context.

There were 6 texts in total, one of them being a prefix of another. Those two texts allowed to compare the rankings of the two annotators, though larger set of annotations would be needed to obtain really reliable outputs. One of the annotators annotated two of the texts, the other one annotated the remaining four.

There were several valuable outputs of this experiment. Some of them could be attributed to more or less severe flaws in task specification and settings, some of them are likely to hold even if the task is tuned.

- The definition of readability is vague. Even though the definition could hardly be more exact (as any criteria like the number of words or the presence of technical terms *presume* the factors of readability), this increases the participants' uncertainty regarding the ranking.

- One of the participants did not feel comfortable ranking any sentence as badly readable. His comments suggested that he felt stupid for not understanding his first text well (the text in question was an expert text on Factor V Leiden trombophilia and the participant does not have any special medical knowledge).

  Such a problem occurs sometimes also in user interface testing and could probably be well reduced by better explaining that the task is in no way to judge the participant himself but to judge the quality of the presented text. Assuring the participants of being not *tested* but *testing* is often accented in user interface testing and could be useful in this case for the same reasons.[1]

[1] I should admit that my primary sources on such testing are specialized courses at Czech Technical University: A4B39TUR Testing of user interfaces (slides should be available from

- Even when the participants did not feel ashamed for not understanding the texts, they used the ranking scale slightly differently, they especially used different extreme values and different 'default' value.

  Performing a kind of normalization would probably be a good way to overcome this issue.

- A little discussion revealed that the readability *is* different from understandability. The participants experienced moments when they could read the text easily, fluently but were not really able to understand it without stopping and re-reading some of its parts.

  A good question is how to measure understandability in the experiment (especially since the understandability is rather a property of a larger part of the text than of individual sentences). This is especially difficult to decide with respect to the fact that presenting any questions during the reading (and ranking) process would likely distract the reader more than the ranking does, and thus would decrease both readability and understandability.

  Using CLOZE tests as suggested by Taylor (1953) could be viable but attention would have to be paid to technical terms. Preparing special comprehension tests would also be a way to go, this would however require deep understanding of each of the texts and could easily get very time-consuming.

4.2 Sentence pair comparison

To get a better idea about readability factors, I prepared an experiment in which the participants were tasked to choose the simpler sentence of a sentence pair. Most of all, I wished to find out whether the common features like sentence length and word commonness really correlate with the perceived simplicity. However, I should admit that no exact hypothesis had been formulated for the experiment; both data preparation and evaluation had been driven by intuition and 'pure numbers' (that is for example deducing something from the fact that three quarters of annotators made the same decision but not computing p-values, Cohen's kappa or other measures).

Both sentences in each sentence pair had the same meaning (or were originally intended to have the same meaning, some of the annotators expressed their doubts about that); they were sampled from translation references by Bojar et al. (2013). This dataset contains many reference Czech translations for 50 English sentences. As the dataset README states, the number of reference translations greatly varies accross sentences – ranging from as little as six translations to as many as over one million.

I sampled some sentences from the reference dataset (based on statistics like sentence length or the length of the longest noun cluster), paired them each with each another and took a part of that pairs.

https://cent.felk.cvut.cz/courses/Y39TUR/?page=slides); A4M39NUR User interface design (slides should be findable at http://nur.felk.cvut.cz/); and B4M39PUR Psychology of user interfaces, now probably transformed to B4M39PUR1 Psychology in HCI (which unfortunately does not have slides available).

The annotators were then presented with the selected pairs and were asked to decide which sentence is simpler. They had four choices available to distinguish between the simpler sentence being strictly better and slightly better but they did not have the choice to rate the sentences as equally complex. There were 100 sentence pairs to compare. There were 7 pairs which occured twice, each time with different ordering of the sentences; the annotators were not informed about this.

The task was done online. I created a simple web application which allowed the annotators to select the simpler sentence by simply clicking the corresponding button. The buttons were organized vertically, the topmost corresponding to the first sentence being strictly better than the second one. The buttons were also labeled, either with the sentence itself (the topmost and bottom-most buttons) or with a text indicating that the upper/lower sentence is slightly better. The annotators did not have a strict time limit to finish annotations, they could fulfill the task at self-paced times. They also had the opportunity to stop annotating and get back to it later. The instructions were to consider general simplification but if in doubts, suppose foreign students coming to study at Czech universities.

In total, 8 annotators participated in annotating the sentences. There were 5 males and 3 females. The annotators included a high school student, university students, high school alumni and university alumni. Only one annotator was not a native speaker of Czech, the others were.

Results

In fact, the most important result of the experiment is that since the sentences compared differed in more than one characteristics, it is hard to really generalize. Even though 100 sentences is a lot, there would be a lot of uncertainty, which is yet bolstered by relatively small number of participants.

Some conclusions can however be made even so. Most of all, the comparison seems to really reflect perceived simplicity and not random factors as for 6 of 7 repeated pairs, 1-2 annotators selected different sentence when the pair was presented in reverse order, the others were consistent; those annotators selecting a different sentence when presented in different order were different each time. The notable exception to this is the pair of "Zlobí se a říká: Ten praví to a ten ono" and "Ten říká to a ten to, zlobí se" (*One says this, another that, he frowns*).[2] Five annotators selected the exact opposite when the pair was presented to them for the second time in reversed order. The rankings were also equally distributed which suggests that those sentences were the same in terms of complexity for the annotators.

An important outcome is that the annotators definitely prefer active voice to passive voice and it seems that this criterion outweights any other. In case one sentence was in active voice and the other one in passive voice, most of the annotators selected the sentence in active voice (I will call it 'active sentence' for now) to be the simpler one. This was true both when the active sentence was first of the pair and when it was second, both when the active sentence was shorter of the sentences in the pair and when it was longer, both when the first

[2]Quotes are omitted for better readability when already quoted.

sentence contained more common vocabulary than the other sentence and when it contained less common vocabulary.

This outcome already implies that the annotators did not consider sentence length to be the most important criterion. It actually seems that sentence length was not more important than vocabulary commonness – longer sentences with more common vocabulary tended to be preferred by the annotators.

However, in case both sentences were in active voice and had a comparable vocabulary in terms of its commonness, the annotators did sometimes prefer the shorter sentence. The annotators also tended to prefer the first sentence in the pair. In 47 pairs, the number of votes for the first sentence being strictly better was strictly greater than the number of votes for the second sentence being strictly better. In was the other way round in 31 pairs. However, a more careful analysis would be needed to check whether first sentences did not happen to share some features which could really make them simpler.

Annotators' feedback

Some of the annotators shared their impressions and thoughs and I would like to summarize them here.

As mentioned before, some of the annotators felt that sometimes the two sentences had a slightly different meaning. They felt confused because of that and found it more difficult to compare such sentences because they considered selecting the sentence with the intended meaning more important. They also experience an incapability to really understand the sentence because it seemed meaningless to them unless they thought out a context in which the sentence could appear.

The annotators also reported that they experienced difficulties comparing longer sentences simply because they were not able to keep both sentences in memory. Imagining the needs of the target group was also an issue for some annotators. On the other hand, some other annotators described some adaptation to the thinking beyond the presented sentences, developing more understanding for the formulations of their 'partner' and incorporating this understanding into their future evaluation.

Some annotators confessed that after ranking some pairs, their attention had fallen greatly and they had found themselves sticking to a simple strategy like selecting the shorter sentence or selecting the first sentence. One annotator also admitted that sometimes he/she had only read the first sentence and if it had been understandable and acceptably well written, he/she had selected this sentence. It is necessary to add that he/she did not do this intentionally, he/she only realized he/she had been doing it sometimes.

Even though this should be rather based on data, I would like to quote one annotator who has suggested that afirmative sentences could be easier to understand than negative sentences and that indicative mode is easier to understand that conditional mode.

We also speculated with some annotators that the decision which sentence is simpler could be affected by the previous sentence pair. However, I have not tried to test this in any way.

4.3 Manual simplification

The main purpose of this experiment was to get an idea about human simplification. The annotators were given a set of complex sentences and were asked to rewrite the sentences in a simpler way.

Three annotators took part in this experiment. Two of them are females, one is male. Two of them are university students, one of them works as a manager and lecturer (and has a university background). None of them has a formal education in teaching Czech but one of them communicates in simplified Czech with various people nearly on a daily basis, one of them does this occasionally.

The annotators were instructed to simplify each of the sentences. The resulting sentences should be easier to understand, the meaning should be preserved. The simplification should be general, usable for anyone who could benefit from it, though a young foreigner coming to study a Czech university was suggested as a representant of the target group. There was no restriction regarding the simplification (except for meaning preservation). Sentence splitting was explicitly allowed. Lexical substitution, eliminination of unimportant words, word re-ordering and explanatory notes were explicitly suggested as means of simplification. The annotators were allowed to suggest multiple simplifications per sentence.

The set of sentences contained 49 sentences selected or sampled from different corpora used in this thesis. Three of those sentences were selected from the blog corpus, the others were equally (± 1) sampled from PDT and the three subcorpora of CNC (syn2000, syn2005, syn2006pub).[3]

The sentences sampled from were sentences meeting at least one of several conditions, e. g. being of a great length, containing many subordinating conjunctions, containing big noun clusters, containing very uncommon words.

In most cases, the annotators suggested only one simplification, there are three sentences for which more than three simplifications had been suggested. In total, there were 150 suggested simplifications. I have analyzed the resulting sentences to identify the strategies and operations used (and not used) by the annotators. I describe them in the following text.

Complex rephrasing

In general, the annotators combined many different strategies to simplify the sentences. To demonstrate this, I will describe the exceptions.

There are 3 simplifications which only employ lexical simplification.

Two of those simplifications are for the complex sentence "Zatím se dramaticky zvýšila jeho entropie" (*So far, its entropy has increased dramatically*). Two of the suggested simplifications were "Zatím se velmi zvětšila jeho neuspořádanost" (*So far, its unorderliness has increased a lot*; using a different expression for the verb) and "Zatím se vysoce zvýšila jeho neuspořádanost" (*So far, its unorderliness has highly grown*).

The third simpification is for the complex sentence "Goldworm tak byl ušetřen žinantních formalit posledního sbohem" (verbatim *Goldworm thus was saved [from] awkward formalities of last farewell*; *Goldworm was thus saved from awkward formalities of the funeral*). The suggested simplification is "Goldworm tak

[3]Those corpora are described in more detail in section 5.1.

byl ušetřen nepříjemných formalit posledního rozloučení" (verbatim *Goldworm thus was saved [from] awkward formalities of last goodbye*; using also a more frequent word for *awkward*). It is worth noting that the other suggested simplifications for this sentence combined other strategies with substituting the euphemism with respective variants of "pohřeb" (*funeral*).

The two complex sentences are complex rather because of the vocabulary, their syntax is easy and they are short.

There is also 1 simplification which only employs de-passivization. The complex sentence is "Kdyby loď vybuchla, bylo by zabito tisíc námořníků" (*If the ship exploded, a thousand sailors would be killed*) and the suggested simplification is "Kdyby loď vybuchla, zemřelo by tisíc námořníků" (*If the ship exploded, a thousand sailors would die*). The other two simplifications take the de-passivization further and rephrase the sentence with ship explosion as the subject: "výbuch lodi by zabil tisíc námořníků" (*an explosion of the ship would kill a thousand sailors*).

There is no other simplification which would only employ one of the strategies described below. Contrarily, some simplifications employ strategies not generalized in this report.

Sentence splitting

Sentence splitting was used by the annotators in most cases. There were 32 sentences for which all annotators suggested splitting and 42 sentences for which at least one of the annotators suggested splitting; 113 proposed simplifications employed sentence splitting.

Most often, splitting the complex sentence into 2 shorter sentences was suggested; this happened in 76 simplifications. In 32 cases, the complex sentence was split into 3 sentences; split into 4 sentences (3 cases) and into 6 sentences has also occured.

