Filozofická fakulta Univerzity Palackého
Katedra anglistiky a amerikanistiky

Faculty of Arts
Palacký University Olomouc

Non-standard Czech in originals and translations from English: A corpus-based study

Bachelor's thesis

Olomouc 2023

Daniela Požárová

Non-standard Czech in originals and translations from English: A corpus-based study (Bakalářská práce).

Autor: Daniela Požárová
Studijní obor: Angličtina pro komunitní tlumočení a překlad
Vedoucí práce: Mgr. Michaela Martinková Ph.D.
Počet znaků: 69 793
Počet stran (podle znaků /1800): 38,7
Počet stran (podle čísel): 40

Abstract and annotation
This thesis is trying to shed some light on the usage of non-standard Czech, namely Common Czech, in translations. The key objective is to find out if translators use Common Czech to the same degree as Czech authors. This is a corpus-based study, using an openly accessible corpus InterCorp. The research is made utilizing a quantitative comparative analysis of features of Common Czech on the lexical level and below. This thesis is divided into a theoretical part with a focus on translation universals, Common Czech, and methodology and a theoretical part focused on analysis. The results indicate that there is certainly a difference in the usage of Common Czech in originals and translations. The results and future possibilities in this field of study are discussed in the conclusion.

Keywords
Non-standard Czech, Common Czech, translation, comparative analysis, corpus.

Abstrakt a anotace
Tato práce se snaží objasnit užívání nespisovné, zejména Obecné češtiny, v překladech. Hlavním cílem této práce je zjistit, jestli překladatelé používají Obecnou češtinu ve stejné míře jako autoři českých originálních textů. Tato studie ke svému výzkumu využívá korpus, konkrétně volně dostupný InterCorp. Výzkum je proveden metodou kvantitativní komparativní analýzy znaků Obecné čestiny v hláskosloví a tvarosloví. Tato práce je rozdělena na teoretickou část, která se zabývá překladovými univerzáliemi, Obecnou češtinou a využitou metodologií, a na praktickou část, která se soustředí na analýzu. Výsledky výzkumu ukazují, že v užití Obecné češtiny v originálech a překladech je opravdu rozdílné. Výsledky a budoucí možnosti výzkumu v tomto odvětví jsou prodiskutovány v závěru.

Keywords
Nespisovná čeština, Obecná čeština, překlad, komparativní analýza, korpus.

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a uvedla úplný seznam citované a použité literatury.

V Olomouci dne 11.12. 2023 **Daniela Požárová**

Děkuji vedoucí diplomové práce Mgr. Michaele Martinkové Ph.D. za ochotu, trpělivost, užitečnou metodickou pomoc a cenné rady při zpracování diplomové práce.

V Olomouci dne 11.12. 2023 **Daniela Požárová**

TABLE OF CONTENTS

1. Introduction....6

2. Translation....8

2.1. Translation Universals....8

2.1.1. Approaches to translation universals....8

2.1.2. T-universals and S-universals....9

2.1.3. Normalization....10

3. Non-standard Czech....10

3.1. Stratification of Czech....10

3.2. Common Czech....11

3.3. Features of Common Czech....11

3.3.1. Syntax....11

3.3.2. Lexical level....11

3.3.3. Phonological level....12

3.3.4. Morphological level....12

4. Register and Level of Politeness....12

4.1. Functional Approach....13

4.2. Approach of Systemic Functional Linguistics (SFL)....13

4.3. Register in relation to this study....13

5. Methodology....14

5.1. Corpus Selection....14

5.2. Data Collection....14

5.3. Data Analysis....15

5.4. Corpus Statistics....15

5.5. Limitations....16

6. Analysis....17

6.1. Change of -é into -ý....17

6.1.1. Inflectional suffix of adjectives....17

6.2. Change from -ý to -ej....20

6.2.1. Suffix in singular masculine adjectives in the nominative....20

6.2.2. Suffix in other adjectives....22

6.2.3. Word base....23

6.3. Prothetic v-....27

6.4. Phonetic shortening of vowel -í....30

6.5. Erasure of syllabic -l....31

6.6. Summary....33

7. Conclusion....37

8. References....39

9. Other sources....40

1. Introduction

Non-standard Czech, characterized by deviations from traditional grammar rules and vocabulary, is a linguistic phenomenon that has become increasingly prevalent in modern times. It is a topic of interest for linguists and scholars who seek to understand its usage patterns and impact on contemporary Czech literature.

Common Czech in particular is a unique linguistic phenomenon. There are many quite big differences between Standard Czech and Common Czech because unlike in other Slavic languages, rules and regulations for Standard Czech were made quite early (Sgall 2012, 18). But Common Czech is getting more prevalent and so it is a topic of many discussions across different fields. Translation studies and Linguistics go hand in hand. The topic of Common Czech and the place it has in the current Czech language should be brought to the attention of translators as well. Translators should be aware of current language trends, so their work is well received by the target audience. This can be accomplished only by producing high-quality work. In current times, with the swift development of machine translation and the use of artificial intelligence the bar for quality in translation is only rising higher. Nevertheless, the human factor is still important and the usage of Common Czech in translations should be analysed. The results might be beneficial for translators, who wish to improve or build their style, and anyone interested in Common Czech and its development.

The human interest in language along with technology expanding at a rapid pace contributes to the development of new research methods and means. It allows access to large databases of data like corpora for many people. All with different backgrounds, interests, and points of view.

This thesis will follow the current trend and explore the field of Translation studies utilizing corpus comparative analysis. It will hopefully contribute new information on the translation process and results of the translator's decision-making from the linguistic perspective.

This study aims to provide a comprehensive analysis of non-standard Czech in both original Czech texts and translations from English, using a corpus-based approach.

The topic of this thesis is a quantitative comparative analysis of Common Czech in original texts, written by Czech authors, and translated texts, written in English, and translated into Czech.

Common Czech is used by an increasing number of Czech speakers and authors. This thesis will study whenever when it comes to Common Czech usage translators prefer to use the method of naturalization. Which is "a tendency of translated texts to conform to target language rather than source language patterns and norms" (Zanettin 2013, 23), making the target text more natural. Or if translators have the tendency to use more formal language variants like Standard Czech. The analysis will create a clear picture of whether translations use similar, greater, or lesser amount of Common Czech. The hypothesis I will try to confirm is that translators tend to use Common Czech less than Czech authors. Translations will be more formal with Standard Czech variants over Common Czech.

The theory section will introduce the necessary topics like translation universals and naturalization, Stratification of the Czech language, Common

Czech, and what are its features and register. Then the methodology chapter will discuss the use of a corpus for comparative analysis, means of use, and touch upon the use of statistics.
The practical part will be the quantitative comparative analysis. The results will be presented in a clear and understandable fashion, using tables and graphs with a commentary. I will also provide examples of the most frequently used Common Czech variants of words in both sub-corpora along with their individual frequencies.

2. Translation

The most interesting question at this moment is probably why would occurrences of Common Czech variants be significantly different in original texts and translated texts.
While a person that is not experienced with translation and perhaps never tried to understand what it entails in depth may assume that translation is an easy process of exchanging words in one language for words in another, it is not so easy.
The authors of a text have the advantage, that not only are they equipped with the knowledge of a certain language, presumably their mother tongue, but they may also use it freely. They naturally express what they intend with the best-suited form.
Translators work with already complete text and have little to no creative input. In an ideal situation with an ideal text, translation balances original form and meaning with the natural target language. But unfortunately meeting this condition is very rare. And so, translators must make difficult choices, like what to keep and what to discard. What is important for the target audience and what is not.
As two languages cannot be the same, translation simply cannot be the same as the original text. The translation is a complicated process, where the key component is the translator and decision-making. They often must decide whether the form or naturalization is more important.
Previous research shows that there are methods translators tend to follow or rules they tend to adhere to. The process of translation has some universal features that can be observed in translated texts (Zanettin 2013, 21). They are called Translation universals.

2.1. Translation Universals

2.1.1. Approaches to translation universals

Research into translation universals is rooted in history when researchers moved from case-by-case to finding general features of translation. Now, contrastive studies supported by statistical research are one of the main tools used to find and either confirm or refute whether something is a generality.
Chesterman (2003, 213-218) differentiates between three approaches to finding generalities.
The first is "the prescriptive approach" (Chesterman 2003, 213). The focus is finding a set of rules that could be universally followed and that would lead to an ideal translation. "The first ideas about translation universals were thus universalist ideals which were based on a notion of sameness with respect to another text or texts." (Chesterman 2003, 214)
The second is "the pejorative approach" (Chesterman 2003, 215). "Here, all translations are regarded as being deficient in some way, and so I call it the pejorative approach. Instead of focusing on defining the ideal, in other words, we focus on the way in which, in practice, translations tend to fall short of this ideal. Contrastively, we focus exclusively on difference, not similarity." (Chesterman 2003, 215)
The last and most current is the "descriptive approach" (Chesterman 2003, 218). This approach is used in this thesis. It became a possibility with the establishment of international comparable corpora. "Corpus-based work that

generates and tests descriptive hypotheses that are thought to hold for all translations, or for all translations of a given sort." (Chesterman 2003, 218) This approach treats translation as a unique text type that is separate from the original. The research usually relies on beforehand set conditions that indicate for what kind and type of translation will the result apply.

All approaches mentioned above have a purpose but also have their problems. The prescriptive approach assumes that all translations can be judged by the same standards and that the texts of the same type are similar enough that if one is analysed the result will be applicable to the entire group (Chesterman 2003, 213). The pejorative approach ignores differences in rhetoric between languages (Chesterman 2003, 217). In recent years linguists and translators do not want to rely on assumptions without clear evidence. Especially when while it was impossible to get it in the past, it is possible and relatively easy now. While the critical approach might offer insight into problems of translation it does not necessarily offer a satisfactory solution to those problems and so it can be seen as a blind path.

The descriptive approach is the most popular nowadays, and it is important to know that it also has problems and limitations. The results of this kind of research indicate what translators tend to do, which in itself can be an overgeneralization (Chesterman 2003, 221). The human factor and decision making is subjective by nature.

