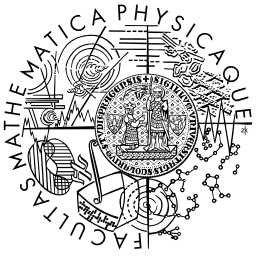


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Gabriela Závodská

Analyzing differences in alternative translations

Institute of Formal and Applied Linguistics

Supervisor of the bachelor thesis: Mgr. Martin Popel, Ph.D.
Study programme: Computer Science

Prague 2024

I declare that I carried out this bachelor thesis on my own, 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 ....

Author's signature

I would like to thank my supervisor, Mgr. Martin Popel, Ph.D., without whose mentorship and patience this thesis would not have come to life. I would also like to express my gratitude to my parents and friends, who supported me throughout my life. A special thank you goes to my partner for their supportive hugs and for believing in me even when I didn't. Last but not least, I would like to thank my therapist for helping me find motivation to finish this thesis.

Title: Analyzing differences in alternative translations

Author: Gabriela Závodská

Institute: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Martin Popel, Ph.D., Institute of Formal and Applied Linguistics

Abstract: This thesis aims to find differences between optimal reference translations (ORT) and standard reference translations (SRT) using the MCC tool developed for their comparison. Various metrics based on morphology, syntax, language models and others are compared for these alternative translations and evaluated using difference and t-test. The ones that contrast enough between the translations are further analyzed using examples. We found a significantly higher use of active voice in the ORTs, substituted by passive voice in the SRTs. A higher syntactic complexity (average number of subordinate clauses) is found in ORT, which we mainly attribute to a higher proportion of adnominal clauses in this translation. Lastly, we included a length comparison of an ORT with two selected machine translations. We found that the ORT is longer in both cases on the document and the segment level, which are shorter mostly because of completely omitting translations of whole phrases or even sentences.

Keywords: translation, optimal reference translation, machine translation, comparison, text analysis

Název práce: Analýza rozdílů v alternativních překladech

Autor: Gabriela Závodská

Ústav: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: Mgr. Martin Popel, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Cieľom tejto práce je nájsť rozdiely medzi optimálnymi referenčnými prekladmi (ORT) a štandardnými referenčnými prekladmi (SRT) pomocou nástroja MCC vyvinutého na ich porovnanie. Pre tieto alternatívne preklady sa porovnávajú rôzne metriky založené na morfológii, syntaxi, jazykových modeloch a iných a vyhodnocujú sa pomocou rozdielu a t-testu. Tie, ktoré sú medzi prekladmi dostatočne kontrastné, sú ďalej analyzované pomocou príkladov. Zistili sme výrazne vyššie využitie priamej reči v ORT, ktorá je v SRT nahradená nepriamou rečou. Vyššiu syntaktickú náročnosť (priemerný počet vedľajších viet) nachádzame v ORT, čo pripisujeme najmä vyššiemu podielu vedľajších viet rozvýjajúcich podstatné meno v tomto preklade. Nakoniec sme zahrnuli porovnanie dĺžky ORT s dvoma vybranými strojovými prekladmi. Zistili sme, že ORT je dlhší v oboch prípadoch na úrovni dokumentu aj na úrovni segmentov, ktoré sú kratšie väčšinou z dôvodu úplného vynechania prekladov celých súvetí alebo dokonca viet.

Klíčová slova: překlad, optimální referenční překlad, strojový překlad, srovnání, analýza textu

Contents

Introduction

When talking about translations, we distinguish various basic approaches – machine and standard professional (human) translations being two of them. However, the possibility that reference-based machine translation evaluation (where the reference translations are standard professional translations) is reaching the end of its validity as some machine translations surpass human translations, gave rise to the creation of so-called optimal reference translations (Zouhar et al., 2024). Standard professional translations will be referred to as standard reference translations, given they are the ones typical for reference-based machine translation evaluation.

The goal here is to automatically find systematic differences and analyze them primarily between the standard reference translations (SRTs) and the optimal reference translations (ORTs), additionally, to a limited extent, between ORTs and machine translations (MTs) from selected systems. The direction of the translations is from English to Czech.

First, the used dataset will be described — the translation types, how they were obtained, and their division —, since its existence is the main motivation for this thesis. In the chapter on related work, the approaches applied in comparing human and machine translation in different language pairs will be examined because some of the methods used in these works are used as inspiration in these experiments. Further on, the exact metrics developed and used for the experiments will be explained in detail.