Unsurprisingly, shorter sentences were usually left unsplit by the annotators while long sentences were split more often. The high rate of sentence splitting is supported by complex sentence selection, their average length was 26.75 words, 24 sentences were longer than 28 words.[4]

Sometimes the split was linked to a complex rephrase, sometimes it was rather a substitution of a comma with a dot (and possibbly some other alterations).

Consider for example this complex sentence: "Je možné, že by mohl vzniknout názor, že Šedé sestry mají pouze klášter na Lomečku, ale pozorní čtenáři ví, že čas od času přinášíme zprávu i o pražském klášteře Šedých sester v Bartolomějské ulici." (*It is possible that an idea would arise that Grey nuns* [a monastery order] *only have a nunnery at Lomeček but an attentive reader knows that sometimes we bring news also from Prague nunnery of Grey nuns in Bartoloměj street*).

All three annotators decided to split the sentence after mentioning the Lomeček location, putting the information about the second nunnery in another sentence. Two of them also made a separate sentence for the fact that the author sometimes brings news from the second nunnery, the third annotator completely cut this fact out.

[4]The average length of a Czech sentence depends on the kind of text and the source reporting the value but ranges from 14 to 20 words.

Another complex sentence is "Dostali jsme hlášení, že na silnici leží poražený chodec, když jsme ale dorazili na místo, zjistili jsme, že asi třicetiletý muž je postřelený." (*We got a report that on the road, there was a struck pedestrian [i. e. struck by a car], when we however got to the place, we found out that the approximately thirty-year-old man was shot.*)

All three annotators decided to split the sentence after mentioning the struck man. All three annotators also decided to only mention that the man was shot in the second sentence; as for his age, they either mentioned this in third sentence or did not mention it at all.

The annotators also sometimes decided to split the sentence at the position of a relative pronoun "který" (*which*). For instance, two annotators decided to split the complex sentence "Připravuje se i novela zákona o cenných papírech, která umožní postihovat černé obchodníky a chránit akcionáře před jejich spekulacemi." (*An amendment to the law on securities, which will allow to punish clandestine dealers and protect shareholders from their speculations, is being prepared*) into a sentence informing about the amendment being prepared and another sentence giving details on the purpose.

Explicit explanations

Some simplifications employed explicit explanations. By this, I mean that the original word was left in the sentence but it was accompanied with an explanation on its meaning, either in brackets or separated by commas.

The explicit explanation was used in 10 simplifications which is not very much compared to 150 simplifications suggested but it presents an alternative to leaving an information out or hoping that the reader would know the expression.

One of the complex sentences simplified in this way mentions plastic surgeons in a snippet "Někteří plastiční chirurgové zkoušejí nové techniky" (*Some plastic surgeons try new techniques*). Two of the annotators did not alter this part of the sentence, the third annotator inserted the explanation "lékaři zabývající se zlepšováním vzhledu" (*doctors dealing with appearence improving*).

The explanations were of varying specificity and exactness. Another sentence for which one of the annotators decided to use explanations was "Samice bývají špatní letci, a samička štětconoše dokonce nemá křídla vůbec a jen sedí na svém prázdném zámotku (který je přilepen na kmen) a čeká na samečka." (*Females use to fly badly, and orgyia female does not even have wings and only sits at its empty cocoon (which clinges to a trunk) and waits for the male*).

One of the annotators transformed this complex sentence into "Samice neumí moc dobře létat. Samice 'štětconoše' (druh hmyzu) ani nemá křídla – jen sedí na kmeni stromu (na svém zámotku – obalu kukly, přilepeném na kmen) a čeká na samce." (*Females use to fly badly. 'Orgyia' female (a kind of insect) does not even have wings – it only sits on a tree trunk (on its cocoon – a wrapper of the chrysalis, [which is] clinged to the trunk) and waits for the male*).

Just to complete the information, the other two annotators left the word orgyia in place, one of the annotators left also the word cocoon in place, the last one took it away at the loss of some information.

An explicit explanation was also used to explain a phrase. One of the complex sentences contained a quotation "Jsem rád, že už to máme z krku" (literally *I am*

happy that we have it off our neck already; *I am happy that it's over*) which was in-place explained by one of the annotators using explanation ("= že už zápas skončil", = *that the match has already ended*).

The last example I will give is a snippet of a complex sentence: "Radostné agape v klášteře ukončuje krásnou slavnost" (*The beautiful celebration in [the] monastery ends with a joyful agape*). One of the annotators used explicit explanation to simplify the word agape: "Radostné agape (společná hostina různě postavených lidí) v klášteře ukončuje krásnou slavnost" (*The beatiful celebration in [the] monastery ends with a joyful agape (a collective fest of people of different social status*).

The other two annotators did not insert any explicit explanations here but they substituted the word with another expression, once the word agape was substituted with the word "hostina" (fest), one with the expression "obřad lásky" (love ceremony).

Implicit explanations

I have mentioned explicit explanations before and now I would like to discuss implicit explanations more. The distinction between lexical substitution and implicit explanation is very hazy since substituting a word with its explanation is still a substitution and vice versa, substituting a word with its synonym is a kind of explanation.

Without claiming this is the good way to distinguish, I decided to treat any substitution which substitutes one word with more words as an explanation. Using this distinction, there are 32 simplifications which employ implicit explanation and 3 sentences for which it is employed in every simplification.

One of those three sentences nicely demonstrates the haziness of the disctinction. A snippet of this complex sentence is "Paní Jarmila s pláčem vzpomíná na Jánovu matku, která zemřela v koncentráku" (*Crying, Mrs Jarmila remembers Jan's mother who has died in concentration camp*, using an informal expression for the concentration camp).

All three annotators substituted the word "koncentrák" with "koncentrační tábor" (*concentration camp*), one of them was even more specific and used "nacistický koncentrační tábor" (*Nazi concentration camp*).

Another case concerned this snippet: "v sousedství sugestivní a stručné metaforiky takřka chandlerovského ražení" (*in the neighbourhood of forceful and consise metaphorics [of] nearly Chandler's ilk*, using a special adjective derived from Chandler's name).

One of the annotators rephrased this to "vedle stručných a výstižných metafor podobného typu, jako používal Chandler" (*next to consise and apt metaphors of the same type as Chandler used*). The other annotators completely omitted Chandler's name and only characterized the word, suggesting "[v porovnání s] výstižným, stručným projevem" (*[in comparison to] apt, consise discourse*) and "stručné a působivé popisy" (*consise and impressive descriptions*).

In some cases, the explanation could be treated as an attempt to translate the complex word to a simpler one. Those cases include substituting "entropie" (*entropy*) with "chaotické chování" (*chaotic behaviour*), substituting "konjunktura" (*boom*) with "ekonomický růst" (*economic growth*), substituting 'dát podnět' (*ini-*

tiate) with "vést [k něčemu]" (*lead [to something]*).

In some other cases, the main purpose of the explanation is to particularize the information. The previously mentioned substitution of "koncentrák" (*concentration camp*) with "nacistický koncentrační tábor" (*Nazi concentration camp*) would be an example of this. Another example is specifying "poražený chodec" (*struck pedestrian [i. e. by a car]*) as "člověk poražený autem" (*person struck by a car*).

There are also cases when the explanation introduces a piece of information not present in the text. For instance take the complex sentence "Šokové vlny, které vyvolala, byly pociťovány ještě dlouho po uzavření příměří mezi Dohodou a Německem v listopadu 1918 a její nezdařený výsledek připravil cestu pro druhý ještě strašlivější konflikt" (*The shockwaves it had caused were still felt long after the armistice between the Allies and Germany had been declared in November 1918 and its unsuccessful result had allowed for [the] second and even more terrible conflict*). Two of the annotators substituted the second conflict with "druhá světová válka" (*Word War II*).

A similar case is a sentence mentioning the word "točna" (ambigous word, meaning *pole*, *turntable*, *revolving stage*, in this context *revolving auditorium*) in the context of [Český] Krumlov. One annotator kept the word as is, one substituted it with "otáčivé divadlo" (*revolving theatre*), the last one substituted it with "otáčivé hlediště" (*revolving auditorium*).

A special case of implicit explanation is substituting metaphor with its meaning. Now there could be another discussion about what is a metaphor and what is not but the point is that the annotators tried to eliminate metaphors from the sentences.

I will recall a different snippet of an already mentioned complex sentence: "starší bratr zůstal navěky v Mauthausenu" (verbatim *older brother stayed forever in Mauthausen*). Even though the verb "zůstat" (*[to] stay*) has a similar meaning to English, it can be also used to express that someone died in the place. All three annotators substituted this verb with more explicit "zemřít" (*[to] die*).

Similarly, two annotators decided to substitute "poslední rozloučení" (verbatim *last adieu*) with "pohřeb" (*funeral*). All three annotators decided to substitute "být zcela na dně" (verbatim *[to] be totally at the bottom*) with various different expressions, "připadat si nejhůř" (*to feel the worst*) or "myslet si, že už nemůže dál" (*think that [he] cannot go further*).

I have identified seven complex sentences containing a metaphor (or something similar enough to metaphor) and the metaphor was explained in simplifications for all of them.

Information reduction

An obvious way to simplify a sentence is to reduce the information given in the sentence. Again there is a hazy distinction, this time between reducing information little enough / too much to preserve the meaning.

It is also arguable what should be considered information reduction as some words add very little to the meaning of the sentence.

For those reasons, the decision whether information reduction was epmloyed is highly subjective. Under this subjective decision, I have identified 24 simplifi-

cations which reduce the information given while preserving the meaning itself.

A perfect example of such reduction would a be previously quoted copmlex sentence: "Dostali jsme hlášení, že na silnici leží poražený chodec, když jsme ale dorazili na místo, zjistili jsme, že asi třicetiletý muž je postřelený." (*We got a report that on the road, there was a struck pedestrian [i. e. struck by a car], when we however got to the place, we found out that the approximately thirty-year-old man was shot.*)

One of the annotators simplified this as "Dostali jsme hlášení, že na silnici leží zraněný chodec. Když jsme dorazili na místo, zjistili jsme, že muž je postřelený." (*We got a report that on the road, there was a hurt pedestrian. When we got to the place, we found out that the man was shot*).

The information on the man's age is lost in the simplification but otherwise, the meaning is the same.

Another way to reduce the information while still preserving the meaning was to eliminate names and other particularities. For example, "chebské kasárny Ernsta Thälmanna" (*Ernst Thälmann barracks in Cheb*) were substituted with "chebské kasárny" (*barracks in Cheb*) by two annotators.

Other strategies

Even though it would take too much space to describe all strategies that have been used by the annotators, there are some more I would like to cover.

De-passivization, transforming passive voice into active voice, was used when simplifying one complex sentence – the only one with passive voice.

Subject insertion was also used. The beginning of the complex sentence "Šokové vlny, které vyvolala," (*Shockwaves which [it] had caused*) was transformed to "První světová válka vyvolávala šok" (*World War I was causing shock*)[5] by one of the annotators.

It is worth noting that the subject had to be guessed from the context, the war is not mentioned explicitly in the sentence.

Another annotator did not insert the original subject but rephrased the sentence so that the subject changed, the simplification begins with "Všichni cítili šokové vlny" (*Everyone felt shockwaves*).

Pronoun disambiguation (co-reference resolution with pronouns) was also used in some cases. One complex sentence contains "i když mě třeba mrzí, že v Krumlově hraje naše divadlo, představení mají velký ohlas, ale jen málokdo přitom ví, že hrají českobudějovičtí herci" (*even though I am sorry that [it is] our theater ensemble [who] is acting, the shows have good responses but only few people know that [it is] actors from České Budějovice [who] are acting.*). Our theater ensemble was replaced by "herci z Českých Budějovic" (*actors from České Budějovice*) by two annotators.

Adjective transformation into another phrase also occured a few times. There are two examples of this in the complex sentence "Volby do sedmičlenného zastupitelstva vypuknou předposlední lednovou sobotu." (*[The] elections to seven-member local government will start on last but one January Saturday*).