Another issue is that even now it is impossible to analyse every text. A researcher needs to use a sample of language. The criteria for which translations can be added to a corpus are not universal and some texts will probably be excluded (Chesterman 2003, 222). Therefore, it is unlikely that the results will be applicable to all kinds of texts.

Scholars must realise that translation is bound to their language and culture, which means that global universality is unlikely to be found (Chesterman 2003, 222).

Finally, translators are always influenced by themselves. What they know, if and how they understand the source material a crucial part of making their translation. Each human being is unique and that holds true for translators as well.

2.1.2. T-universals and S-universals

While researching the third "descriptive" approach Chesterman (2003, 219) gathered existing universals and organized them into two groups of universals. Both focus on differences but they are different according to what texts are being compared.

S-universals are those that can be traced when comparing a translation with its source text. (Chesterman 2003, 218)

T-universals can be found when comparing original texts with translated texts, both in the target language. (Chesterman 2003, 218)

Chesterman (Chesterman 2003, 219) gives the following examples of universals in each group:

Potential S-universals:
- Lengthening: translations tend to be longer than their source texts
(cf. Berman's expansion; also Vinay and Darbelnet 1958: 185; et al.)
- The law of interference (Toury 1995)
- The law of standardization (Toury 1995)
- Dialect normalization (Englund Dimitrova 1997)

- Reduction of complex narrative voices (Taivalkoski 2001)
- The explicitation hypothesis (Blum-Kulka 1986; Klaudy 1996; Øverås 1998) (for example there is more cohesion in translations)
- Sanitization (Kenny 1998) (more conventional collocations)
- The retranslation hypothesis (later translations tend to be closer to the source text; see the special issue of Palimpsestes on 'retranslation': Bensimon 1990)
- Reduction of repetition (Baker 1993)
- S□bj□□tivization (Ch□vali□□1995; s□□b□low)

Potential T-universals:
- Simplification (Laviosa-Braithwaite 1996)
– Less lexical variety
– Lower lexical density
– More use of high-frequency items
- Conventionalization (Baker 1993)
- Untypical lexical patterning (and less stable) (Mauranen 2000)
- Under-representation of TL-specific items (Tirkkonen-Condit 2000)

2.1.3. Normalization

The universal feature that is truly relevant to this thesis is normalization also called sanitization. While Chesterman (Chesterman 2003) placed sanitization with potential S-universals, he does not explain what it is.
Zanettin (2013, 23) describes normalization as follows:

> Normalization, also sometimes referred to as "conventionalization", "standardization", "conservatism" and "sanitization" is the (alleged) tendency of translated texts to conform to target language rather than source language patterns and norms, producing more conventional rather that unusual target strings. Sanitization for instance has been defined as the conservative rendering of creative source language features. Indicators of lexical normalization include degree of lexical and collocational creativity and degree of formality.

While normalization can be observed by comparing source and target texts, the degree of normalization or lack of it can only be assessed when comparing translated texts with originals in the same language.
As the features of Common Czech in this thesis that are being researched are at the lexical level and lower, we need to look at universal features on this level as well.
Differences between Standard Czech and Common Czech are in the degree of formality and register variation. Those are also "indicators of lexical normalization" according to Zanettin (Zanettin 2013, 23).

3. Non-standard Czech

3.1. Stratification of Czech

The Czech language has its own stratification according to social, regional, and function varieties (Krčmová 2017). Krčmová (2017) divides the Czech language into structural varieties and non-structural varieties.
Non-structural varieties are characterized by a specific lexicon. Typical examples are slang, argot, or jargon.
Structural varieties are created by complete language structure. These varieties can be standard or non-standard. Standard varieties are more prestigious and representative. At the top is Standard Czech, followed by Colloquial Czech.
Standard Czech is a regulated variety, which means there are rules of usage made and actualized by the Institute for the Czech Language. This variety is

the most representative and is described in dictionaries, grammar books, and textbooks. It is taught in schools and is often used in both writing and speech. Colloquial Czech is mainly used in a spoken medium and thus is less regulated. Both varieties are quite stable because regulation prevents spontaneous changes.
Non-standard varieties are less prestigious, and they are not regulated at all. They are susceptible to change. For example, the language used by children at the school age differs from generation to generation. The closest to the Standard Czech varieties is Common Czech, followed by Interdialects and Dialects.

3.2. Common Czech

Common Czech has a special place in the stratification of the Czech language. It started as an Interdialect in the Bohemian region. Later, it began to spread to other parts of the country.
Krčmová (2017) states that while Common Czech was used in private conversations at first, it became more mainstream in Prague and then started to be used in media. As it spread and became more popular it lost both regional and social characteristics.
Eventually, it became a separate variety of non-standard Czech, and some linguists, especially in Bohemia, view it as a sub-standard variety quite close to Colloquial Czech. There are even some linguists who say Colloquial Czech is only a bridge between Standard Czech and Common Czech (Sgall 2012, 12).
Common Czech will have an increasingly important place in the Czech language. While dialects and interdialects weaken and slowly disappear, Common Czech does not (Sgall 2012, 17). In current times Common Czech is still a source of controversy among Czech linguists. It is a topic that challenges understanding of the Czech language and it needs more research.
To present typical features of Common Czech, Krčmová (2017) separated them according to language level.

3.3. Features of Common Czech

3.3.1. Syntax

Common Czech can be described as a spontaneous, spoken, and often expressive form of language. It occurs in the form of dialogue. On the level of syntax, Common Czech shares features with any unprepared spoken text.
Krčmová (2017) lists the following examples:

- Repetitiveness
- Ellipse
- Derangement of sentence structure
- Incorrect verb and preposition correspondence
- Non-subject-predicate sentences
- Independent sentence constituents

3.3.2. Lexical level

Words belonging to Common Czech are those that have no regional or social characteristics. They are commonly used in spoken texts.

They are not accepted in Standard Czech because they are of foreign origin or are too expressive. These words usually have Standard Czech synonyms.

- Foreign origin: špitál (a hospital), šponovat (strain, tighten)
- Expressive: tutovka (a sure thing), nalejvárna (taproom)

3.3.3. Phonological level

At this level, thanks to previous research on a case-by-case basis, we can group differences between Standard Czech and Common Czech according to how and where in the word comes to a change.
These changes are phonological and written.

- Change of -é into -í/ý in the inflectional suffix of adjectives
- Change of -é into -í/ý in original Czech word bases
- Change of -ý into -ej in a suffix of singular masculine adjectives in the nominative
- Change of -ý into -ej in a suffix of other adjectives
- Change of -ý into -ej in word bases
- Prothetic v-
- Phonetic shortening of í
- Erasure of syllabic -l
- Phonetic simplification of consonant clusters (erasure, assimilation)

3.3.4. Morphological level

Common Czech at the morphological level can be defined by an absence of certain attributes of Standard Czech. For example, transgressive, passive, or past conditional.
There is a tendency to unify word forms.

- Tendency to move from morphological alternations in the word base
- Coming closer to hard and soft types of words in flection
- Conditional bysme
- Endings unification in the industrial plural with -ma
- No differences of gender in the flection of plural adjectives and pronouns

4. Register and Level of Politeness

The level of politeness is important to both authors and translators. In literature, all dialogues and means of expression have their level of politeness. It helps the reader to be better immersed in the story and fully understand the circumstances of the characters. Even if a reader does not know a character, just like in real life they can infer many things about it from the dialogue used. Native speakers recognize different textual varieties of their mother tongue naturally and that is enhanced by formal schooling (Biber and Conrad 2009,2-3). It is more difficult in a foreign language and should be deliberately studied along with grammar if a person wishes to fully understand the language.
If a translator wishes to do a good job, they must first analyse what text varieties can be found in the source text.
Textual varieties can be seen as broadly as differences between languages, but on a smaller scale, they are the differences between one speaker and another, one text and another (Biber and Conrad 2009, 4).
One such textual variety is register.

Each language has its means of expressing politeness. But generally, the words we use can be sorted into categories according to how, when, where, and with whom we use them. It depends on cultural and situational context (Biber and Conrad 2009, 5).
There are two main different approaches to register and register study.

4.1. Functional Approach

The leading figure in this approach is Douglas Biber.
Biber and Conrad (2005, 6) have described register as follows: "A register is a variety associated with a particular situation of use. The description of a register covers three major components: the situational context, the linguistic features, and the functional relationships between the first two components". It means that items in a register can be described with lexical and grammatical means. They can also be described with properties of the situation in which they exist. It is important to remember that each item is in a specific register for a reason, it is suited for a certain purpose in both linguistic and contextual ways (Biber and Conrad 2009, 6).
According to Argamon (2019, 104), this approach is continuing on through register analysis. The linguists are trying to describe the character of a register in dimensions with a scale of variation (Argamon 2019, 104).
The goal of this research is to study "how to computationally classify texts according to register, as well as how to analyse register characteristics computationally." (Argamon 2019, 101)

4.2. Approach of Systemic Functional Linguistics (SFL)

The SFL approach further develops the notion of situational context by implying that register creates distinctive language varieties for the purpose of communication depending on the context (Argamon 2019, 104)
Register is "determined by the contextual variables of 'field' (the type and domain of social discourse), 'tenor' (the relationship between the speaker and audience, and their relevant social roles), and 'mode' (parameters of textual organization, such as the communication channel and discourse goals)." (Argamon 2019, 104)
This approach is developing further through register synthesis (Argamon 2019, 101). The goal is to generate texts using register, either starting from the meaning itself or translating one register to another while the meaning stays the same (Argamon 2019, 101)

4.3. Register in relation to this study

Register is crucial to Translation studies. According to Matthiessen (Matthiessen, Wang and Ma 2019, 103), it is through register and its analysis that students of translation truly hone their skills. The same could be said about foreign language learning.
Register and level of politeness are tightly related to language stratification. Lowering one's register can very well mean going from Standard Czech, which is polite speech, to Common Czech or slang, which might be considered impolite in certain social groups and situational contexts.
Common Czech is a register, that can have a few sub-registers. Unfortunately, there is not yet enough research done into Common Czech itself, but perhaps that can be changed by using computing methods and register. While this

study only deals only with linguistic features in a quantitative capacity and so cannot be considered register analysis, it would be a way to continue this research further. The analysis still uses register, for example, the sub-register of written Common Czech that is further specified by use in fiction.