The results of the experiments are inspected to identify differences between distinct translation types. The experiments typically consist of computing the metrics for the compared alternative translations, comparing their differences on the document level, and performing a t-test on the segment level. A tool was developed in Python to perform these experiments. For a more in-depth analysis of metrics that give interesting results, concrete examples from the dataset are inspected. On some occasions, the Pearson correlation coefficient of the differences in selected metrics is computed to explore to what extent, if any, they cause a difference between two alternative translations together.

The approach presented in this thesis has an advantage in the format of the metrics used. While usually, comparison-based researches of translations focus on a few selected metrics, this thesis implements an extensive range of metrics focusing on various linguistic aspects, mainly based on the Universal Dependencies. Except for two metrics, one measuring the untranslated tokens and the other measuring nominalizations, they can be used to compare translations in other language pairs.

Grammarly[1] was used to correct and rephrase sentences, and DeepL[2] was used to help with translations of some examples from Czech to English.

[1] `https://app.grammarly.com/`

[2] `https://www.deepl.com/en/translator`

1 Motivation and Data

The motivation to look for differences between different types of translation arose with the emergence of the so-called optimal reference translations. The main intention was to compare them with standard reference translations only. However, given their availability, machine translations were included too.

The dataset consists of 51 selected articles from the Fifth Conference on Machine Translation (WMT20) news texts.[1] For each article, it contains the source text (English), three standard reference translations, and two optimal reference translations (Czech) (Kloudová, Mraček, et al., 2023b). Two selected machine translations (Czech) were added to this set from the mentioned WMT20 dataset. Therefore, the final set consists of one source text and eight different translations for each of the 51 selected articles. All of these are segment-aligned, with the segments typically corresponding to sentences, although there are occasional deviations due to some sentences being translated into multiple ones or vice versa. In total, the dataset encloses 579 segments.

Example 1: Example of the same segment from the original and each translation (`sent_id=163`)

source	”That is perhaps a reason why e-cigarette use remains relatively low in Scotland.
N1	„To je také možná důvodem, proč se elektronické cigarety ve Skotsku tolik nepoužívají.“
N2	„To je možná jeden z důvodů, proč nejsou elektronické cigarety ve Skotsku tolik rozšířené.“
P1	„To je možná důvod, proč je používání e-cigaret ve Skotsku relativně málo rozšířené.“
P2	”Toto je možná důvod, proč je používání e-cigaret ve Skotsku poměrně málo rozešířené.
P3	“To je možná důvod, proč se e-cigarety ve Skotsku tolik nepoužívají.“
MT1	„To je možná důvod, proč je používání elektronických cigaret ve Skotsku stále relativně nízké.
MT2	„To je možná důvod, proč je používání e-cigaret ve Skotsku relativně nízké.

Out of these, a test set consisting of 6 articles (1.76 %) with 68 segments (11.74 %) is created by extracting the last six articles. Its purpose is to facilitate the testing of a possible classifier in the future.

We denote by the term *document* the set of the remaining training articles corresponding to a single translation.

1.1 Standard Reference Translations (SRTs)

A standard reference or professional translation is a translation created by a human translator (Zouhar et al., 2024).

The dataset includes three SRTs labelled P1, P2 and P3. P1, created by a professional from a translation agency, was included in the original WMT2020 dataset, serving as the reference human translation created to evaluate machine

[1] `https://statmt.org/wmt20/`

translations in the WMT shared task. Translations P2 and P3 were also created by translators from a translation agency for further research on reference translations. For all these translations, the requirement was to avoid using machine translation, but Kloudová, Bojar, et al. (2021) confirmed the suspicion that P1 is a post-edited machine translation by a series of automatic and manual analyses.

Given the goal of identifying systematic differences between standard and optimal reference translations, we chose to concatenate all three SRTs into a single file, which will be further referenced as SRT.

1.2 Optimal Reference Translations (ORTs)

An optimal reference translation, as defined by Zouhar et al. (2024), is a translation that stems from multiple (at least two) SRTs, and the process of its creation consists of a team of linguists and translatologists choosing the best solution from the available SRTs. However, due to the low quality of the SRTs P1, P2 and P3, the team had to apply a different strategy, as Kloudová, Mraček, et al. (2023a) describe: new translations were created, revised, discussed and edited until a mutual understanding of the optimality of the result was reached. It is important to note that optimality refers to the quality of the outcome and the methodology to achieve it, with no consideration of its time complexity.

The two resulting translations, called N1 and N2, differ in the following (Kloudová, Mraček, et al., 2023b):

- N1 is truer and closer to the source in meaning and language, especially syntax. It also consistently adds the '-ová' suffix to female surnames.
- N2 is looser, more idiomatic and fluent, more true to the Czech reporting style, with the use of explanations and localizations.