All three annotators decided to eliminate the adjective "sedmičlenný" (*seven-member*) and mention the number of representatives in another sentence. One

[5]The difference in verb aspect is caused by the rest of the sentence.

of the annotators also decide to substitute "lednovou sobotu" (*January Saturday*) with "sobotu v lednu" (*Saturday in January*).

4.3.1 Feedback

The annotators' feedback suggested that the task turned out to be much more difficult than they had expected. This can be partially explained be high-quality simplifications provided by them, partially it demonstrates the need of some automation.

The feedback also suggests that after simplifying a few sentences, the annotator tends to adapt to the sentence complexity and begin to alter the sentences less, which possibly results in more complex resulting sentences.

A manual analysis of the simplifications also shows that the annotators have some tendencies: one annotator tends to use more explanations, another one tends to produce sentences shorter than the complex ones, ...

5. Lexical simplification

I have introduced lexical inplification in section 3.3 and now I will describe it in more detail. I omit substitution selection in this description since I have not experimented with explicit selection. This decision was made partially due to the fact that I am not aware of any general-domain word sense disambiguation tool which would perform well for Czech, partially because ranking could still capture the sense adequacy in the score.

Even though this might contradict the idea of simplification, substituting a complex word with itself is permitted in lexical simplification (at least as I describe it). This especially allows to partially recover from incorrect complex word identification.

I also describe the corpora I have used in the course of working on the thesis, mostly to allow for linear reading as the corpora are referred to in this chapter.

5.1 Corpora used in the thesis

I would like to describe the corpora used in the thesis. I have used several corpora, some of them were used only to learn some statistics, some were used also for sentence sampling and especially for exmerimental simplification runs. I have also created a small corpus of complex sentences.

- **PDT**, Prague Dependency Treebank is a manually annotated treebank. Annotated texts are news articles, the text is annotated at four different layers including morphological and analytical (syntactic) layer. This treebank contains 453, 574 tokens in 49, 431 sentences, although not all tokens are annotated at all layers. Data can be obtained online, see Bejček et al. (2013).

- **syn2000**, **syn2005** and **syn2006pub** are subcorpora of the Czech National Corpus. Texts in all three subcorpora are of varied genres. Data can be queried through an online interface.[1] Texts are lemmatized and annotated, see Čermák et al. (a), Čermák et al. (b) and Čermák et al. (c). However, I have worked with raw texts provided by my supervisor and had the corpora tagged by MorphoDiTa. The corpora contain respectively 120, 977, 151 tokens, 122, 891, 595 tokens and 362, 988, 882 tokens.

- **corp** is the name I use to refer to PDT, syn2000, syn2005 and syn2006pub corpora altogether. In case I wanted to learn some probabilities or frequencies or wanted to get general substitution statistics, I usually used those four corpora.

- **blog** is a simple corpus of the posts[2] I publish at my website. The main reason for using it was the familiarity with the texts which proved useful during evaluation of experimental substitution runs on this corpus. This corpus does contain some spelling and stylistic errors.

[1] http://www.korpus.cz/

[2] For completeness, calling the posts *tweets* would be much more accurate, especially because of their length.

- **cswiki** is just the dump of Czech Wikipedia,[3] processed by the wp2txt tool[4] and tagged by MorphoDiTa. This corpus has been used mostly to learn some statistics.

- **complex** is included here for completeness. This is the corpus I have created to test and evaluate the simplification process and is described in more detail in the introductory part of chapter 6.

5.2 Complex word identification

Complex word identification is a step of lexical simplification. In this step, complex words are identified so that substitutions can be generated for them during later steps.

There is no easy answer to the question what a complex word is. A complex word is any word which is difficult to read or understand, probably a long word, an uncommon word or a word made by an uncommon combination of characters.

I have experimented with complex word identification based on frequency, word length and character probabilities.

All words

A simple baseline solution is to identify all words as complex words. It is virtually impossible to substitute each word of a sentence with another one and stll obtain a meaningful sentence. However, since there will be no substitutions available for some of the words and many words will probably score as already-the-best in the ranking step, such a simple approach could still give acceptable results. The real performance will of course highly depend on the numbers of complex and non-complex words in the text.

Content words

Another simple baseline is to identify not all words but a well-defined subset of them. I have experimented with treating all nouns, adjectives, verbs and adverbs as complex words.

Lemma frequency

Complex word identification based on lemma frequency assumes that uncommon words are complex; it disregards other statistics such as word length.

I based lemma frequency on stripped lemmata as this is what I obtain from individual synonym dictionaries and it is generally possible to determine for synonyms obtained by generation from embeddings.

Reference frequencies are computed using the 'corp' set, that is the PDT, syn2000, syn2005 and syn2006pub corpora.

[3] http://dumps.wikimedia.org/cswiki/20170120/cswiki-20170120-pages-articles.xml.bz2, more recent versions are available at https://dumps.wikimedia.org/cswiki/

[4] https://github.com/yohasebe/wp2txt

A word is considered complex if and only if its frequency does not meet a required treshold, which was set empirically.

For the sake of completeness, I should note that word sense is not disambiguated in any way, the frequency used can therefore be higher than the actual frequency of the given complex word in the given sense.

Character count

Character count is a simple statistic. The basic assumption is that the longer a word is, the more complex it is. Despite the simplicity of this definition, there are things to be decided – mostly which key should have the characters counted, form, stripped lemma or lemma.

I decided to try counting characters of both the original form and of its stripped lemma. I did not however consider lemmata. Since lemmata contain additional information, their length does not have to correspond to the length of the word as perceived by the reader. While taking hints on form derivations into consideration could be an interesting experiment (since the presence of such hint might be a sign of complicated morphology and thus reader's difficulties with parsing the word), auxiliary strings describing the sense do not really relate to the complexity of the word in question.

All characters are treated equally, though some weights could possibly be introduced to better reflect the higher or lower commonness of some characters. They could be based not only on character commonness but also on some other features like penalizing characters with diacritics.

Syllable count

Syllable count is another statistics which assumes that the complexity is mostly driven by word length. The length is however measured in syllables instead of characters.

It is clear that syllable count can be lower than character count as several characters usually make up a single syllable. In Czech, syllable count could also be lower under some other circumstances: there are non-syllabic words "k", "s", "z" and "v" (prepositions *to/towards*, *with*, *from*, *in*). Nevertheless, this is not an issue for complex word identification as their length is small anyway.

Identically to character count, all syllables are treated equally. Again, some weights could possibly be introduced to add some distinction. For example, syllables constituted by syllabic consonants like "pl-ný" (*full*) or "prst" (*finger*) could be treated as more difficult than syllables constituted by a vowel like "a-hoj" (*hello*).

Unigram character probability

Unigram probability of a word is the product of unigram probabilities of characters that make up this word, i. e. $P(dog) = P(d)P(o)P(g)$. This probability assumes that uncommon characters make up complex words. Such probability also implicitly penalizes long words as their probability is naturally lower than probability of short words.

Unigram counts I use for probability computation are based on the 'corp' set. All words tagged as punctuation and words not containing any character of the Czech alphabet were eliminated. All other words were lowercased and then used to compute the unigram probabilities.

Since no other words were eliminated, it is likely that some non-Czech words were used. However, there should be little enough of them for the results to still be reliable.

Though this is rather an implementation issue, it is good to keep underflowing in mind. Character probabilities are low and they get even lower upon multiplication.

Bigram character probability

Bigram probability of a word is the product of bigrams present in the word, i. e. P(dog) = P(d)P(do)P(og)P(g). The assumption for bigram probability is very similar to unigram probability: uncommon character combinations make a complex word.

Bigram counts I use are again based on the 'corp' set, the same methodology was used to obtain them as to obtain unigram counts. Word boundary is treated as a character in this case.

5.3 Substitution generation

Substitution generation is a step of lexical simplification. Its input is a complex word (possibly with its context) and its output is a set of possible substitutions for word.

The substitutions should have a meaning similar to the original complex word. It could happen that some of the substitutions do have a similar meaning but are not suitable for the context for some reasons; still the substitutions can be generated at this step (and then be eliminated at later steps, preferrably during substitution selection).

I have experimented with obtaining synonyms and near-synonyms as substitutions via two groups of ways. The first is dictionary-based, the second is based on word embeddings.

5.3.1 Dictionary-based generation

Dictionary-based substitution generation is very easy in principle. The complex word is searched for in a synonym dictionary and if found, the substitutions are taken directly from the dictionary.

If the dictionary distinguishes *synsets* (a set of synonyms defined by its meaning) and it is possible to map the sense of the complex word to those synsets, only substitutions from relevant synsets should be taken; substitutions from all synsets of the complex word should be considered otherwise.

I have experimented with four different dictionaries which I now describe in more detail. The description of individual resources is followed by tables giving exact counts of words and synsets in all of them as well as by some additional comments on comparison of those resources.

Czech WordNet (PDT adaptation)

I have used Czech WordNet 1.9 by Pala et al. (2011). As the name suggests, Czech WordNet is a Czech analogue of the original WordNet (Miller (1995)).

The original Czech WordNet was developed by the Centre of Natural Language Processing at the Faculty of Informatics, Masaryk University, the version used was slightly modified by the Institute for Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University.

WordNet has already been used for substitution generation in English, see for example Paetzold and Specia (2015). Czech WordNet is however much smaller (according to the WordNet website,[5], there are 117,000 synsets in the WordNet; Czech WordNet consists of only 23,904 synsets), which is likely to have negative impact on performance.

Czech WordNet includes nouns, verbs, adjectives and adverbs; adverbs are the least covered yet included part of speech. Synsets are distinguished in WordNet, some hypernym and hyponym relations are also described.

By manual inspection, the words covered by Czech WordNet are very unbalanced with respect to domain. There is also a great amount of synsets containing only one word, which are useful to capture the meaning but not for substitution generation.

Synonym dictionary A

Another source of synonyms was a synonym dictionary. I have little knowledge about this dictionary's background. It was created several years ago and used in a commercial company. I have been given it in private and I am not allowed to share it, but I have used it in experiments.

Even though the dictionary cannot be shared, it can demonstrate the effect of using a dictionary as well as the improvement of the results by adding more resources.

This dictionary covers most of Czech parts of speech: nouns, adjectives, pronouns, numerals, verbs, adjectives and conjunctions. The synsets are not distinguished explicitly but words are given in groups corresponding to synsets; no meaning explanation or synset identification is given.

By sampling and manual inspection, the dictionary is of general domain, including a lot of loanwords, with little technical terms.

Synonym dictionary B

Another synonym dictionary I have used is the web version of Slovník synonym[6]. According to the information on the website, its content is based on the printed edition and further extended by the website users. All synonyms added by users are marked; by simple inspection of randomly sampled user suggestions, the user-suggested synonyms are all fine-quality and really synonymous.

This dictionary covers most of Czech parts of speech. Synonyms are given as lists of synonyms for each of the entries, the synonyms are sometimes grouped by meaning/senses but synsets are not explicitly given. The words covered in this

[5] https://wordnet.princeton.edu/

[6] http://www.slovnik-synonym.cz/

	dictionary A	dictionary B	wordnet	babelnet
# lemmata	21391	26673	25295	179757
# synsets	18211	15318	23094	102307
avg syn/word	3.14	1.85	1.35	2.39
med syn/word	2	1	1	2
avg word/syn	3.68	3.21	1.48	4.20
med word/syn	3	3	1	3
avg cands/word	8.76	5.54	2.39	7.30
med cands/word	6	4	2	5
# synsets of size 1	0	0	16074	21884
# with 1 synset	8477	15352	19985	88041
# without candidates	0	0	10982	10349
# multiword expressions	0	12438	7680	50248

Table 5.1: Word, synset and candidate statistics of synonym sources

dictionary are rather of general-domain, the dictionary covers also some dialectical and other non-standard words.