5. Methodology

The focus of this study is making a quantitative analysis of the Common Czech features in original and translated texts. This chapter serves to introduce what methods and means will be used to make this analysis.

5.1. Corpus Selection

There are several things that lead to the selection of the right corpus for research. The analysis in this study will compare Standard Czech and Common Czech, which means that the corpus used can be monolingual. Also, there needs to be the possibility to differentiate between original and translated texts.
This study will use a comparable monolingual corpus of Czech, namely InterCorp, to conduct the research.
In selecting a corpus, it is important to ensure that it is representative of the population being studied. InterCorp consists of contemporary Czech literature from a variety of sources, including literary journals, publishers, and online archives. The texts were selected based on their relevance to the study, including the usage of non-standard Czech in both originals and translations from English.
To regulate the size of the language sample the research will be limited to two sub-corpora—one of original Czech fiction and the other of English fiction translated into Czech.
The fiction category was chosen because it is the most suitable for making these comparisons. It uses both Standard Czech and Common Czech, and there are plenty of original and translated texts.
Both sub-corpora were made from InterCorp version 13. The sub-corpus of original Czech texts consists of 19,417,319 tokens. To clarify things, I will refer to this sub-corpus by the name Original fiction. The sub-corpus of translated Czech texts has 32,298,897 tokens. I will refer to this corpus by the name Translated fiction.

While efforts were made to include a diverse range of texts from different genres and authors, the corpus may not be fully representative of all contemporary Czech literature, which could result in biases and inaccuracies in the analysis.

5.2. Data Collection

Data collection involves the selection of texts from the corpus and the extraction of relevant data for analysis. The texts were selected based on their relevance to the study, and the relevant data was extracted from the selected texts, including examples of non-standard Czech, contextual information, and metadata.

The corpus includes an automated search of tokens based on so-called tags. The tag basically describes the token by linguistic means. The search window offers a selection of tags that are often searched, but there is also a possibility to manually insert tags as needed. To complete the analysis both pre-selection and manual tag selection were used.
While the search engine of the corpus automatically identifies all tags that were used to describe a feature of Common Czech being searched, these methods may not capture all instances of Common Czech or accurately identify the type and context of its usage. This could lead to errors and inaccuracies in the data, which could undermine the validity and reliability of the analysis.
To check all tokens of the result for accuracy would be time-consuming. The results will be sorted by word forms and checked. All errors found will be removed from the used examples and subtracted from the numerical results.

5.3. Data Analysis

Data analysis involves using quantitative and qualitative methods to identify patterns and trends in the usage of non-standard Czech in contemporary literature.
Quantitative analysis involves the use of statistical methods to identify the prevalence and usage patterns of non-standard Czech in the corpus. Qualitative analysis, on the other hand, involves examining the specific contexts in which non-standard Czech is used.
This thesis will use quantitative analysis. The research will examine how many hits of the researched token are in the corpus. What is the relative frequency of the token and if the difference between relative frequencies is statistically significant.
The researcher's personal biases and presumptions may have an impact on how the outcomes are interpreted. The qualitative analysis may also not be able to capture the nuances and complexities of non-standard Czech usage in cases where its usage is highly context-dependent and variable.

5.4. Corpus Statistics

Quantitative corpus-based studies rely on a large amount of data. That is one of the reasons why linguists often, but not always, make use of statistics in such studies. Even thou the use of statistics has its problems, it is generally the recommended approach. The human brain is not built to work with complex frequency data in a consistent and dependable way (Jenset 2008, 17). The corpus itself often uses statistical procedures to work with data. The analysis in the following chapter will rely on the usage of frequency, so statistics are necessary. This section serves to introduce the basics of statistics used in corpus-based studies.
Corpus statistics can be sorted into categories.
Descriptive statistics aim to describe and nothing else. In this category is above mentioned frequency. Specifically relative or normalised frequency, that answers the question ‘how often might we assume we will see
the word per x words of running text?’ (McEnery and Hardie 2012, 49). Relative frequency needs a normalisation base, which in this analysis will be 1,000,000. “Normalised frequencies based on ‘occurrences per thousand

words' or, as here, 'occurrences per million words' are the most commonly encountered in the literature; many corpus search tools generate these figures automatically." (McEnery and Hardie 2012, 50). InterCorp does generate relative frequency automatically.
Inferential statistics will also be used in this study. Specifically, the analysis will use a test of significance. "Most things that we want to measure are subject to a certain amount of 'random' fluctuation. We can use significance tests to assess how likely it is that a particular result is a coincidence, due simply to chance." (McEnery and Hardie 2012, 51). If the results prove to be significant, we can assume that there is some kind of pattern, that could be applied even outside of the tested language sample.
While comparing the quantities of varieties of Common Czech in both sub-corpora, I will ensure that the differences are statistically significant. I will use the online tool Sigil to automatically calculate the significance test.
I have decided to research the phonological features of Common Czech listed in section 3.3.3 Phonological level because they are easiest to look for in the corpus. Each feature will be studied separately.

5.5. Limitations

There are several limitations to the methodology used in this study. The corpus used may not be fully representative of all contemporary Czech literature, which could result in biases and inaccuracies in the analysis. Automated methods used for data collection and analysis may not capture all instances of non-standard Czech or accurately identify the type and context of its usage, which could lead to errors and inaccuracies in the data. Even when manually checked for such errors and inaccuracies, the results might not be without fault.
Also, since data interpretation is subjective, the researcher's prejudices and presumptions may play a role.
The use of statistics is also not without issues. Using an online tool for significance testing may be risky unless we know exactly what method and criteria the tool uses for calculation. While the result might not be entirely wrong, there might be a slight deviation. There is also the fact that significance tests might produce a false result even when the cut-off point of significance is 95%, like in this thesis. Researchers try to avoid this by using even higher cut-off points like 99.99% (McEnery and Hardie 2012, 52).
Despite these limitations, this study provides valuable insights into the usage patterns of non-standard Czech in contemporary literature, highlighting its role and significance in the Czech language and culture. By acknowledging these limitations and taking steps to address them, future studies can build upon this methodology to provide even more robust and comprehensive analyses of non-standard Czech usage.

6. Analysis

6.1. Change of -é into -ý

This change can occur in various parts of speech and various parts of the word, but not all these changes are caused by the use of Common Czech. It can also happen in Colloquial speech or to differentiate terminology of different fields, but that change mainly occurs in the base of the word as opposed to Common Czech.
The one significant indicator of Common Czech usage is when the -é to -ý change occurs in the suffix of adjectives.

6.1.1. Inflectional suffix of adjectives

Czech is an inflectional type of language; suffixes of adjectives change depending on the grammatical categories of case, gender, number, and person.

Case	Masculine Gender		Feminine Gender	Neuter Gender
	Animate	Inanimate		
1. nominative	mladý	mladý	mlad**á**	mladé
2. accusative	mladé**ho**	mladé**ho**	mladé	mladé**ho**
3. genitive	mladé**mu**	mladé**mu**	mladé	mladé**mu**
4. dative	mladé**ho**	mladý	mlad**ou**	mladé
5. vocative	mladý	mladý	mlad**á**	mladé
6. locative	mladé**m**	mladé**m**	mladé	mladé**m**
7. instrumental	mladý**m**	mladý**m**	mlad**ou**	mladý**m**

Table 1 - Standard Czech singular adjective flection

Table 1 - Standard Czech singular adjective flection above demonstrates how regular singular adjectives with so-called hard endings -hý, -chý, -ký, -rý, -dý, -tý, or -ný behave in Standard Czech. Every form with orange highlighted -é in the suffix can be changed to a Common Czech variant with -ý in the suffix.
While making a corpus query we must take into consideration the fact that there are Standard Czech forms of adjectives with the suffix -ý. To ensure the results are correct, the Standard Czech variants with the suffix -ý were excluded from the search.
I have divided the researched items into two groups for convenience. The first group is made from adjectives in the singular, and the second group is made from adjectives in the plural.

	Original fiction	Translated fiction
	Hits	
Suffix -é	261 684	429 918
Suffix -ý	13 674	10 299
	i.p.m.	
Suffix -é	13 476.83	13 310.61
Suffix -ý	704.22	318.87

Table 2 – Change from -é into -ý: Inflectional suffix of adjectives in singular. Comparison of original and translated texts.

As we can see in Table 2 – Change from -é into -ý the Standard Czech variant in both sub-corpora is present in similar quantities, with 13 476.83 i.p.m in original texts and 13 310.61 i.p.m. in translated texts. Conversely, the common Czech variant occurs more in original texts with a frequency of 704.22 i.p.m. than in translated texts with 318.87 i.p.m. The difference is statistically significant.

1) Significance test result: $G^2 = 3{,}731.42523***$ - difference is significant at $p<.001$

	word	freq	i.p.m.
1	jinýho	515	26.52
2	jasný	260	13.39
3	celý	241	12.41
4	starýho	198	10.2
5	možný	194	9.99
6	dobrý	154	7.93
7	novýho	136	7
8	starý	131	6.75
9	jiný	123	6.34
10	plný	119	6.13

Table 3 – Change from -é into -ý: Inflectional suffix of adjectives in singular. Original fiction examples.

	word	freq	i.p.m.
1	jinýho	550	17.03
2	starýho	197	6.1
3	celý	190	5.88
4	plný	155	4.8
5	dobrý	145	4.49
6	jasný	126	3.9
7	celým	122	3.78
8	možný	118	3.65
9	starý	102	3.16
10	velký	100	3.1

Table 4 - Change from -é into -ý: Inflectional suffix of adjectives in singular. Translated fiction examples.