BabelNet

BabelNet (Navigli and Ponzetto (2012)) is a collection of national wordnets, Wikipedia entries, Wiktionary, Wikidata, ImageNet, FrameNet and other resources. It covers 271 languages including Czech.

BabelNet covers all parts of speech; unlike other resources, it also covers some proper names and named entities in general. Synsets are distinguished.

By manual inspection of a few entries, the quality is not very good and there are lots of artifacts among the synonyms. However, a lot of queries result in a set of synonyms, including queries for very rare or technical terms.

Dictionary statistics and comparison

A large portion of lemmata present in all of the sources is one-synset-only (see table 5.1), this portion being biggest in WordNet. Generating substitutions, or more exactly selecting substitutions, for such lemmata should be easier as there is no need to disambiguate word sense. However, this is rather a theory which does not hold in reality. The fact that a lemma is only present in one synset does not generally mean that it has only one sense, it rather means that other senses are not described in the given source.

To give some examples, the word "doplnit" is present in only one synset in Czech WordNet. It has one synonym, "dolít" (*to refill [by pouring in more, for liquids]*), no definition is given. The meaning could really be *to refill* (and maybe the word "dosypat" (*to refill [for solid stuff]*) would also be a synonym) but it can also be *to fill in* (e. g. fill in gaps in sentences, a common task in language tests) or *to add more information.*

The word "test" has one synonym, "písemka" (*exam paper*). The meaning, however, is not limited to written tests, physical test could also be "test".[7]

[7]On the other hand, it would probably not be used in reference to oral exam.

The word "znamenat" also has one synonym and no definition in Czech WordNet. The synonym given is "zaznamenat" (*to take [sth.] down*) which suggests that the captured sense is *to write down*, *to note down*. The word has another meaning, *to mean*. Given my experience as native speaker and results of searching the CNC[8], the other sense is much more common. This implies that incorrect substitution set would be generated often.

Another one-synset-only word is "pustit" which also has one synonym, "spustit" (*to launch, to lower, to start*). As English translations suggest, there are more meanings. Word uses include "pustit film" (*play the movie*), "pustit někoho dovnitř" (*let somebody in*), "pustit barvu" (*color bleed*) or "pustit něco na zem" (*drop something to the ground*).

The last example is the word "adresář" which has the word "seznam" (*list*) as its synonym. While the word can be used to refer to a list of [postal] adresses, a very common meaning is *a directory* in the computer sense.

To give some examples using also source other than WordNet, the word "obsazovat" has only one synset in dictionary A, its synonyms suggest the meaning *to occupy*. The word can however also be used as synonym to "osazovat" (which could translate as *to assing a person [e. g. to operate a station or a cashier's desk]*). On the other hand, the word "obsazovat" has imperfective aspect and its perfective variant "obsadit" is synonymous to the perfective variant of "osazovat" according to dictionary A.

Another example could be the word "znaménko", which has *[artificial] mark* as its suggested meaning in dictionary A, though it could also mean *birthmark*. Dictionary B gives a lot of informal variants but it only suggests the word "flek" to mean *a stain* though it could also mean *a job*.

On contrary, some words are given as synonyms for the very same synset, though the senses could possibly be distinguished. An example of this is the word "potkat". All other words of the same synset are "vidět" (*to see*), "setkat se" (*to meet*), "potkat se" (*to meet, to encounter*) and "uvidět" (*to see, to spot*). While the word "potkat" can be used for both encountering something by seeing it and for meeting someone, a distinction between those senses would be desirable when generating synonyms.

It is worth noting that except for dictionary A, the sources treat reflexive verbs, consisting of the verb itself and a reflexive pronoun, as special entries. In this thesis, only the verb itself is searched for during substitution generation which leads to poor results for those verbs.

Some examples of words for which irrelevant substitutions are generated because of reflexive pronoun neglect are "flákat" (*to hit, to do something sloppily*) instead of "flákat se" (*to slack off*), "představit" (*to introduce*) instead of "představit si" (*to imagine*) or "naučit" (*to teach*) instead of "naučit se" (*to learn*).

Another reason for a word to have only one synset, especially in WordNet, is being special (or rather technical or obscure) enough not only to have just one sense, but to have no candidate substitutions at all. This is not the case of dictionary A and dictionary B, as synsets are given as groups of at least two words in

[8]By querying for the lemma "zaznamenat", I was not able to find a single instance of the word having the sense of noting down in a returned sample. After querying for `[lemma="znamenat"][word="do"]`, which restricts the results to the word followed by the preposition "do" (*in, into, untill*), I identified two instances of 27 returned to have this sense.

	dictionary A	dictionary B	wordnet	babelnet
dictionary A	-	11537	12534	11850
dictionary B	16819	-	17953	19001
wordnet	16438	16575	-	13550
babelnet	170216	172085	168012	-

Table 5.2: Additional lemmata obtained when combining synonym sources

those sources, but it happens in both WordNet and BabelNet. Examples of such words include "polypovití" (*hydrozoa*), "normovaný polynom" (*normalized polynomial*) or "polymixin" (*polymixin*). My further analysis has shown that 10,050 words with just one synset do not have candidate substitutions at the same time in Czech WordNet.

It it also worth noting that both WordNet and BabelNet contain synsets consisting of only one lemma. Of course they do, there would be no words without candidates otherwise, on the other hand, the number of such synsets is concerningly big. Moreover, being in a synset without other lemmata does not necessarily mean that no usable substitutions exist, it could rather mean that the substitutions are not included in the source in question.

The word "skončit" (*to end*, *to finish*) could be an example of this. It is present in three synsets in Czech WordNet, each time with no synonyms. However, word "ukončit" (*to end*, *to terminate*) or "přestat" (*to stop*, *to quit*) could be used as synonyms in some cases. Similarly, there is no synonym for the word "bumbat" (*to drink*, used in motherese) but the word "pít" (*to drink*) could be used, the word "plat" (*pay*) could be used for "žold" (*soldier's pay*) etc.

Multiword expressions, here defined as any literal (lemma) containing a space (or an underscore for BabelNet), seem to be a great issue. Except for the dictionary A, multiword expressions are a big portion of all lemmata. Since those are not handled well in this thesis (as they are neither correctly ranked nor is the correct form derived from them), they are virtually thrown away, thus reducing the available sources by a quarter to almost a half. Such reduction also influences the contribution an additional source makes – if a new source brings 16k lemmata, of which 12k are multiword expresssions, only 4k are really used on top of the previous source. Better multiword expression handling is therefore recommended as future work.

Even though the first three sources are of comparable sizes (see table 5.1), they contain different lemmata which means that combining them does increase the number of words for which substitutions can be generated (see table 5.2). For each source pair, less than half of the lemmata present in the first source is also present in the other one. This is true also for pairing with BabelNet, which is of much bigger size in terms of different lemmata contained.

Using additional source not only helps to cover more lemmata but also to generate more substitutions for the lemmata already present in the original source (see tables 5.3, 5.4, 5.5, 5.6).

For example, the table 5.4 shows that when taking all lemmata from dictionary A and searching them in dictionary B, substitution sets for 264 lemmata are strict superset of substitution set which would be obtained from dictionary A.

	dictionary A	dictionary B	wordnet	babelnet
dictionary A	-	11610	16327	13393
dictionary B	17083	-	20898	20149
wordnet	16615	16575	-	14133
babelnet	170562	172092	173340	-

Table 5.3: Counts of lemmata resulting in strict subset of substitutions when using different synonym source

	dictionary A	dictionary B	wordnet	babelnet
dictionary A	-	264	177	346
dictionary B	73	-	0	7
wordnet	3793	2945	-	5328
babelnet	1543	1148	583	-

Table 5.4: Counts of lemmata resulting in strict superset of substitutions when using different synonym source

	dictionary A	dictionary B	wordnet	babelnet
dictionary A	-	16	154	25
dictionary B	16	-	0	0
wordnet	154	0	-	465
babelnet	25	0	465	-

Table 5.5: Counts of lemmata resulting in the same set of substitutions when using different synonym source

	dictionary A	dictionary B	wordnet	babelnet
dictionary A	-	9501	4733	7627
dictionary B	9501	-	5775	6517
wordnet	4733	5775	-	5369
babelnet	7627	6517	5369	-

Table 5.6: Counts of lemmata resulting in incomparable set of substitutions when using different synonym source

Numbers of lemmata resulting in strict subsets (table 5.3) are very high when compared to numbers of lemmata resulting in strict supersets. This can be partially explained by different lemma coverage, new substitution set for a lemma not present in the new source cannot be greater than substitution obtained from the original source.

However, numbers of lemmata resulting in incomparable substitution sets are also high (table 5.6), especially with respect to the numbers of non-common lemmata. Combining sources therefore leads to bigger substitution sets and is more likely to contain the ideal substitution.

Unused: Wikipedia redirects

My supervisor suggested that Wikipedia redirects could also be a good source of substitutions as their purpose is to caputre all the ways a user might ask for relevant content and bring the user to the relevant page. I agreed this was an interesting idea and decided to explore those redirects further.

I have downloaded Czech Wikipedia redirect dump,[9] extracted the redirects themselves and inspected them by just browsing through the file and by simple sampling.

A lot of redirects is not usable as they map terms to more general entries, like mapping "1. Newtonův pohybový zákon" (*Newton's first law*) to "Newtonovy pohybové zákony" (*Newton's laws of motion*) or mapping "Tradiční znaky" (*Traditional characters*) to "Čínské znaky" (*Chinese characters*). Mapping terms to pages containing their definition also occurs, like mapping "Mezinárodní kód země" (*International country code*) to "ISO 3166-1".

Mapping terms to more canonical ones is also common, e. g. mapping "18. května" to "18. květen" (*18 May*, Czech variants differ in case) or mapping "Motokáry" (*Go-karts*) to "Motokára" (*Go-kart*, plural to singular).

Those kinds of redirects make the source hard to use. However, by my inspection, I have found several interesting mappings, e. g. "psalterium" to "žaltář" (*psalter*), "chajda" to "chata" (*cottage*), "kmitočet" to "frekvence" (*frequency*, though the first term is limited to frequency as physical quantity while the second could also be used e. g. for the frequency of lemmata in a corpus) or "chladnička" to "lednička" (*fridge*).

Some redirects also caputre orthographic differencies (e. g. mapping "theologie" to "teologie" (*theology*) or "televise" to "televize" (*television*)) or different word ordering (e. g. mapping "Mytologie Inuitů" (*Mythology of Inuits*) to "Inuitská mytologie" (*Inutit mythology*). Those could also make valuable substitutions.

Because of the source size (there are 234, 152 redirects in the dump) and hard-to-generalize inapplicable mappings, I have disregarded this source. However, I wish not to discourage any future efforts to make use of this source as it seems that valuable substitutions could be extracted from it.

[9] http://dumps.wikimedia.org/cswiki/20170201/cswiki-20170201-redirect.sql.gz, more recent versions are available at https://dumps.wikimedia.org/cswiki/

5.3.2 Embedding-based generation

When combining natural language processing and machine learning, the traditional approach to word representation is one-hot encoding. For vocabulary of size V, words are represented are V-dimensional vectors with zeros at $V - 1$ positions and number one at one position, matching the word's position in the vocabulary.

Word embeddings are real-valued vectors of much lower dimension (while V could be up to $50,000$, embedding dimension does not usually exceed $1,000$) which still uniquely encode the given word.

The paper of Mikolov et al. (2013a) has caused a great increase in the use of embeddings, probably because of the speed at which embeddings can be trained. It introduced two neural network architectures usable for efficient training.

The paper also noticed that those embeddings seem to capture also many semantic properties of the represented words. A very common example is the possibility to perform a kind of algebraic operations on the embeddings, the most famous probably being $emb(\text{king}) - emb(\text{man}) + emb(\text{woman}) \simeq emb(\text{queen})$. Mikolov et al. (2013b) give even more abstract examples like $emb(\text{Vietnam}) + emb(\text{capital}) \simeq emb(\text{Hanoi})$. Embeddings of Czech words have already been shown to also partially capture these properties by Svoboda and Brychcín (2016).