When we compare the individual word forms of Common Czech in both sub-corpora, a similar division of relative frequency can be seen. Table 3 and Table 4 show the occurrence of the ten most prevalent Common Czech word forms in both sub-corpora. Out of the ten word forms, there are eight that both sub-corpora share and only two that are different.
All items present in both sub-corpora in the first ten have a higher relative frequency in original texts than in translated texts. That holds when we compare items of equal rank and even when we compare shared items.
In both tables the word *jinýho* (a different one) is ranked first, meaning it is the most used. In Original fiction, this word has a frequency of 26.52 i.p.m. but in Translated fiction, it has only 17.3 i.p.m.

The items in Original fiction ranked one to four have a relative frequency in the double digits and rank ten has a relative frequency 6.13 i.p.m. While in Translated fiction the only item with double digits relative frequency is the first rank and there is quite a large drop in i.p.m. to the second rank, which is 6.1. This shows that when it pertains to the occurrence of the Common Czech -ý in the suffix of adjectives, the difference between the two sub-corpora is not only in the quantity of Common Czech word forms but in their variety and distribution as well.

Case	Masculine Gender		Feminine Gender	Neuter Gender
	Animate	Inanimate		
1. nominative	mlad**í**	mlad**é**	mlad**é**	mlad**á**
2. accusative	mlad**ých**	mlad**ých**	mlad**ých**	mlad**ých**
3. genitive	mlad**ým**	mlad**ým**	mlad**ým**	mlad**ým**
4. dative	mlad**é**	mlad**é**	mlad**é**	mlad**á**
5. vocative	mlad**í**	mlad**é**	mlad**é**	mlad**á**
6. locative	mlad**ých**	mlad**ých**	mlad**ých**	mlad**ých**
7. instrumental	mlad**ými**	mlad**ými**	mlad**ými**	mlad**ými**

Table 5 - Standard Czech plural adjective flection

The table above demonstrates how regular plural adjectives with so-called hard endings -hý, -chý, -ký, -rý, -dý, -tý, or -ný behave in Standard Czech. Every form with orange highlighted -é in the suffix can be changed to a common Czech variant with -ý in the suffix.

	Original fiction	Translated fiction
	Hits	
Suffix -é	66 008	114 064
Suffix -ý	5 429	3 601
	i.p.m.	
Suffix -é	3 399.44	3 531.51
Suffix -ý	279.6	111.49

Table 6 - Change from -é into -ý: Inflectional suffix of adjectives in plural. Comparison of original and translated texts.

The results of the second group, adjectives in the plural, follow the same scheme as the adjectives in the singular but it is clear, that they occur less often.
The only change is that the Standard Czech variant occurs in higher frequency in the translated texts with 3 531.51 i.p.m. than in original texts with 3 399.44 i.p.m. The Common Czech variant in original texts has a frequency of 279.6 i.p.m. In translated texts, the frequency is lower with 111.49 i.p.m.
The difference is also statistically significant.

2) Significance test result: G^2 = 1,881.54725*** - difference is significant at p<.001

rank	word	freq	i.p.m.
1	celý	173	8.91
2	jiný	151	7.78

3	starý	125	6.44
4	nový	114	5.87
5	plný	103	5.31
6	malý	98	5.05
7	různý	93	4.79
8	krásný	80	4.12
9	velký	78	4.02
10	černý	75	3.86

Table 7 - Change from -é into -ý: Inflectional suffix of adjectives in plural. Original fiction examples.

Rank	word	freq	i.p.m.
1	celý	127	3.93
2	plný	105	3.25
3	starý	95	2.94
4	velký	94	2.91
5	zatracený	77	2.38
6	malý	77	2.38
7	jiný	72	2.23
8	nový	72	2.23
9	pěkný	54	1.67
10	bílý	52	1.61

Table 8 - Change from -é into -ý: Inflectional suffix of adjectives in plural. Translated fiction examples.

Table 7 and Table 8 show the first ten ranks of Common Czech adjectives in plural that change in the suffix. Out of the ten word forms, there are seven that both sub-corpora share and only three that are different.
All items present in both sub-corpora in the first ten have a higher relative frequency in original texts than in translated texts. That holds when we compare items of equal rank and even when we compare shared items.
The first rank in the second group is also taken by the same item in both sub-corpora. The word *celý* (whole/entire) has a frequency of 8.91 i.p.m. in original texts. In translated texts the frequency is again lower, 3.93 i.p.m. In comparison with the first group of singular adjectives, the relative frequencies are lower and drop more gradually, but the fact that rank ten in Original fiction has a relative frequency of 3.86 i.p.m., which is more than rank two in Translated fiction, with 3.25 i.p.m., remains the same.
When it comes to change from -é to -ý in inflectional suffixes of adjectives overall relative frequencies show that the Common Czech suffix -ý is used less in translated texts.

6.2. Change from -ý to -ej

6.2.1. Suffix in singular masculine adjectives in the nominative

In singular masculine adjectives in the nominative, the change from standard variant -ý to common variant -ej occurs at the very end of the word.
We can look at Table 1 - Standard Czech singular adjective flection reference. This change can occur in word forms with ending -ý highlighted in green.

	Original fiction	Translated fiction
	Hits	
Suffix -ý	115 808	183 944
Suffix -ej	12 638	9 959
	i.p.m.	
Suffix -ý	5 964.16	5 695.06
Suffix -ej	650.86	308.34

Table 9 – Change from -ý to -ej: Suffix in singular masculine adjectives in the nominative. Comparison of original and translated texts.

Table 9 shows the frequencies of Standard Czech and Common Czech variants of adjectives in both sub-corpora. In both corpora, the i.p.m. of the Standard Czech variants are again similar. Original fiction has slightly more with 5 964.16 i.p.m. and Translated fiction has 5 695.06 i.p.m. The relative frequency of Common Czech variants is 650.86 i.p.m. in Original fiction and 308.34 i.p.m. in Translated fiction.
This difference is statistically significant.

3) Significance test result: $G^2 = 3{,}130.37489***$ - difference is significant at $p<.001$

Rank	word	freq	i.p.m.
1	starej	519	26.73
2	malej	369	19
3	celej	320	16.48
4	mladej	287	14.78
5	jinej	262	13.49
6	dobrej	250	12.88
7	velkej	182	9.37
8	hodnej	170	8.76
9	blbej	163	8.4
10	jedinej	161	8.29

Table 10 - Change from -ý to -ej: Suffix in singular masculine adjectives in the nominative. Original fiction examples.

Rank	word	freq	i.p.m.
1	starej	496	15.36
2	celej	339	10.5
3	dobrej	306	9.47
4	velkej	252	7.8
5	malej	247	7.65
6	jinej	163	5.05
7	hodnej	158	4.89
8	mladej	149	4.61
9	jedinej	147	4.55
10	černej	141	4.37

Table 11 - Change from -ý to -ej: Suffix in singular masculine adjectives in the nominative. Translated fiction examples.

Table 10 and Table 11 show the first ten ranks of singular masculine adjectives in the nominative in Common Czech from both sub-corpora. Out of the ten word forms, there are nine that both sub-corpora share and only one that is different.
All items present in both sub-corpora in the first ten have a higher relative frequency in original texts than in translated texts. That holds when we compare items of equal rank and even when we compare shared items.
Both groups have the same word form in rank one. The word *starej* (old) has a relative frequency of 26.73 i.p.m. in the original texts and 15.36 i.p.m. in the translated texts.
Relative frequencies are quite evenly distributed between ranks and there are no visible abnormalities.

6.2.2. Suffix in other adjectives

This change can occur in other adjectives as well, but it is not as frequent. Because these examples are each in a different case, they have a corresponding suffix ending. That means that the place of the change is in the suffix but not at the very end.
We can look at Table 1 and Table 5 for reference. This change can occur in all word forms with -ý highlighted in blue.

	Original fiction	Translated fiction
	Hits	
Suffix -ý	48 800	76 652
Suffix -ej	1688	885
	i.p.m.	
Suffix -ý	2 513.22	2 373.21
Suffix -ej	86.93	27.4

Table 12 - Change from -ý to -ej: Suffix in other adjectives. Comparison of original and translated texts.

The numbers indicate that both Standard Czech and Common Czech variants occur less in translated fiction. The Standard Czech variants have similar relative frequencies. Original fiction has slightly more with 2 513.22 i.p.m. than Translated fiction with 2 373.21 i.p.m. The relative frequencies of Common Czech variants are 86.93 i.p.m. in Original fiction and 27.4 i.p.m. in Translated fiction.
This difference is statistically significant.

4) Significance test result: $G^2 = 828.30551***$ - difference is significant at $p<.001$

Rank	word	freq	i.p.m.
1	jinejch	29	1.49
2	ženskejm	22	1.13
3	rudejch	20	1.03
4	starejch	20	1.03
5	různejch	19	0.98
6	jinejm	19	0.98

7	jinejma	16	0.82
8	ženskejma	14	0.72
9	dlouhejch	14	0.72
10	malejma	13	0.67

Table 13 - Change from -ý to -ej: Suffix in other adjectives. Original fiction examples.

Rank	word	freq	i.p.m.
1	starejch	20	0.62
2	jinejma	11	0.34
3	starejm	11	0.34
4	velkejma	11	0.34
5	jinejm	10	0.31
6	bílejch	10	0.31
7	velkejch	10	0.31
8	starejma	9	0.28
9	jinejch	9	0.28
10	bílejma	9	0.28

Table 14 - Change from -ý to -ej: Suffix in other adjectives. Translated fiction examples.

Table 13 and Table 14 show the first ten ranks of other adjectives in the Common Czech variant in both sub-corpora. Out of the ten, there are four that both share and six that are different. All word forms are in the plural. It can be deduced that with the exception of singular masculine adjectives in the nominative, other singular adjectives occur in the Common Czech variant quite rarely.

All items present in both sub-corpora in the first ten have a higher relative frequency in original texts than in translated texts. That holds when we compare items of equal rank and even when we compare shared items.

The individual examples shared by both sub-corpora are in different ranks, but we can still compare some of them. The word *jinejm* (to different ones) is in sixth place in Original fiction with 0.98 i.p.m. and fifth in Translated fiction with 0.31 i.p.m., but the fifth and sixth place in both tables have the same relative frequency and are interchangeable. They can be treated as if they are in the same rank.

As the occurrence of other adjectives in the Common Czech variant is rare and none of the relative frequencies are higher than 1.5 i.p.m. there are no visible abnormalities when it comes to their distribution.

6.2.3. Word base

Changes can occur in the base of a word. The problem with research of this change is that there is no way to isolate the group of words in which the change can happen. Therefore, there is no general group of examples to gain data about frequencies in the corpora.

This change cannot be judged as a whole, but if we hold to the premise that relative frequencies of individual examples reflect relative frequencies of the general group, as shown in previous chapters, we can somewhat indicate what would the results for the general group look like.

I researched three examples from each relevant part of speech. The examples chosen are the most frequent word forms in which there was a change to a Common Czech variant.

	Word	Original fiction	Translated fiction
		Hits	
-ý	být	15 360	28 793
-ej	bejt	1 838	1 198
-ý	cítit	8 458	16 101
-ej	cejtit	319	199
-ý	přemýšlet	2 313	5 185
-ej	přemejšlet	169	185
		i.p.m.	
-ý	být	791.05	891.46
-ej	bejt	94.66	37.09
-ý	cítit	435.59	498.5
-ej	cejtit	16.43	6.16
-ý	přemýšlet	119.12	160.53
-ej	přemejšlet	8.7	5.73

Table 15 - Change from -ý to -ej: Word base – verb. Comparison of original and translated texts.

The most frequent verbs with the change from -ý to -ej in the word base that are present in both sub-corpora are *být* (be), *cítit* (smell/feel), and *přemýšlet* (think).
When comparing original and translated texts, the frequencies show that the Standard Czech variant occurs less in original texts than in translations. The Common Czech variant is less frequent in translated texts. That holds true for all three adjectives chosen to represent this change. It is clearly visible in the example *být* (be). The Standard Czech variant has a relative frequency of 791.05 i.p.m. in Original fiction, while it is 891.46 i.p.m. in Translated fiction. The Common Czech variant is more frequent in Original fiction with a relative frequency of 94.66 i.p.m. than in Translated fiction with 37.09 i.p.m. All differences in relative frequencies of Common Czech variants between the sub-corpora are statistically significant.

5) *Bejt* significance test result: $G^2 = 656.10289***$ - difference is significant at $p<.001$
6) *Cejtit* significance test result: $G^2 = 15.25531***$ - difference is significant at $p<.001$
7) *Přemejšlet* significance test result: $G^2 = 122.29664***$ - difference is significant at $p<.001$

	Word	Original fiction	Translated fiction
		Hits	
-ý	sýr	353	635
-ej	sejr	19	9
-ý	mýdlo	334	451
-ej	mejdlo	42	22
-ý	býk	287	571
-ej	bejk	78	98
		i.p.m.	
-ý	sýr	18.18	19.66
-ej	sejr	0.98	0.28
-ý	mýdlo	17.2	13.96
-ej	mejdlo	2.16	0.68
-ý	býk	14.78	17.68
-ej	bejk	4.02	3.03

Table 16 - Change from -ý to -ej: Word base – noun. Comparison of original and translated texts.

The most frequent nouns in which the change is possible are *sýr* (cheese), *mýdlo* (soap/party), and *býk* (a bull).
Common Czech variants of all three nouns occur very little in both sub-corpora. But they are still more frequent in Original fiction. This is clearly visible in the example *mejdlo* (soap/party). In Original fiction, the Common Czech variant has a frequency of 2.16 i.p.m. while in Translated fiction it is only 0.68 i.p.m.
The differences in Common Czech varieties *sejr* (cheese) and *mejdlo* (soap/party) are statistically significant. But the difference of *bejk* (a bull) is not. That means that the relative frequency in both sub-corpora is too similar.

8) *Sejr* significance test result: $X^2 = 9.71633$** - difference is significant at $p<.001$
9) *Mejdlo* significance test result: $X^2 = 20.33827$*** - difference is significant at $p<.001$
10) *Bejk* significance test result: $X^2 = 3.15964$ - difference is not significant

The same research could be applied to adjectives as well, but there are not enough occurrences of adjectives that could be used as a representative of this change in both sub-corpora to provide relevant results. While there are not any examples of adjectives, we can assume that they would be derived from the nouns where this change occurs. The frequencies might therefore be similar. This is only a theory that at present cannot be proven, for the occurrences of both relevant nouns and adjectives in the corpora are too few.

	Original fiction	Translated fiction
	Hits	
-ý	20 711	28 972
-ej	519	357
	i.p.m.	
-ý	1 066.63	897
-ej	26.73	11.05

Table 17 - Change from -ý to -ej: Word base – pronoun. Comparison of original and translated texts.

The pronouns can be put into a general research group. Table 17 clearly shows that pronouns are generally used less in Translated fiction. That holds true for both Standard Czech variants and Common Czech variants. The relative frequency of the Common Czech variants is 26.73 i.p.m. in Original fiction and 11.05 i.p.m. in Translated fiction.
This difference is statistically significant.

11) Significance test result: $G^2 = 168.67933***$ - difference is significant at $p<.001$

Rank	word	freq	i.p.m.
1	svejch	236	12.15
2	svejma	85	4.38
3	mejch	78	4.02
4	svejm	52	2.68
5	tvejch	24	1.24
6	mejma	23	1.19
7	tvejma	11	0.57
8	tvejm	5	0.26
9	mejm	5	0.26

Table 18 - Change from -ý to -ej: Word base – pronoun. Original fiction examples.

Rank	word	freq	i.p.m.
1	svejch	132	4.09
2	svejma	73	2.26
3	mejch	53	1.64
4	svejm	43	1.33
5	tvejch	26	0.81
6	mejma	14	0.43
7	mejm	8	0.25
8	tvejm	4	0.12
9	tvejma	3	0.09

Table 19 - Change from -ý to -ej: Word base – pronoun. Translated fiction examples.

The first six places in both sub-corpora are held by the same word forms. The first place is held by the word form *svejch* (one's own). It is a reflexive pronoun that does not have a clear equivalent in English. The relative

frequency is 12.15 i.p.m. in Original fiction and 4.09 i.p.m. in Translated fiction. The relative frequency is higher in Original fiction.

6.3. Prothetic v-

This change is made by inserting v- at the beginning of a Standard Czech word.
This is another change that cannot be researched in a general group, but only in individual examples. I have again split them according to a part of speech.

	Word	Original fiction	Translated fiction
		Hits	
Standard	oko	24 178	39 486
Prothetic v-	voko	61	47
Standard	okno	10 628	10 984
Prothetic v-	vokno	35	14
Standard	otázka	4 425	8 892
Prothetic v-	votázka	14	21
		i.p.m.	
Standard	oko	1 245.18	1 222.52
Prothetic v-	voko	3.14	1.46
Standard	okno	547.35	340.07
Prothetic v-	vokno	1.8	0.43
Standard	otázka	227.89	275.3
Prothetic v-	votázka	0.72	0.65

Table 20 – Prothetic v-: Noun. Comparison of original and translated texts.

The three most frequent nouns that exist with a prothetic v- in the sub-corpora are *oko* (eye), *okno* (window), and *otázka* (question). The Standard Czech variants are more used in original texts, except for the word *otázka*, which is used more in translated texts. The Common Czech variants are all used less in Translated fiction. The most frequent of the three examples in both sub-corpora is the word *voko* (eye). The relative frequency in Original fiction is 3.14 i.p.m. while in Translated fiction it is only 1.46 i.p.m.
The difference in the frequency of Common Czech words *voko* and *vokno* is statistically significant, but the difference in frequency of the word *votázka* is not.

12) *Voko* significance test result: X^2 = 15.71658*** - difference is significant at p<.001
13) *Vokno* significance test result: X^2 = 22.56669*** - difference is significant at p<.001
14) *Votázka* significance test result: X^2 = 0.01570 - difference is not significant

	Word	Original fiction	Translated fiction
		Hits	
Standard	otevřít	6 553	11 864
Prothetic v-	votevřít	11	9
Standard	odejít	4 490	10 347
Prothetic v-	vodejít	73	66
Standard	objevit	4 765	10 940
Prothetic v-	vobjevit	3	1
		i.p.m.	
Standard	otevřít	337.48	367.32
Prothetic v-	votevřít	0.56	0.28
Standard	odejít	231.24	320.35
Prothetic v-	vodejít	3.76	2.04
Standard	objevit	245.4	338.71
Prothetic v-	vobjevit	0.16	0.03

Table 21 - Prothetic v-: Verb. Comparison of original and translated texts.

The three most frequent verbs in the corpora that can occur with prothetic v- are *otevřít* (open), *odejít* (leave), and *objevit* (discover). All Standard Czech variants are more frequent in Translated fiction. The Common Czech variants are more frequent in Original fiction. The most frequent of the three examples in both sub-corpora is the word *vodejít* (leave). The relative frequency in Original fiction is 3.76 i.p.m. while in Translated fiction it is 2.04 i.p.m.
The only statistically significant difference in Common Czech examples is in the word *vodejít* (leave). The difference in frequency of the words *votevřít* (open) and *vobjevit* (discover) is not statistically significant.

15) *Votevřít* significance test result: $X^2 = 1.90734$ - difference is not significant
16) *Vodejít* significance test result: $X^2 = 12.65709$*** - difference is significant at $p<.001$
17) *Vobjevit* significance test result: $X^2 = 1.06223$ - difference is not significant

	Word	Original fiction	Translated fiction
		Hits	
Standard	ožralý	50	54
Prothetic v-	vožralý	83	17
Standard	ošklivý	696	1 546
Prothetic v-	vošklivý	44	24
Standard	ostrý	1 459	2 316
Prothetic v-	vostrý	31	19
		i.p.m.	
Standard	ožralý	2.58	1.67
Prothetic v-	vožralý	4.28	0.53
Standard	ošklivý	35.84	47.87
Prothetic v-	vošklivý	2.27	0.74
Standard	ostrý	75.14	71.71
Prothetic v-	vostrý	1.6	0.59

Table 22 - Prothetic v-: Adjective. Comparison of original and translated texts.