The task on which embeddings are trained is predicting words based on context (either predicting a word given the surrounding words or predicting surrounding words given the word in the middle). This implies that the property which embeddings capture the best is similarity in terms of context use, the same similarity which is sought by training word classes. The semantic similarity is probably only a consequence of the contextual similarity.

Words having similar embeddings therefore need not have similar meaning, embeddings of antonyms are often very similar to each other. The other way round, words of similar meaning need not have similar embeddings, especially if they are used in slightly different contexts.

Nevertheless, similar words tend to occur in similar contexts and thus have similar vectors. For this reason, I have experimented with substitution generation using embeddings. The basic idea is that embeddings most similar to the embedding of the complex word belong to its synonyms. I have also discovered that a similar approach has already been attemped by Paetzold and Specia (2015).

Models

I used two corpora as the source text on which training was performed. The first one was the 'corp' corpus, the second one was the 'cswiki' corpus.

For each of those two corpora, I trained models using forms, stripped lemmata and lemmata. For each corpus-key combination, I trained three models: a model trained on the text as it is, a model where only content words are left and all other words are eliminated, and a model where content words are left in place and other words are substituted with an auxiliary symbol.

I use Word2Vec implementation from the gensim tool[10] (version 0.13.4) to train the models. I used default training parameters.

[10] https://radimrehurek.com/gensim/models/word2vec.html

Substitution generation

During substitution generation, the complex word is searched for in the model, and if found, 5 most similar embeddings are considered to be synonymous.

The number 5 was chosen arbitrarily and possibly is not the best threshold. Some more experiments to test the effect of this threshold would be recommended.

It is also arguable whether the strategy of selecting N most similar embeddings is the best. Selecting all embeddings which have similarity over an established treshold or selecting embeddings until a drop in similarity occurs could also be viable and would deserve more extensive testing.

Since the training task is word prediction based on context, vector similarities really relate to context similarity more than to synonymity. In particular, this implies that a great portion of suggested substitutions are actually antonyms.

This has already been noticed by some authors and some solutions has been proposed to adapt the models for synonymity, in particular I would like to mention Mrkšić et al. (2016) and Faruqui et al. (2014). I would like to test the effect of model adaptation in a future work.

5.4 Substitution ranking

Substitution ranking is the step of lexical simplification when all remaining substitutions are ranked and the best substitution is selected. The ranking should reflect the substitution simplificity, which also covers the suitability for the given context.

It seems reasonable to suppose that similar strategies would produce good results for complex word identification and substitution ranking as both steps deal with complexity/simplicity ranking and overall suitability for the given text.

For this reason, I have used most of complex word identification strategies also to rank substitutions, namely lemma frequency, character count, syllable count, unigram character probability and bigram character probability. The strategies of identifying all words or content words are not applicable to substitution ranking.

In addition to those strategies, I have tried using Mistrík formula for substitution ranking. The formula is described in 3.2.2 but to recall, it has the following form:

$$R = 50 - S \cdot V \cdot \frac{L}{N}$$

The variables represent the number of syllables per word (S), the number of words per sentence (V), the number of distinct words (L) and the total number of words (N).

However, this strategy is likely to produce the same results as syllable count since only two features of a sentence can be changed by word substitution. Those two are the number of syllables per word (which is measured by the syllable count strategy too) and the number of distinct words (which is unlikely to change).

6. Results

I have implemented and evaluated several methods for complex word identification, substitution generation and substitution ranking. Here I would like to present the results.

The evaluation has been done on a special small corpus, the 'complex' corpus. This corpus is well reusable for evaluation of complex word identification and partially reusable for evaluation of the other two steps.

The corpus is divided into three sets, it contains 154 sentences equally divided between the three files (52 + 51 + 51). Sentences were preselected automatically using sentence length constraints (the sentences should not be too long because longer sentences are more demanding for the annotators judging them and therefore allow only smaller amount of evaluation iterations) and word frequency heuristics (sentences containing infrequent words are likely to contain a complex word).

After preselection, I chose the final 154 sentences which contained at least one word which I considered complex. Those sentences were divided into the files and made up the corpus.

Complex words were then annotated in the corpus and those annotations were used to evaluate complex words identification. Substitution generation was evaluated by generating and checking substitutions for all complex words as identified by the annotators. Substitution ranking was then evaluated by checking the resulting sentences when using approved substitutions from substitution generation evaluation and the ranking strategy in question.

It is worth mentioning that using substitutions suggested by humans (in contrast to using substitutions piped from the previous step) for evaluation of substitution ranking could also bring important results.

6.1 Complex word identification

For complex word identification, I marked words which I considered complex. After that, I tasked three annotators to check the annotations and possibly add or remove some complex words.

The annotations are therefore not truly independent, it is reasonable to suppose that should the annotators mark complex words from scratch, their annotations would slightly differ from the annotations they really delivered.

Nevertheless, I believe that the annotations are of quality good enough to be used in such evaluation. In total, 193 words were annotated as complex by all of the annotators, 217 are annotated as complex by a majority and 251 are marked as complex by at least one annotator. Complex word identification strategies were primarily evaluated against the words marked as complex by at least one annotator but results against the other two variants are also included.

It is however worth remarking that because of sentence preselection, there is a bias towards identification based on word frequency.

For the sake of completeness, I would like to mention that only single words were allowed to be annotated as complex. I have discussed this a little in section 3.3, too. While substituting only single words and leaving multiword expressions

in place makes a lot of sense, multiword expressions can be complex and thus deserve a substitution.

The annotators followed the instructions to only mark single words but suggested also some multiword expressions in their comments. Namely, they suggested to treat "chai latté" (*chai latte*), "s odřenýma ušima" (literally *with scraped ears*, *by the skin of one's teeth*), "táborový řečník" (literally *camp talker*, *platform orator*), 'mše svatá" (*Holy Mass*) and "černá magie" (*black magic*) as complex.

Discussion

The performances of complex word identification strategies are given in table 6.3. I should remind that those values are obtained by evaluating against words marked as complex by at least one annotator. Precision, recall and F-measure are defined as usual, i. e. precision is $P = C/M$, recall is $R = M/G$, where C is the number of words that were correctly marked as complex, M is the number of words that were marked as complex and G is the number of words marked as complex by the annotators. F-measure is then $F = 2 \cdot P \cdot R/(P + R)$.

F-measure as defined treats precision and recall as equally important, thought it is arguable whether this is the case. Recall could be considered more important because while simplifying a non-complex word is not likely to complexify the text, leaving a complex word in the text in a way hurts simplicity.

	precision	recall	F-measure
train			
all	0.17	1.00	0.29
content	0.26	0.98	0.41
frequency	0.93	0.70	0.80
character length (form)	0.39	0.88	0.54
character length (stripped lemma)	0.39	0.82	0.53
syllable length (form)	0.42	0.79	0.55
syllable length (stripped lemma)	0.44	0.72	0.55
unigram character probability	0.35	0.83	0.49
bigram character probability	0.27	0.90	0.41
test			
all	0.18	1.00	0.30
content	0.29	0.99	0.45
frequency	0.96	0.66	0.78
character length (form)	0.41	0.90	0.57
character length (stripped lemma)	0.44	0.86	0.58
syllable length (form)	0.46	0.85	0.60
syllable length (stripped lemma)	0.50	0.73	0.60
unigram character probability	0.36	0.83	0.50
bigram character probability	0.26	0.87	0.41

Table 6.1: Performances of complex word identification strategies (evaluated against words marked as complex by at least one annotator

Nevertheless, the frequency-based strategy completely outperforms any other strategy. I should however remind once more that the methodology of sentence preselection creates a bias towards this strategy. Even though the result of frequency-based identification is definitely encouraging, it should not be overestimated unless verified on data that do not suffer from such a bias.

It is also notable that the frequency-based strategy is the only one that has precision greater than recall (and again, recall could be considered more important than precision). This could be due to some correlations between frequency and complexity as well as due to parameter tuning. After an arbitrary threshold was set, frequency-based strategy had the precision of almost 1 and the precision was reduced by treshold shift. The threshold for other strategies, on the other hand, resulted in the recall of 1 (or almost 1) which was later reduced by threshold shift.

Another notable outcome is the recall of the content strategy. This strategy, reporting any content word as complex, achieved the recall of 0.98 on the training set and 0.99 on the testing set. In a way, this is expactable as usually, there are for example much more nouns than prepositions in a language, and therefore much more complex nouns than complex prepositions. On the other hand, this result demonstrates that omitting other parts of speech or treating them with just a small set of rules would be justifiable.

Evaluation against words marked as complex by a majority of the annotators

	precision	recall	F-measure
train			
all	0.15	1.00	0.26
content	0.24	0.97	0.38
frequency	0.92	0.76	0.83
character length (form)	0.35	0.88	0.50
character length (stripped lemma)	0.35	0.81	0.49
syllable length (form)	0.39	0.80	0.51
syllable length (stripped lemma)	0.39	0.72	0.51
unigram character probability	0.32	0.85	0.47
bigram character probability	0.24	0.90	0.38
test			
all	0.15	1.00	0.26
content	0.25	0.99	0.40
frequency	0.94	0.76	0.84
character length (form)	0.35	0.88	0.50
character length (stripped lemma)	0.36	0.85	0.51
syllable length (form)	0.38	0.84	0.52
syllable length (stripped lemma)	0.41	0.73	0.53
unigram character probability	0.30	0.83	0.44
bigram character probability	0.22	0.86	0.35

Table 6.2: Performances of complex word identification strategies (evaluated against words marked as complex by a majority of annotators)

and agaist words marked as complex by all annotators was also performed, see tables 6.1 and 6.2. All strategies of course lead to the same results, only the data that the results are evaluated against are different.

By definition, fewer words are considered complex in those evaluations. Precision therefore cannot be higher (the number of words marked as complex is still the same but some words might no longer be considered complex). The greater drop in precision, however, the greater ability to capture a potencial complexity for a reader. Contrarily, the lesser drop, the greater fixation at universally complex words.

Recall, on the other hand, could either increase or decrease. The increase is probably more expectable as fewer words can be retrieved. Similarly to drop in precision, a drop in recall is a sign of capturing words that are complex only for some readers.

6.2 Substitution generation

All words marked as complex by at least one annotator were used in the evaluation set (as mentioned before, there are 251 such complex words). Substitution generation was tested using those words, it was evaluated first separately for dictionary-based sources and for embedding-based sources, then jointly.

	precision	recall	F-measure
train			
all	0.13	1.00	0.23
content	0.21	0.98	0.35
frequency	0.80	0.75	0.77
character length (form)	0.30	0.86	0.44
character length (stripped lemma)	0.30	0.78	0.43
syllable length (form)	0.33	0.77	0.46
syllable length (stripped lemma)	0.36	0.70	0.45
unigram character probability	0.28	0.85	0.42
bigram character probability	0.21	0.89	0.34
test			
all	0.14	1.00	0.24
content	0.22	0.99	0.36
frequency	0.87	0.79	0.83
character length (form)	0.31	0.89	0.46
character length (stripped lemma)	0.32	0.84	0.47
syllable length (form)	0.34	0.83	0.48
syllable length (stripped lemma)	0.36	0.71	0.48
unigram character probability	0.27	0.83	0.40
bigram character probability	0.20	0.86	0.32

Table 6.3: Performances of complex word identification strategies (evaluated against words marked as complex by all annotators)

Substitutions were generated for each word using all sources. A list of all substitutions generated by a set of sources (dictionary-based sources, form-based embeddings, stripped lemma-based embeddings and lemma-based embeddings) was presented to annotators, the format was a complex sentence with emphasized complex word followed by a list of suggested substitutions, one substitution per line, and an empty line, followed by another complex sentence, etc. The set of sources was known to the annotators, the information about which substitution was generated by which source was not.