The addition of the prothetic v- can be seen in adjectives as well. The three most frequent adjectives with this change are *ožralý* (drunk), *ošklivý* (ugly), and *ostrý* (sharp). Both Standard Czech and Common Czech variants are used more in Original fiction. The most frequent example in Original fiction is *vožralý* (drunk) with 4.28 i.p.m., but it is the least frequent in Translated fiction out of the three examples. The most frequent there is the word *vošklivý* (ugly) with 0.74 i.p.m.
The difference in the relative frequency of the Common Czech variants is statistically significant in all three examples.

18) *Vožralý* significance test result: $X^2 = 86.18188$*** - difference is significant at p<.001
19) *Vošklivý* significance test result: $X^2 = 20.24910$*** - difference is significant at p<.001
20) *Vostrý* significance test result: $X^2 = 11.72963$*** - difference is significant at p<.001

	Original fiction	Translated fiction
	Hits	
Standard	26 849	40 628
Prothetic v-	988	365
	i.p.m.	
Standard	1 382.73	1 257.88
Prothetic v-	50.88	11.3

Table 23 - Prothetic v-: Pronoun. Comparison of original and translated texts.

The pronouns can be again put in a general group. The Standard Czech varieties are slightly more frequent in Original fiction. The Common Czech varieties are more frequent in Original fiction as well. The relative frequency is 50.88 i.p.m. in original fiction and 11.3 i.p.m. in translated fiction.
The difference is quite big and statistically significant.

21) Significance test result: $G^2 = 701.67800***$ - difference is significant at $p<.001$

	word	Freq.	i.p.m.
1	vona	409	21.06
2	voni	374	19.26
3	vono	199	10.25
4	vony	6	0.31

Table 24 - Prothetic v-: Pronoun. Original fiction examples.

	word	Freq.	i.p.m.
1	vona	150	4.64
2	voni	116	3.59
3	vono	97	3
4	von	2	0.06

Table 25 - Prothetic v-: Pronoun. Translated fiction examples.

The first three ranks are filled with the same words. The first rank is held by the word *vona* (she) in both sub-corpora. The relative frequency in original texts is 21.06 i.p.m. and in translated texts 4.64 i.p.m.

6.4. Phonetic shortening of vowel -í

The change from a Standard Czech to a Common Czech variant is in shortening the length of the vowel -í into an -i. This change can happen in the base of a word as well as in the suffix.
Because making a general research group is impossible, I once again searched for individual examples.
This change can occur in verbs, nouns, and adjectives. Unfortunately, there were not enough examples for nouns and adjectives in the sub-corpora. Therefore, only the verbs can be compared.
This change cannot occur in the infinitive of the verb, but it occurs in other word forms. I have decided to use the lemma or the infinitive in Table 27 below, so there is no need to see each word form individually.
For a better understanding of non-Czech speaking readers, Table 26 below shows all word forms of a relevant verb with -í according to relevant grammatical categories of person and number in the present tense.

Infinitive		věřit
Singular	1. person	věřím
	2.person	věříš
	3. person	věří
Plural	1. person	věříme
	2.person	věříte
	3. person	věří

Table 26 – Verb word forms

	Word	Original fiction	Translated fiction
		Hits	
Long í	věřit	2 189	3 765
Shortened i		58	2
Long í	vědět	29 935	44 203
Shortened i		948	101
Long í	vidět	8 694	9 529
Shortened i		130	19
		i.p.m.	
Long í	věřit	112.73	116.57
Shortened i		2.99	0.06
Long í	vědět	1 541.66	1 368.56
Shortened i		48.82	3.13
Long í	vidět	447.74	295.03
Shortened i		6.7	0.59

Table 27 – Phonetic shortening of vowel -í: Verb. Comparison of original and translated texts.

The three most frequent examples are the words *věřit* (trust/believe), *vědět* (know), and *vidět* (see). The Standard Czech variant of *věřit* is more frequent in Translated fiction while *vědět* and *vidět* are more frequent in Original fiction. The Common Czech variants of all three are more frequent in Original fiction. The most frequent example is the verb *vědět* (know) with 48.82 i.p.m. in Original fiction and 3.13 i.p.m. in Translated fiction.
All three differences are statistically significant.

22) *Věřit* significance test result: X^2 = 86.93169*** - difference is significant at p<.001
23) *Vědět* significance test result: X^2 = 1,246.14983*** - difference is significant at p<.001
24) *Vidět* significance test result: X^2 = 154.85845*** - difference is significant at p<.001

6.5. Erasure of syllabic -l

Syllabic -l can be erased in masculine verbs where the final -l is preceded by a consonant.

	Original fiction	Translated fiction
	Hits	
Syllabic -l	131 624	251 136
Erasion	7 846	6 553
	i.p.m.	
Syllabic -l	6 778.69	7 775.37
Erasion	404.07	202.89

Table 28 – Erasion of syllabic -l: Comparison of original and translated texts.

The Standard Czech variants are more frequent in Translated fiction. The Common Czech variants on the other hand are more frequent in Original fiction.
The difference is statistically significant with a frequency of 404.07 i.p.m. in original texts and 202.89 i.p.m. in translated texts.

25) Significance test result: G^2 = 1,696.97714*** - difference is significant at p<.001

Rank	word	freq	i.p.m.
1	řek	794	40.89
2	moh	670	34.51
3	nemoh	365	18.8
4	přines	179	9.22
5	neřek	178	9.17
6	vykřik	157	8.09
7	vytáh	120	6.18
8	zahlíd	113	5.82
9	drnk	88	4.53
10	vylez	87	4.48

Table 29 - Erasion of syllabic -l: Original fiction examples.

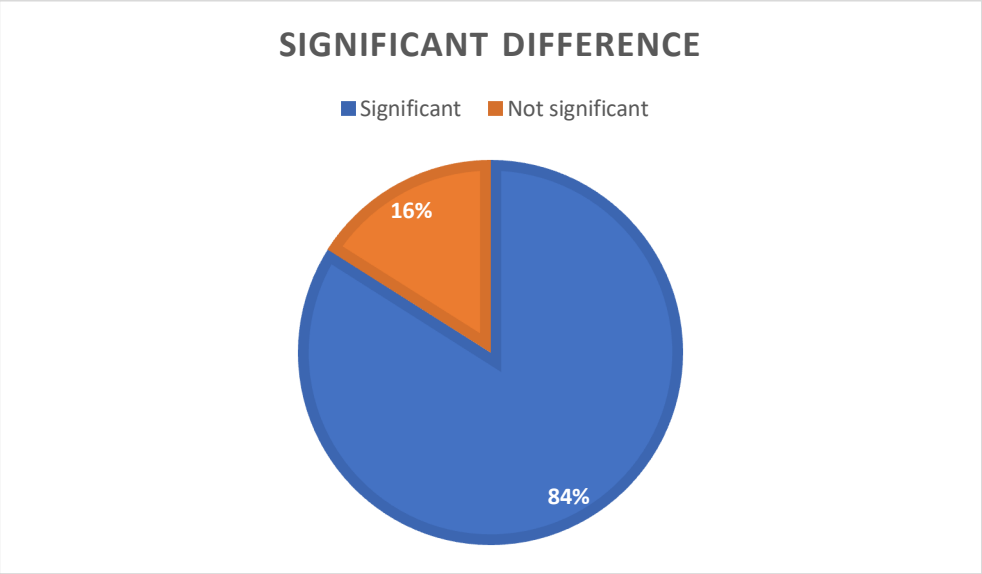
Rank	word	freq	i.p.m.
1	řek	1 135	35.14
2	moh	1 013	31.36
3	nemoh	348	10.77
4	přines	257	7.96
5	neřek	244	7.55
6	vylez	87	2.69
7	proved	76	2.35
8	pomoh	70	2.17
9	utek	70	2.17
10	vlez	68	2.11

Table 30 - Erasion of syllabic -l: Translated fiction examples.

The first five ranks in both sub-corpora are filled with the same words. The first rank is held by the word *řek* (said/told). The frequency is 40.89 i.p.m. in Original fiction and 35.14 i.p.m. in Translated fiction.

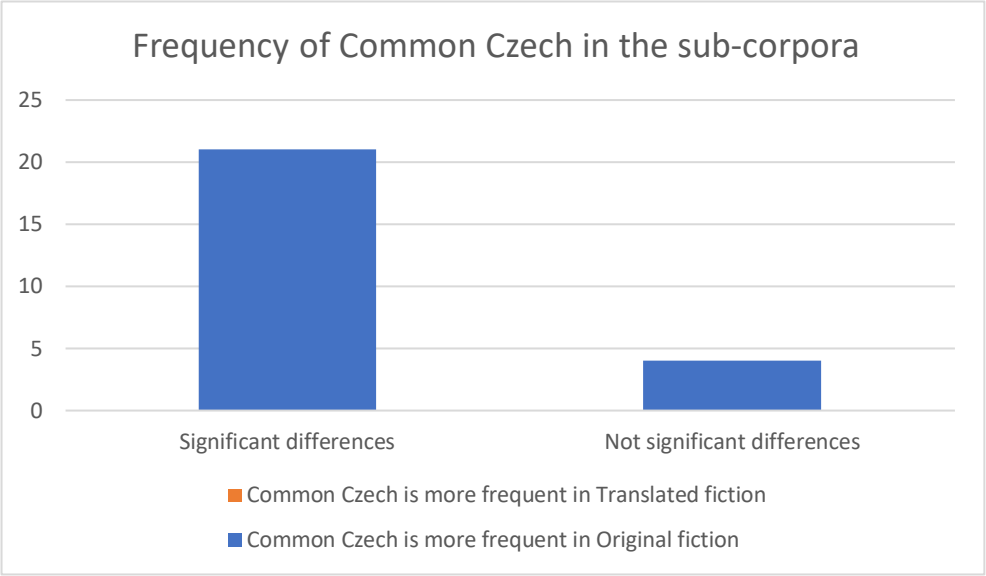
6.6. Summary

I shall finish this analysis by making a general overview of results from all previous chapters. This section will contain graphs and percentages to make certain that the results are clear and understandable.



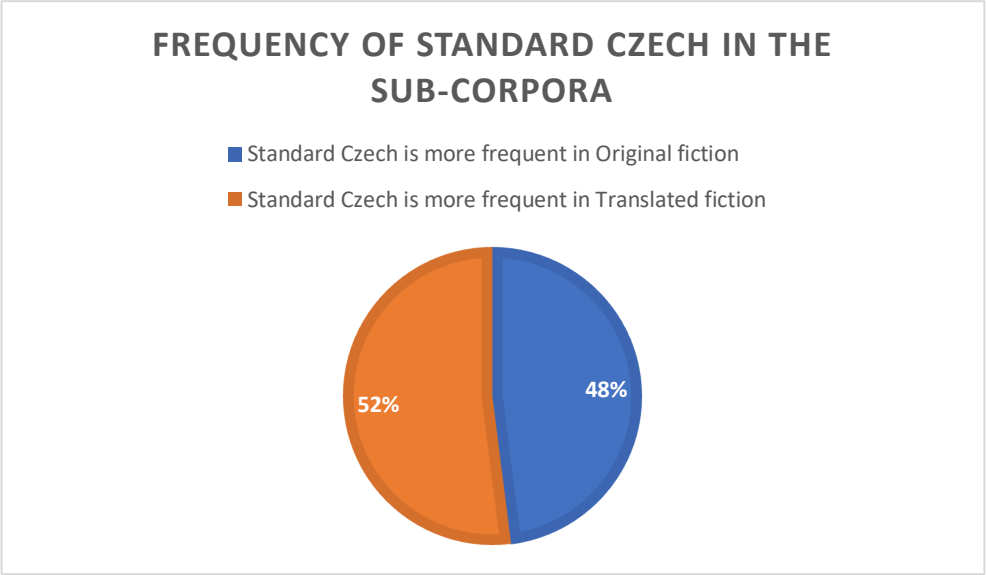
Graph 1 – Significant differences in this analysis.

There are 25 tests of significance of differences between occurrences of specific Common Czech tokens or a group of tokens in two sub-corpora in this analysis. The sub-corpora are Original fiction with original texts written in Czech and Translated fiction with texts translated from English to Czech. Out of them all, 21 tests proved that the difference in the relative frequency of those specific tokens is statistically significant. That is 84%. Significant differences are portrayed in blue in Graph 1 above. It means that it is highly probable that these results were not simply gained by chance, but the is a visible pattern in the use of Common Czech.
There are only 4 tests that proved to be not significant. That is 16%. All of them were tests of a single word. It is possible that if the sub-corpora were enlarged the result might be different. There would need to be generally more data including specific data of this one token. There is of course the possibility that the the token is used relatively similarly and is evenly distributed in original and translated texts alike.



Graph 2 – Frequency of Common Czech tokens in the sub-corpora.

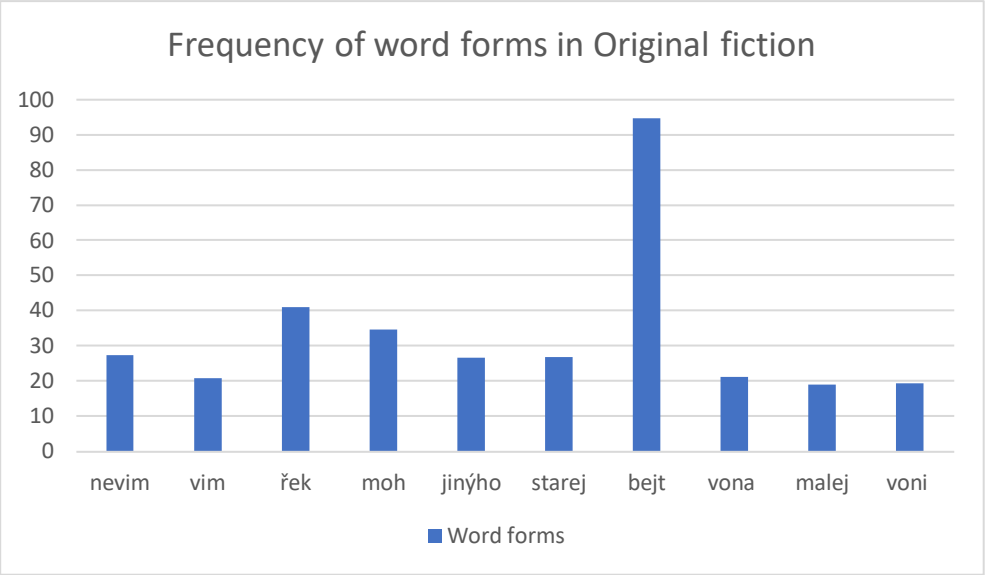
Graph 2 shows that all features of Common Czech are more frequent in the sub-corpus Original fiction. This result was made by viewing the relative frequency of all tokens with normalisation base of 1,000,000. In both categories – tokens where the difference is significant and tokens where the difference is not significant – there is not a single case, where the Common Czech variant is more frequent in translated texts. Results show that translators do not tend to use Common Czech as often as authors.



Graph 3 – Frequency of Standard Czech tokens in the sub-corpora.

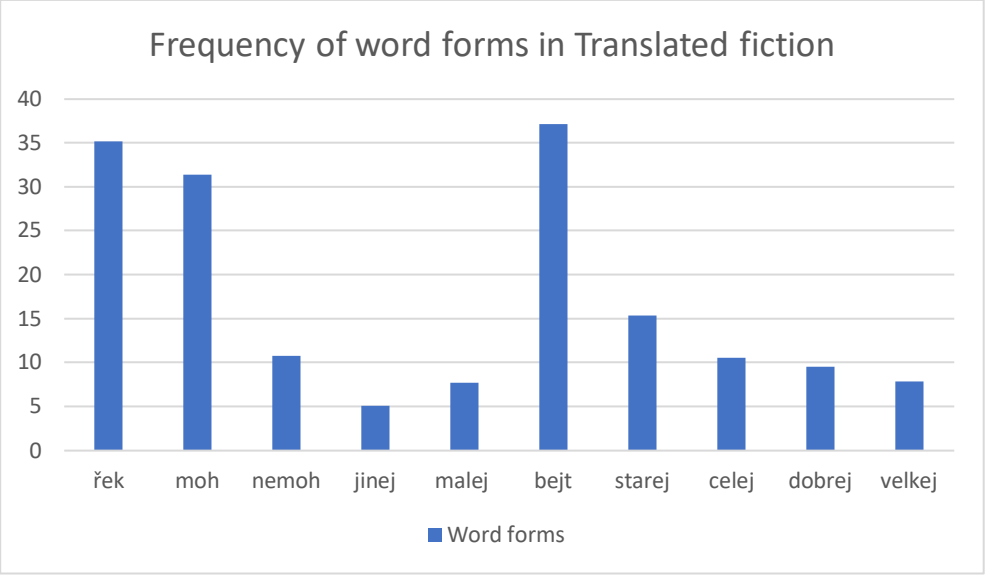
When it comes to Standard Czech variants, they are slightly more frequent in Translated fiction. Graph 3 shows that this holds true for 52% of the cases in this analysis.
It must be mentioned that there was not any significance testing for Standard Czech differences in the sub-corpora. This result would probably change if there were only significant differences.

When we look at the individual examples there is a clear ranking visible for the word forms used in Original fiction.

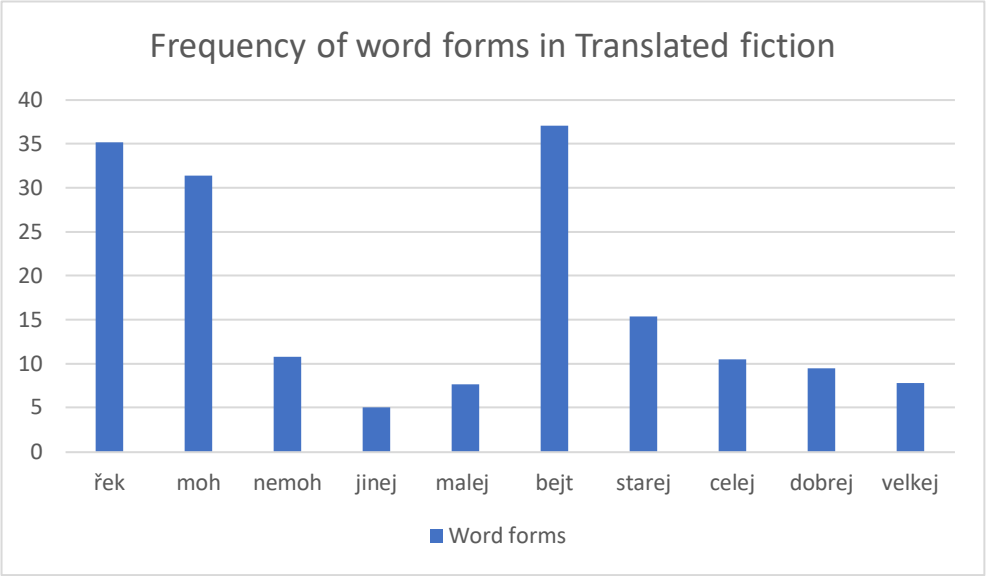


Graph 4 – Frequency of individual word forms in the sub-corpora Original fiction

Graph 4 shows the ten most frequent examples of Common Czech word forms across the researched categories in the sub-corpus Original fiction. It is obvious that the first rank holds the word *bejt* (be). It is not surprising because this is a widely used word in Standard Czech as well. The second rank is held by the word *řek* (said). The third rank Is held by the word *moh* (can/could).
It is worth mentioning that the first four ranks are held by verbs. The second and third ranks are both examples of syllabic -l erasure.
A similar graph can be made for the examples from the sub-corpus Translated fiction.