The annotators' task was to delete lines containing substitutions which could not be really used as a substitution in the given context. They were instructed to keep all synonyms, no matter whether they were simpler or more complex than the original complex word. They were also instructed to keep the very same word if it was present. All lines preserved by at least one annotator were considered as rated to be synonyms.

Precision and recall were then computed for each source. The term *generated substitution* in the following discussion refers to a substitution generated by the given source. The term *correct substitution* refers to a generated substitution which was rated as true synonym by humans (i. e. rated as a true synonym by at least one annotator).

Precision is computed as the ratio between generated substitutions and correct substitutions. Recall is computed as the ratio between correct substitutions and correct subtitutions generated by all sources altogether.

Both precision and recall are given both macro-averaged and micro-averaged. Here macro-averaging refers to computing precision/recall for each complex word and then averaging those values. Micro-averaging refers to counting all correct substitutions and all generated substitutions and using those values directly (no distinction between individual complex words is done).

Evaluation of dictionary-based sources

Even though the annotators were asked to accept any complex word as its own substitution, substitutions equal to stripped lemma of the complex word were disregarded during evaluation.

In total, no substitution was generated for 81 complex words. To check the existence of a substitution, I asked some annotators to give substitutions for those complex words (in the context of their sentences). There were 5 complex words for which no annotator could give a substitution.

Those 5 complex words (and their whole sentences) were:

- "Zakryl mi rukou ústa a kývnul hlavou, *jakože* rozumí." (*He covered my mouth with [his] hand and nodded *as if* he understood.*)
- "Z *tmavnoucího* žita plály vlčí máky, minul žebřiňák naložený snopy. (*Field poppies were glowing in *darkening* rye, he went past a hay wagoon loaded with sheafs.*)
- "Procházíte-li těmi končinami, nemůžete minout zplundrované *sudetoněmecké* hřbitovy." (*If you walk through those regions, you cannot miss plundered *Sudeten German* graveyards.*)

source	none	incorrect	mixed	correct	usable
dictionary A	146	14 (06 %)	79 (31 %)	12 (05 %)	90 (36 %)
dictionary B	160	16 (06 %)	54 (22 %)	21 (08 %)	75 (30 %)
wordnet	184	21 (08 %)	20 (08 %)	26 (10 %)	46 (18 %)
babelnet	167	11 (04 %)	51 (20 %)	22 (09 %)	73 (29 %)

Table 6.4: Substitution generation results of dictionary-based sources; counts of words are given; none = no substitution generated, incorrect = only irrelevant substitutions generated, mixed = both irrelevant and relevant substitutions generated, correct = only relevant substitutions generated, usable = mixed + correct

source	macro precision	macro recall	micro precision	micro recall
dictionary A	0.45	0.45	0.38	0.55
dictionary B	0.48	0.30	0.39	0.30
wordnet	0.51	0.13	0.46	0.12
babelnet	0.52	0.34	0.39	0.30

Table 6.5: Precision and recall of substitution generation for dictionary-based synonym sources

- "V této *mnišsky* prosté cele se zrcadlil svět." (*The world reflected itself in this *monasticly* plain room.*)
- "Než se *nadál*, nějaký lump mu pilu ukradl." (*Anon, some scroundel stole his saw.*, expressing the word anon with a phraseme, part of which is the complex word)

Those 5 complex words are counted towards the 81 complex words with no substitution, which however leaves 76 complex words for which no substitution was generated using the dictionaries but some were suggested by humans.

The counts of words for which no substitution was generated are even higher when evaluating the sources individually (see table 6.4).

While no substitution is generated for 81 (76) words when using all sources, which leaves about one third of words without substitution, 146-184 words, which makes almost two thirds of all complex words, are left without substitutions when using only one of the sources.

The precision is not very good and macro precision is generally better than micro precision (see table 6.5). Macro precision ranges from 0.45 to 0.52, micro precision ranges from 0.38 to 0.46. Words with many senses (and thus many incorrect substitutions) could explain the difference in macro and micro precision.

It is probably worth noting that even though the evaluation is binary (generated substitution either is a synonym or it is not), the task itself is not as the task is not to decide whether a word is a synonym for another word but to generate all synonyms for a given word.

Recall is generally worse than precision, being as low as 13 % for WordNet and in general not exceeding 50 %. This is not necesserily bad as even though a greater set of synonyms increases the chances of selecting an ideal substitution, any simpler substitution would be well usable.

source	absolute	relative
dictionary A	36 %	86 %
dictionary B	30 %	82 %
wordnet	18 %	69 %
babelnet	29 %	87 %

Table 6.6: Ratios of words with at least one relevant suggestion compared to all words in testset (absolute) and to all words with at least one substitution generated (relative)

corpus	text	none	incorrect	mixed	correct	usable
corp	orig	160	60 (24 %)	29 (12 %)	2 (01 %)	31 (12 %)
corp	drop	163	54 (22 %)	32 (13 %)	2 (01 %)	34 (14 %)
corp	subst	163	54 (22 %)	32 (13 %)	2 (01 %)	34 (14 %)
wiki	orig	186	56 (22 %)	9 (04 %)	0 (00 %)	9 (04 %)
wiki	drop	188	53 (21 %)	10 (04 %)	0 (00 %)	10 (04 %)
wiki	subst	188	52 (21 %)	11 (04 %)	0 (00 %)	11 (04 %)

Table 6.7: Substitution generation results of embeddig-based sources trained on word forms; counts of words are given; none = no substitution generated, incorrect = only irrelevant substitutions generated, mixed = both irrelevant and relevant substitutions generated, correct = only relevant substitutions generated, usable = mixed + correct

I should note that neither macro nor micro precision was affected by words for which no substitution was generated. In a sense, the resulting numbers for a source are more relevant if less words are left without any substitution.

On the other hand, detecting that no substitution was generated is far easier than detecting the incorrectness of the substitution generated. If no substitution is generated, the complex word obviously cannot be simplified as there is nothing to replace it, yet it will not be replaced with a nonsuitable word either. For this reason, the ratio of words with at least one correct substitution to all words with at least one substitution is also important. Those ratios however seem to be promising for all sources (see table 6.6).

Evaluation of embedding-based sources

Embeddings trained on word forms perform much worse than dictionary-based strategies. While the number of words that do not have any substitution generated is similar (see tables 6.4 and 6.7), embedding-based strategies produce only incorrect substitutions more often, up to five times more. This especially means that they succeed less often.

Embeddings trained on stripped lemmata completely outperform any other strategy in terms of words without substitutions (see table 6.8). The model trained on the original text of the 'corp' corpus resulted in only two words with no substitution. However, the words with substitutions add both to words having only incorrect substitutions and to words having both correct and incorrect substitutions. Words having correct substitutions only are extremely rare with

corpus	text	none	incorrect	mixed	correct	usable
corp	orig	2 (01 %)	149 (59 %)	98 (39 %)	2 (01 %)	100 (40 %)
corp	drop	5 (02 %)	146 (58 %)	95 (38 %)	5 (02 %)	100 (40 %)
corp	subst	5 (02 %)	145 (58 %)	98 (39 %)	3 (01 %)	101 (40 %)
wiki	orig	39 (16 %)	177 (71 %)	35 (14 %)	0 (00 %)	35 (14 %)
wiki	drop	40 (16 %)	176 (70 %)	35 (14 %)	0 (00 %)	35 (14 %)
wiki	subst	40 (16 %)	175 (70 %)	36 (14 %)	0 (00 %)	36 (14 %)

Table 6.8: Substitution generation results of embeddig-based sources trained on stripped lemmata; counts of words are given; none = no substitution generated, incorrect = only irrelevant substitutions generated, mixed = both irrelevant and relevant substitutions generated, correct = only relevant substitutions generated, usable = mixed + correct

corpus	text	none	incorrect	mixed	correct	usable
corp	orig	111 (44 %)	37 (15 %)	99 (39 %)	4 (02 %)	103 (41 %)
corp	drop	113 (45 %)	39 (16 %)	95 (38 %)	4 (02 %)	99 (39 %)
corp	subst	113 (45 %)	40 (16 %)	95 (38 %)	3 (01 %)	98 (39 %)
wiki	orig	120 (48 %)	90 (36 %)	41 (16 %)	0 (00 %)	41 (16 %)
wiki	drop	120 (48 %)	88 (35 %)	43 (17 %)	0 (00 %)	43 (17 %)
wiki	subst	120 (48 %)	88 (35 %)	43 (17 %)	0 (00 %)	43 (17 %)

Table 6.9: Substitution generation results of embeddig-based sources trained on lemmata; counts of words are given; none = no substitution generated, incorrect = only irrelevant substitutions generated, mixed = both irrelevant and relevant substitutions generated, correct = only relevant substitutions generated, usable = mixed + correct

embeddings, though this could be attributed to the way of selection – five words are always returned and few words have that many synonyms.)

However, the low count of words without substitutions need not be a good thing. No substitution happens if the complex word is not present in the model used. Higher numbers of words without substitutions obtained by using embeddings trained on lemmata suggest that the complex words are not in the model. However, something that has its lemma stripped to the same string is in the model so some substitutions get returned. Unfortunately, two words sharing stripped lemma but not lemma are most likely homonyms[1] so their substitutions are not interchangeable.

Indeed, embeddings trained on lemmata have higher numbers of words without substitutions but very similar numbers of lemmata with at least one correct substitution (see tables 6.9 and 6.8). In other words, words which move from the 'no substitution' category when switching from lemma-trained embeddings to stripped-lemma-trained embeddings most likely move to the 'no correct substitution' category.

Lemmata seem to be the best to train embeddings on if the goal is substitution generation. Embeddings trained on lemmata perform better than both embeddings trained on sripped lemmata (especially in terms of words with incorrect substitutions only) and embeddings trained on forms (especially in terms of words without substitution). This actually complies with some linguistic intuition.

It is probably worth noting that models trained on the 'corp' corpus perform better than models trained on the 'cswiki' corpus. This can probably be attributed to the relatively small size of Czech Wikipedia.

It should also be noticed that while both non-content word dropping and substituting has a little effect, the effect is not very big (and is not always positive, those text alterations seem to hurt the performance of models trained on lemmata).

Joint evaluation

Dictionary-based strategies perform much better in terms of precision which could again be caused by the way of embedding-based substitution selection – selecting either five or zero substitutions each time. With dictionary-based strategies, there is a possibility of generating for example two or three substitutions.

The increased precision of some strategies (compare tables 6.4 and 6.10) is notable because this could be a sign of some inconsistencies among the annotations.

The recall is generally not very good. Surprisingly, it is the best for embeddings trained on stripped lemmata using the 'corp' corpus. However, this can be explained by the number of words that get both correct and incorrect substitutions generated when using those embeddings.

[1]Homonyms are words which are spelled/pronounced the same way but have a different meaning.

strategy	macro - precision	- recall	micro - precision	- recall
dictionary A	0.46	0.19	0.39	0.27
dictionary B	0.51	0.12	0.41	0.15
wordnet	0.54	0.05	0.48	0.06
babelnet	0.52	0.14	0.39	0.14
corp/form/orig	0.22	0.06	0.22	0.06
corp/form/drop	0.23	0.06	0.23	0.07
corp/form/subst	0.24	0.07	0.24	0.07
wiki/from/orig	0.10	0.02	0.10	0.02
wiki/from/drop	0.10	0.01	0.10	0.02
wiki/from/subst	0.10	0.01	0.10	0.02
corp/slem/orig	0.25	0.23	0.25	0.20
corp/slem/drop	0.25	0.23	0.25	0.20
corp/slem/subst	0.26	0.23	0.26	0.20
wiki/slem/orig	0.09	0.06	0.09	0.06
wiki/slem/drop	0.09	0.05	0.09	0.06
wiki/slem/subst	0.08	0.05	0.08	0.06
crop/lemm/orig	0.31	0.14	0.31	0.14
crop/lemm/drop	0.34	0.14	0.34	0.15
crop/lemm/subst	0.33	0.14	0.33	0.15
wiki/lemm/orig	0.11	0.03	0.11	0.05
wiki/lemm/drop	0.10	0.04	0.10	0.04
wiki/lemm/subst	0.11	0.04	0.11	0.04

Table 6.10: Precisions and recalls of substitution generation strategies; embedding-based strategies described as corpus/key/preprocess; orig = no changes done to the text, drop = non-content words dropped, subst = non-content words substituted

strategy	correct rankings
frequency	113 (50 %)
character length (stripped lemma)	75 (33 %)
syllable length (stripped lemma)	72 (32 %)
unigram character probability	51 (23 %)
bigram character probability	47 (21 %)
Mistrík formula	72 (32 %)

Table 6.11: Substitution ranking results – the number of times substitution selected by the strategy was considered simpler by humans

6.3 Substitution ranking

Evaluation of substitution ranking is not a straightforward task. It would make sense to count how many times both humans and the ranking strategy in question rank the same substitution as the best. It would make sense to count how many times the strategy selects a substitution which is not worse than a substitution selected by humans. It would make sense to count the ratio of commonly ordered substitution pairs between the strategy and humans. And this could go forth.