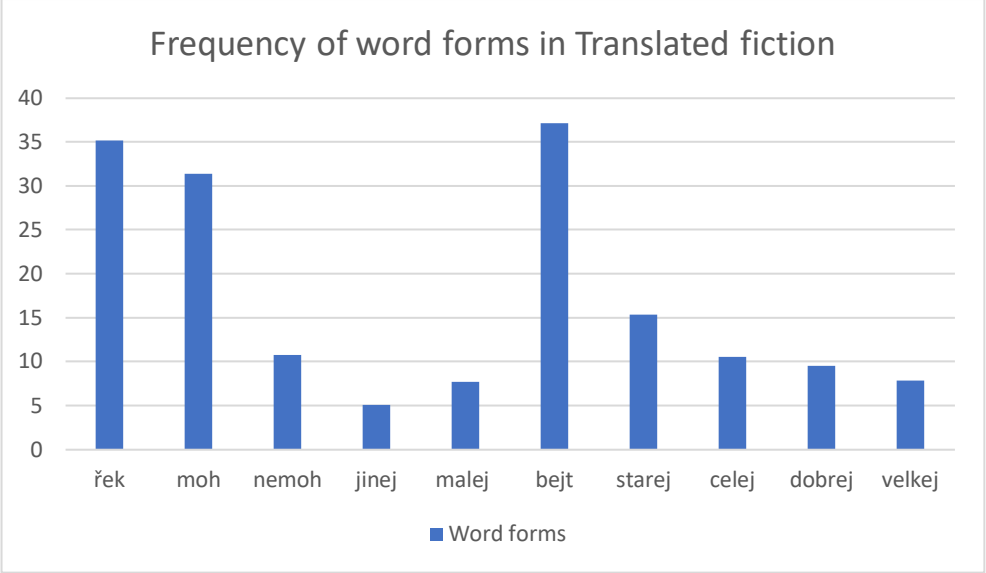
Graph 5 - Frequency of individual word forms in the sub-corpora Translated fiction



Graph 5 shows the ten most frequent examples of Common Czech word forms in the sub-corpus Translated fiction.

The first three ranks are the same as in Original fiction: *bejt* (be), *řek* (said/told), *moh* (can/could). But the difference between the first rank and the second is much less drastic.

While Graph 4 has quite evenly distributed examples of several different features of Common Czech, in



Graph 5 the majority of the top ten ranks are held by adjectives with the change of -ý to -ej. This might indicate that Common Czech is not only used less in translated texts but there are features that are used more than others, leading to uneven distribution.

Graph 1 – Significant differences in this analysis

7. Conclusion

The usage of non-standard Czech in modern literature has become an everyday occurrence, and this linguistic trend can be identified by differences from common vocabulary and grammar conventions. Because this demonstrates the diversity and depth of the language, non-standard Czech is a crucial component of the Czech language and culture.

This research aimed to find out if translators tend to use the method of naturalization with relatively similar or more frequent usage of Common Czech variants in translated texts as in original Czech texts. Or if translators tend to make the translated text more formal and use less Common Czech than original texts written in Czech.

The conclusion is that translators tend to make their target text more formal and use less Common Czech than original Czech texts. This is proven by the results of comparative quantitative analysis in this thesis. 100% of all researched tokens relevant to one of the features of Common Czech are more frequently used in original texts than in translated texts. 84% of the results are statistically significant, which means that the hypothesis of this thesis holds true for the smaller sample of language represented by the sub-corpora, but also it can be assumed to apply more generally. Of course, that cannot be made into a clear claim without empirical evidence.

The analysis of individual examples showed that features of Common Czech in translated texts might not be evenly distributed, and translators tend to one more over the others. It could also imply that they are using more adjectives because of the influence of the English source text.

This study used a corpus-based technique to examine how the non-standard Czech language was employed in both the source texts and the translations from English. The corpus used for analysis was InterCorp. It was further reduced to two sub-corpora, Original fiction featuring texts written in Czech and Translated fiction featuring texts written in English and translated into Czech.

The research was not entirely smooth and there were a few difficulties to overcome. While researching groups of tokens that have a feature of Common Czech there were a few false results, that were identified in the manual check and removed. Not all features of Common Czech could be encapsulated into one group described by a few of the same tags. This problem was solved by taking the most prevalent examples of a certain feature.

To shed light on the function and importance of non-standard Czech in the Czech language and culture, recurring patterns, and trends in its usage in modern literature were able to be identified through the corpus analysis.

This allowed for a deeper comprehension of the use of non-standard Czech in both the originals and translations from English.

The study provides a deeper understanding of the differences between Standard and Common Czech and their usage in different types of texts. It also emphasizes the importance of considering the difference between Standard and Common Czech when translating from or into Czech. This research is important for anyone interested in the differences between Standard and Common Czech, as well as for translators, linguists, and Czech language learners.

It is also important to use non-standard Czech in English translation. Thanks to translators who were able to include non-standard Czech, the characters' language and actions were more rooted in their Czech cultural context. This results in a more natural and seamless reading experience. The difficulties of translating non-standard Czech, however, were also emphasized because there might not always be a precise English equivalent. Translators must carefully evaluate the cultural subtleties and context of both languages while juggling the requirement to remain authentic to the original material with the necessity to adapt it for a Czech audience.

Additionally, this research has shown just how important it is to take non-standard Czech into account when translating from English to Czech because it has a significant effect on the overall tone and authenticity of the translated text.

The significance of non-standard Czech in translations from English is anticipated to grow with the increasing globalization of literature, making this another area of study that deserves additional research. There is still a lot to learn about this topic.

There are several ways for future authors to build on this thesis and this research. The author can focus on Common Czech and whether the translator's place of birth and situational context influence their use of Common Czech in their translations. They can make a complete register analysis of the linguistic features and situational context and try to interpret what function they fulfill together.

They can focus on Standard Czech and whether translators compensate for lesser Common Czech usage with Standard Czech or other means.

They can also research whether these results can be duplicated in different conditions. Those might be a larger sample of texts, a different genre of texts or a different corpus. There is also the possibility to focus more in depth on a single feature of Common Czech or perhaps limit their research on a specific time period. For more experienced authors, there is also the possibility to make their own corpus, which would allow them to fully control the size of the researched sample and all criteria for choosing which texts to include in their research. However, there are legal and ethical considerations, such as fair use and intellectual property rights, privacy and confidentiality, and transparency and accountability, that must be taken into account when building a corpus. There is also the possibility to do similar research on texts from a different medium, like spoken texts.

This thesis successfully reached a conclusion, and all results indicate that our hypothesis was correct. In general, this work extends our knowledge of the function of non-standard Czech in current Czech language use and translation techniques.

Hopefully, it brings new data, results, and a point of view that will be beneficial and helpful to the fields of Translation studies and Linguistics in the future.

8. References[1]

Argamon, Shlomo S. 2019. "Register in computational language research. " In *Register Studies* 1:1 (2019), pp. 100–135. URL: https://doi.org/10.1075/rs.18015.arg

Baroni, Marco, and Stefan Evert. 2008. "Statistical methods for corpus exploitation. " In *A. Lüdeling and M. Kytö (eds.), Corpus Linguistics. An International Handbook*, article 36. Mouton de Gruyter, Berlin. URL: https://www.stephanie-evert.de/PUB/BaroniEvertHSK38_manuscript.pdf

Biber, Douglas, and Susan Conrad. 2009. *Register, Genre and Style.* Cambridge: Cambridge University Press.

Chesterman, Andrew. 2003. "Contrastive textlinguistics and translation universals." In *Contrastive Analysis in Language: Identifying Linguistic Units of Comparison*, edited by Dirk Willems, Bart Defrancq, Tom Colleman, and Dominique Noël, 213-229. London: Palgrave MacMillan.

Evert, Stefan. 2007. "Statistics tutorial. " In *ICAME Conference 2007.* Stratford upon Avon. URL: https://www.stephanie-evert.de/PUB/Handout_ICAME_Statistics.pdf

Gammelgaard, Karen. 1999. "Common Czech in Czech Linguistics." *Slavonica*, 5:2, 32-51. DOI: 10.1179/sla.1999.5.2.32

Jenset, Gard B. 2008. "Basic statictics for corpus linguistics. " In *Research Gate*. DOI:10.13140/2.1.1684.6084

Matthiessen, Christian M.I.M., Bo Wang and Yuanyi Ma. 2019. "Expounding register and registerial cartography in systemic functional linguistics: an interview with Christian M.I.M. Matthiessen. " In *WORD*, 65:2, 93-106, DOI: 10.1080/00437956.2019.1599544

McEnery, Tony, and Andrew Hardie. 2012. *Corpus linguistics.* Cambridge: Cambridge University Press.

Krčmová, Marie. 2017. "OBECNÁ ČEŠTINA. " In *Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - New encyclopedic dictionary of Czech.* URL: https://www.czechency.org/slovnik/OBECNÁ ČEŠTINA

Krčmová, Marie, and Jan Chloupek. 2017. "NÁRODNÍ JAZYK." In *Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - New encyclopedic dictionary of Czech.* URL: https://www.czechency.org/slovnik/NATIONAL LANGUAGE

Sgall, Petr. 2012. "Obecná Čeština. " In *Linguistica online.* URL: http://www.phil.muni.cz/linguistica/art/sgall/sga-001.pdf

Skřivánková, Jana, 2009. "Skloňování přídavných jmen." In *Moje čeština.* URL: https://www.mojecestina.cz/article/2009092802-sklonovani-pridavnych-jmen

[1]Note to translation: All Czech texts used in this thesis do not have a published translation. All translation used in quotes and paraphrases are mine.

Zanettin, Federico. 2013. "Corpus Methods for Descriptive Translation Studies." In *Procedia - Social and Behavioral Sciences* 95: 20-32. ISSN 1877-0428.

9. Other sources

Rosen, A., M. Vavřín and A. J. Zasina. 2020. Korpus InterCorp – čeština, verze 13 z 1. 11. 2020. Ústav Českého národního korpusu FF UK, Praha. URL: WWW http://www.korpus.cz

Sigil: Corpus Frequency Test Wizard. URL: http://sigil.collocations.de/wizard.html