I have decided to count how many times the strategy selects any substitution that humans consider simpler than the complex word. Scoring high therefore does not necessarily imply that the strategy ranks correctly in terms of mutual ordering but it does imply that the strategy produces results usable for text simplification.

I took the substitutions rated as correct by humans and asked some annotators to select those which were simpler than the complex word. All substitutions considered simpler by at least one annotator were treater as simpler.

After that, I rated the substitutions using all strategies and counted how many times the best substitution was considered simpler by humans. The ratios reported are not counted against the 251 complex words but against 224 words for which a simpler substitution exsts according to humans.

Frequency-based strategy outperforms all other strategies (see table 6.11). This complies with high precision of this strategy also for complex word identification (though it is still necessary to keep the preselection bias in mind), it also complies with some intuition.

Mistrík formula and syllable length scored the same, as expected. A longer snippet of text would probably needed to really distinguish performaces of those two strategies.

7. Notes on implementation

A lot of programming has been part of this thesis and I would like to share some notes on the programming part and on the tools used. Few scripts are actually attached to the thesis, not only because they are mostly Perl and sometimes really write-only but also because most of them were meant only to measure some statistics or randomly sample some results.

Still, there are also scripts worth (or almost worth) sharing. I should however start with notes on third-party software I have been using.

7.1 Third-party software used in the work

I have used several pieces of third-party software in the course of the work on the thesis. I will list them here and give some hints on their installation and usage.

MorphoDiTa

MorphoDiTa[1] is a tool for morphological analysis, morphological generation and tagging. See Straková et al. (2014) for details.

All corpora I have used have been tagged with MorphoDiTa, except for the PDT corpus which contains manual annotation. I also use MorphoDiTa to generate a correct form of a selected substitution and because of that, it is a crucial part of the simplification pipeline.

As for installation and setup, I cloned MorphoDiTa from its repository[2] and '`cd src && make`' then did the trick. To really work, MorphoDiTa actually requires a trained model, such models can be downloaded from the MorphoDiTa homepage. I use the one provided by Straka and Straková (2016).

As the user manual suggests, tagging can be started by running `run_tagger tagger_file [file:[output_file]]`. Unfortunately, to the extent of my knowledge, MorphoDiTa does not support any kind of masking so tagging a whole corpus requires specifying each file with its associated output file on the command line.[3] (It is of course possible to let MorphoDiTa tag several files without specifying the output, contents of the files however will not be separated in any way.) It is also worth noting that on the contrary, any string gets tagged as if were word, including document boundary marks if they are present in the corpus. I use vertical output, not XML.

Tagging of the 'corp' corpus took approximately 10 hours on the machine I used. This time could be greatly reduced since MorphoDiTa is (seems to be) perfectly parallelizable. However, I do not know any better way to parallelize than executing the tagger several times, each time with different file arguments.

Beware that some of my scripts use MorphoDiTa's Perl bindings. The folder `bindings/perl/` therefore has to be included in `@INC`. The scripts attemp to add `../prgs/MorphoDiTa/bindings/perl/` so in case you are unfamiliar with

[1] http://ufal.mff.cuni.cz/morphodita

[2] https://github.com/ufal/morphodita

[3] I admit I usually ended up writing a script for that.

Perl and/or do not know how to alter its `@INC`, you can just follow the directory structure described later.

Gensim

Gensim[4] is a tool for natural language processing and information retrieval. See Řehůřek and Sojka (2010) for details. I use the tool for training embedding models and for obtaining synonyms from those models.

Gensim is written in Python and requires NumPy and Scipy on top of Python itself to work. Gensim's author suggests to install using either `EasyInstall` (`easy_install -U gensim`) or `pip pip install -upgrade gensim`. I ran into some issuess when trying to install that way but this had possibly a lot to do with my non-root access to the machine in question. When installing from tar source, I had to manually fulfill some dependencies, namely install boto[5] and smart_open[6].

The tool does not require any special file format to train a model from text, I however sticked to one sentence per line because of simple sentence yielding.

wp2txt

I have used wp2txt[7] to extract text from the Wikipedia dump. I used the suggested way to install it, i. e. via `gem install wp2txt`.

This tool is not necessary as no reported statistics depend on it but I mention it here for completeness.

7.2 User manual

I should warn that since this thesis was more exploratory and experimental than implementational, the scripts are not very user-friendly. Scripts located in the `dev` subfolder are even less so.

I suppose most people would be interested in one specific script: snadnik.pl. This script is able to control a simplification pipeline and simplify a prepared text. To run it, MorphoDiTa must be installed and its Perl bindings must be either on Perl's `@INC` path or in the directory `../prgs/MorphoDiTa/bindings/perl/` relative to the `scripts` folder.

The input to this script is a tagged document, i. e. output of MorphoDiTa tagger (selecting vertical output istead of XML output is necessary). Conventionally, it would be located under `data/corpora/tagged/corpus_name/` but this is not needed.

The user is expected to select a complex word identification strategy, substitution generation and substitution selection strategy via command line options, i. e. by typing `-cwi=(strategy)`, `-sg=(strategy)` and `-sr=(strategy)`. All three strategies must be selected.

[4]https://radimrehurek.com/gensim/
[5]https://pypi.python.org/pypi/boto
[6]https://pypi.python.org/pypi/smart_open
[7]https://github.com/yohasebe/wp2txt

For both `cwi` and `sr` options, the following values are valid:

- `freq`, frequency-based strategy,
- `char_form`, character length applied to word form,
- `char_slem`, character length applied to stripped lemma,
- `syll_form`, syllable length applied to word form,
- `syll_slem`, syllable length applied to stripped lemma,
- `uniprob`, unigram probability,
- `biprob`, bigram probability.

Two more values are valid for the `cwi` option:

- `all`, identification of all words,
- `content`, identification of all content words.

One more value is valid for the `sr` option:

- `mistrik`, Mistrík formula.

As for the the `sg` option, two values are valid:

- `wordnet`, using Czech Wordnet,
- `dict`, using a dictionary.

If the value `freq` is specified for either `cwi` or `sr` option, the option `db-path` can also be specified to give the path to a stripped lemma frequency database. Such database is one file, with one line per each stripped lemma, and the format of the line is stripped lemma and the corresponding count, separated by a tab.

Similarly, `dict-path` can be specified for any value of the `sg` option. If the value `wordnet` is used, the file referred to by `dict-path` has to be an XML-formatted WordNet. Otherwise, the dictionary should have one synset per line and words in one synset should be separated by a tab.

On top of the strategies, the user can specify both `-input (file)` and `-output (file)`.

The script writes its output to the specified output file. The format is smilar to the input format, except that there are additional comments on each line.

7.3 Miscellaneous notes

Some notes are not important enough to create a special section for them, not general enough to fit elsewhere in the text but neither unimportant enough to just leave them out. I would like to mention all of them here, though it could be a little messy.

- When computing character length of any kind of key, I do it by a simple call to the Perl `length` function, which should take care of dealing with logical characters (opposed to physical bytes).
- For historical reasons, dictionary A is referred to as the 'synonyma' dictionary somewhere in code. For similar historical reasons, dictionary B is referred to as the 'slovnik' dictionary.
- Derivation of the substitution's tag is quite simple and mainly driven by an attempt to require a grammatically correct word but also allow to generate some output. At first, I started with copying the tag from the complex word but this approach can lead to a situation when no form is generated. This happens for example when substituting a noun having feminime gender with a noun having neuter gender and leaving the gender unchanged in the tag.

8. Conclusion

I have experimented with text simplification, particularly with text simplification in Czech. To the extent of my knowledge, this is a pioneering work for Czech, though text simplification has already been studied for some other languages. As such, this thesis, among other things, reviews readability measures that could perhaps be used for Czech texts.

I have conducted three experiments involving humans and arranged several annotation sessions to get better insight into perceived simplificity and its factors. I especially value the diversity of my annotators as they were of various ages and educational backgrounds and included a non-native speaker, a person suffering from a mental disorder, a person dealing with hearing-impaired people on a daily basis and dyslexic people. The experiments have brought valuable outcomes.

Besides experimenting with people, I have implemented several strategies for lexical simplification, a subtask of text simplification. I have paid special attention to substitution generation, the process of generating substitutions for a complex word but I have also implemented and evaluated various methods for complex word identification and for substitution selection.

The precision of complex word identification strategies is not very good except for the frequency-based strategy. A bias however exists towards this strategy and its precision should be confirmed in more neutral settings. On the other hand, the recall of those strategies generally ranges from 0.75 to 0.85, which is good, especially given that substituting a non-complex word is likely to be a less severe problem than leaving a complex word in place.

Substitution generation strategies generate substitutions for about one third of complex words when used in isolation but manage to generate substitutions for more than two thirds of complex words when combined. This suggests that resource development is likely to help this task. I have also experimented with strategies based on word embeddings. Embeddings, especially if trained on lemmata, seem to be an interesting alternative to dictionary-based approaches.

The frequency-based strategy scores 50 % in substitution ranking, i. e. it manages to choose a truly simpler substitution half of the time. Other strategies perform worse but they still score around 30 %. While this might seem little, it is 30-50 % more than not being able to automatically simplify at all.

I have also implemented a script which allows to try any of the tested strategies on one's own corpus. This script in attached to this thesis via the enclosed CD.

I have not solved the task of text simplification – and I had not expected to – but I have presented a few points any future work can be based on.

8.1 Future work

While researching the task of text simplification and performing various experiments, I got some ideas on what could be viable and worth a try. I would therefore like to suggest some future works that could bring additional (and hopefully better) results.

Language modelling

Language modelling could greatly help both with substitution ranking and with substitution selection, either explicit or implicit as a part of ranking. Some of my preliminary experiments suggest that modelling could be really helpful (albeit unfortunately not miraculous) and further experiments would be very contributive.

Dealing with multiword expressions

As I have pointed out several times, this thesis only deals with single words and does not consider multiword expressions. Such expressions might however require simplification or, by contrast, be a good substitution for a complex word.

Some decisions would have to be made (for example, what a multiword expression has to fulfill to be treated as a multiword expression) and some steps including morphological generation and the whole step of substitution ranking would have to be adapted to correctly deal with such expressions. On the other hand, the benefits could be substantial.

Embedding-related experiments

I have used word embeddings to generate substitutions in this thesis and the results are promising. Substitutions generated by embeddings trained on lemmata are more likely to be correct than substitutions generated by dictionary-based strategies.

More experiments could bring better insight into the capabilities and limitations of embeddings. Those experiments might include alterations of the model to better adapt it for synonymity as well as training with different architectures, context windows, vector dimensions, ...

Explanation generation

If explanation generation is employed, the complex word is left in place and an explanation is generated to make it possible to understand that word. This strategy is sometimes used by humans as described in section 4.3. While people use their intuition, text comprehension and their own knowledge when generating explanations, artifitial methods could try to make use of knowledge bases like Wikipedia.

De-passivization

Even though some sentences in passive voice cannot be transformed to active voice, transformations of some others should be feasible. Describing some rules to complete such de-passivization and evaluating them would definitely be an interesting experiment.

Sentence splitting

I have dealt with lexical simplification but I believe sentence splitting is a promising approach to syntactic simplification. As I mentioned in section 3.4, splitting could get very tricky sometimes, however I believe it would bring some results.

Slightly generalized search for words in a dictionary

Dictionary-based strategies tend to give acceptable results in substitution generation. Still, they often fail to generate any suggestions. Based on my experience, sometimes this is because the complex word itself really is not in the dictionary searched through but a very similar word is – a verb with different aspect or, more likely, an adjective instead of a verb and the like.

Retrieving substitutions from Wikipedia redirects

I mentioned Wikipedia redirects in section 5.3.1. I have experimented with this resource and decided not to use it in this thesis, however I believe some interesting and important substitutions could be mined from it, though some heuristics would have to be employed to filter out a lot of unuseful material.

Pronoun insertion

Pronoun insertion might be very complicated under some circumstances. Yet it could be done easily under some other circumstances. Even though I believe there are more valuable experiments, I suggest researching the circumstances which allow to re-insert a dropped pronoun.

Foreign word detection

I consider this a minor suggestion only but sometimes reading difficulties are caused by a presence of a foreign word. Such word generally cannot be simplified but it could be detected and either translated, explained or enhanced with hints on pronnunciation.

As I have said, I believe those ideas could lead to better results and could help take the simplification further. After all, the first steps have already been done.

Bibliography

Marcelo Adriano Amancio and Lucia Specia. An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 123–130, 2014.

Mandya Angrosh, Tadashi Nomoto, and Advaith Siddharthan. Lexico-syntactic text simplification and compression with typed dependencies. In *COLING*, pages 1996–2006, 2014.

María Jesús Aranzabe, Arantza Díaz De Ilarraza, and Itziar Gonzalez-Dios. Transforming complex sentences using dependency trees for automatic text simplification in basque. *Procesamiento del lenguaje natural*, 50:61–68, 2013.

Gianni Barlacchi and Sara Tonelli. Ernesta: A sentence simplification tool for children's stories in italian. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 476–487. Springer, 2013.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague dependency treebank 3.0, 2013. URL `http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Or Biran, Samuel Brody, and Noémie Elhadad. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 496–501. Association for Computational Linguistics, 2011.

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. Many czech references for 50 sentences selected from WMT11 data, 2013. URL `http://hdl.handle.net/11858/00-097C-0000-0023-10B2-F`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Peter Brodsky and Heidi Waterfall. Characterizing motherese: On the computational structure of child-directed language. In *Proceedings of the Cognitive Science Society*, volume 29, 2007.

David Caplan. *Language: Structure, processing, and disorders.* The MIT Press, 1992.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Citeseer, 1998.

Han-Bin Chen, Hen-Hsen Huang, Hsin-Hsi Chen, and Ching-Ting Tan. A simplification-translation-restoration framework for cross-domain smt applications. In *COLING*, pages 545–560, 2012.

Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, page 30. ACM, 2013.

William Coster and David Kauchak. Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation*, pages 1–9. Association for Computational Linguistics, 2011a.

William Coster and David Kauchak. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669. Association for Computational Linguistics, 2011b.

Walter Daelemans, Anja Höthker, and Erik F Tjong Kim Sang. Automatic sentence simplification for subtitling in dutch and english. In *LREC*, 2004.

Jan De Belder and Marie-Francine Moens. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM, 2010.

Jan De Belder, Koen Deschacht, and Marie-Francine Moens. Lexical simplification. In *Proceedings of ITEC2010: 1st international conference on interdisciplinary research on technology, education and communication*, 2010.

Louise Deléger and Pierre Zweigenbaum. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 2–10. Association for Computational Linguistics, 2009.

Siobhan Devlin and Gary Unthank. Helping aphasic people process online information. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 225–226. ACM, 2006.

Noemie Elhadad and Komal Sutaria. Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 49–56. Association for Computational Linguistics, 2007.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.

Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32 (3):221–233, 1948.

Edward Fry. A readability formula that saves time. *Journal of Reading*, 11, 1968.

Dee Gardner and Elizabeth C Hansen. Effects of lexical simplification during unaided reading of english informational texts. *TESL REPORTER*, 40(2):27, 2007.

Kevin R Gregg and Stephen D Krashen. *The input hypothesis: Issues and implications*. JSTOR, 1986.

Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

Donald P Hayes and Margaret G Ahrens. Vocabulary simplification for children: A special case of 'motherese'? *Journal of child language*, 15(02):395–410, 1988.

James E. Hoard, Rirchard Wojcik, and Holzhauser Katherina. *Computers and Writing*, chapter An Automated Grammar and Style Checker for Writers of Simplified English, pages 278–296. Springer Netherlands, 1992.

Bui Thanh Hung, Nguyen Le Minh, and Akira Shimazu. Sentence splitting for vietnamese-english machine translation. In *Knowledge and Systems Engineering (KSE), 2012 Fourth International Conference on*, pages 156–160. IEEE, 2012.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 9–16. Association for Computational Linguistics, 2003.

Eva Janoušková. *Analýza učebnic zeměpisu*. PhD thesis, Faculty of Education, Masaryk University, 2008.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. A semantic and syntactic text simplification tool for health content. In *AMIA Annu Symp Proc*, volume 2010, pages 366–70, 2010.

Robin Keskisärkkä. Automatic text simplification via synonym replacement. Master's thesis, Lipköping University, 2012.

Sigrid Klerke and Anders Søgaard. Simple, readable sub-sentences. In *ACL (Student Research Workshop)*, pages 142–149, 2013.

Marie Komorná. *Psaná čeština českých neslyšících*. Česká komora tlumočníků znakového jazyka, 2008.

Michael H Long and Steven Ross. Modifications that preserve language and content. 1993.

Annie Louis and Ani Nenkova. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352, 2013.

William Anderson McCall and Lelah Mae Crabbs. *Standard Test Lessons in Reading...* Number 5. Teachers College, Columbia University, Bureau of Publications, 1926.

G. Harry McLaughlin. Smog grading: A new readability formula. *Journal of Reading*, 12:639–646, 1969.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892*, 2016.

Courtney Napoles, Benjamin Van Durme, and Chris Callison-Burch. Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 91–97. Association for Computational Linguistics, 2011.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

Gustavo Paetzold and Lucia Specia. Text simplification as tree transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 116–125, 2013.

Gustavo Paetzold and Lucia Specia. Lexenstein: A framework for lexical simplification. In *ACL (System Demonstrations)*, pages 85–90, 2015.

Karel Pala, Tomáš Čapek, Barbora Zajíčková, Dita Bartůšková, Kateřina Kulková, Petra Hoffmannová, Eduard Bejček, Pavel Straňák, and Jan Hajič. Czech WordNet 1.9 PDT, 2011. URL `http://hdl.handle.net/11858/00-097C-0000-0001-4880-3`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Sarah E. Petersen and Mari Ostendorf. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23:80–106, 2009.

Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. Association for Computational Linguistics, 2008.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The penn discourse treebank 2.0. In Nicoletta (Conference Chair) Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odejk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Dana Prchalová. Zkoumání čtivosti – srovnání způsobů měření obtížnosti textu. Bachelor's thesis, Faculty of Arts, Charles University, 2013.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219. Springer, 2013.

Violeta Seretan. *Acquisition of Syntactic Simplification Rules for French.* Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7. URL `https://archive-ouverte.unige.ch/unige:30961`. ID: unige:30961.

Matthew Shardlow. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, (Special Issue on Natural Language Processing):58–70, 2014.

Cynthia M Shewan and Gerald J Canter. Effects of vocabulary, syntax, and sentence length on auditory comprehension in aphasic patients. *Cortex*, 7(3): 209–226, 1971.

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

Advaith Siddharthan. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298, 2014.

Malcolm Smith and Richard Taffler. Readability and understandability: Different measures of the textual complexity of accounting narrative. *Accounting, Auditing & Accountability Journal*, 5(4), 1992.

Lucia Specia. Translating from complex to simplified sentences. In *International Conference on Computational Processing of the Portuguese Language*, pages 30–39. Springer, 2010.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 347–355. Association for Computational Linguistics, 2012.

Sanja Štajner, Biljana Drndarevic, and Horacio Saggion. Corpus-based sentence deletion and split decisions for spanish text simplification. *Computacion y systemas*, 17(2):251–262, 2013.

Milan Straka and Jana Straková. Czech models (MorfFlex CZ 161115 + PDT 3.0) for MorphoDiTa 161115, 2016. URL `http://hdl.handle.net/11234/1-1836`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P14/P14-5003.pdf`.

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16*, number 085, pages 375–386. Linköping University Electronic Press, 2013.

Lukáš Svoboda and Tomáš Brychcín. New word analogy corpus for exploring embeddings of czech words. *arXiv preprint arXiv:1608.00789*, 2016.

Wilson L Taylor. "cloze procedure": a new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433, 1953.

S Rebecca Thomas and Sven Anderson. Wordnet-based lexical simplification of a document. In *KONVENS*, pages 80–88, 2012.

Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. Facilita: reading assistance for low-literacy readers. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 29–36. ACM, 2009.

Kristian Woodsend and Mirella Lapata. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the conference on empirical methods in natural language processing*, pages 409–420. Association for Computational Linguistics, 2011.

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

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

Nikolay Yakovets and Ameeta Agrawal. Simple: Lexical simplification using word sense disambiguation., 2013.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368. Association for Computational Linguistics, 2010.

Olga Zelinková. *Poruchy učení*. Portál, 2015.

Olga Zelinková and Miloslav Čedík. *Mám dyslexii*. Portál, 2013.

F. Čermák, R. Blatná, J. Hlaváčová, J. Klímová, J. Kocek, M. Kopřivová, M. Křen, V. Petkevič, V. Schmiedtová, and M. Šulc. Syn2000: žánrově vyvážený korpus psané češtiny, a. Dostupný z WWW: http://www.korpus.cz.

F. Čermák, D. Doležalová-Spoustová, J. Hlaváčová, M. Hnátková, T. Jelínek, J. Kocek, M. Kopřivová, M. Křen, R. Novotná, V. Petkevič, V. Schmiedtová, H. Skoumalová, M. Šulc, and Z. Velíšek. Syn2005: žánrově vyvážený korpus psané češtiny, b. Dostupný z WWW: http://www.korpus.cz.

F. Čermák, D. Doležalová-Spoustová, J. Hlaváčová, M. Hnátková, T. Jelínek, J. Kocek, M. Kopřivová, M. Křen, R. Novotná, V. Petkevič, V. Schmiedtová, H. Skoumalová, M. Šulc, and Z. Velíšek. Syn2006pub: korpus psané publicistiky, c. Dostupný z WWW: http://www.korpus.cz.

A. Attachments

A.1 Contents of enclosed CD

File structure of the enclosed CD is as follow:

- thesis.pdf – this text
- experiments/
  - sent_readability_ranking/
    - ranking_scanned.pd – scanned evaluation sheets
  - sent_pair_comparison/
    - dump of 100 evaluated pairs, complete dump of ratings
    - sources of web application, a script for database initialization
  - manual_simplification
    - task zip archive as distributed to the annotators
    - all resulting sentences
- eval_cwi/
  - complex word annotations (as suggested by me + changes by other annotators)
  - corpus files with marked (non-)complex words
  - corpus files with marked (non-)complex words as produced by individual strategies
- eval_sg/
  - substitutions as produced by individual groups of strategies
  - substitutions as filtered by annotators
- eval_sr/
  - all correctly generated substitutions for complex words
  - simpler substitutions as filtered by annotators
- corpora/
  - a sample corpus
- scripts/
  - various scripts including snadnik.pl and scripts it depends on
  - _unsorted/ – some of the scripts used in the course of the work, in case anyone was interested in them