However, N2 has not been created for each segment; in some cases, the authors felt it would result in an excessively artificial form. Therefore, in the data used, the missing segments of N2 are substituted by the corresponding segments from N1.

In pursuit of the thesis objective, only N1 will be utilized for comparison with standard reference translations, as it is closer to the source and more universal. For this purpose, a file was created by repeating N1 three times, resulting in a file of the same length and aligned with SRT, further referred to as ORT.

1.3 Machine Translations (MTs)

The selected machine translations from the WMT20 set are from the following systems:

- CUNI-DocTransformer (MT1)
- Online-B (MT2)

CUNI-DocTransformer is a system developed by Popel (2020) at Charles University for the WMT20 shared task. As the name suggests, this system works

on a document level, creating a translation for multiple sentences at once. The official results of the shared task show that DocTransformer outperformed all the other systems, thus surpassing the baseline T2T-2018 system created the previous year (Barrault et al., 2020). It is a representation of a constrained system. On the other hand, Online-B is the best-performing one of the unconstrained systems, meaning it was trained on data different from those provided by the shared task. It is a freely available online system.

To obtain and format this data (ORTs, SRTs and MTs) run the following command from the tool folder:

```
make data
```

2 Related work

In this chapter, we describe the three articles that most significantly influenced the metrics and approaches applied in this thesis.

Bestgen (2021) compares neural machine translations and human translations of news texts translated from French to English. The focus is on a single feature, the bigrams of words that frequently occur together but do not necessarily depend on each other. The examples provided are "dramatic increase, depend on, out of" (Bestgen, 2021). The proportion of this type of bigrams to the total number of bigrams was used for evaluation via a t-test with a level of significance < 0.001. This evaluation inspired the use of the t-test in this thesis for the analysis of metric significance.

From the work of Evert and Neumann (2017), which focuses on the comparison of English and German translations with the source (in both directions), the inspiration for the morphological metrics (further described in Section 3.1) was drawn. They apply metrics such as coordinating conjunctions per number of tokens, nouns per number of tokens or adjectives per number of tokens, which motivated the use of all part-of-speech (POS) tags described in the Universal Dependenciess (UDs). Furthermore, the use of passive voice per number of verbs, colloquialisms per number of tokens, finite verbs per number of sentences and others inspired the inspection of UD's Features. The use of type-token ratio (TTR) and lexical density was also retrieved from here.

As opposed to this thesis, Evert and Neumann (2017) apply a multivariate approach to analyzing the metrics to identify "systematic properties of text"(Evert and Neumann, 2017), which are hard to observe based on a single metric. We try and keep this in mind during analysis and attempt to find connections between metrics.

In the study of Liu and Afzaal (2021), texts translated from Chinese to English are compared from a syntactic point of view with non-translated ones written originally in English. They evaluate all of the metrics separately, using p-values among others. Some of these metrics served as inspiration for this thesis. What we define as syntactic complexity is inspired by the metric of "clauses per sentence"(Liu and Afzaal, 2021). The measures focusing on the amount of coordination and subordination led to measuring the proportions of each of the Dependency Relations defined by UD.

3 Metrics

This thesis aims to identify differences between translations on the level of various metrics as defined here. The motivation behind their use stems from prior research findings for different language pairs[1] or from a belief in their potential utility. This chapter will categorise the metrics according to their linguistic roles: morphological, syntactic, language model-based, and others. In the subsequent chapter, they will be classified according to their way of calculation – measuring for words or segments.

Before computing the metrics, the input texts are preprocessed. Preprocessing consists of tokenization, tagging and parsing using UDPipe version 2.4 (Straka and Straková, 2017), which comes with the Udapi[2] Python release. For the source English text, the `english-ewt-ud-2.4-190531` model is used and for the Czech translations, the `czech-pdt-ud-2.4-190531` model. The segments, on occasion, consist of multiple sentences, but the parser treats each segment as a sentence. Therefore, each segment is assigned only one root, and the relations between the sentences are left for the parser to determine. The input texts are thus converted to a CoNLL-U format,[3] which is then saved to a .conllu file with the same name to speed up later computations. The tool prioritizes locating the input file with a .conllu suffix, using the .txt suffix only if the former is unavailable.

The metrics are computed for given units: segments, articles, or the whole document. A simple difference and paired t-test were used to compare the results. The difference captures which translation has a higher value of the metric, while the t-test compares the similarity of the metrics for two alternative translations on the segment level and in selected cases on the article level. Metrics, where the difference is 0, or the p-value is `nan` — which can happen when the sentence-level results are equal — are ignored.

3.1 Morphological Metrics

Morphological metrics inspect the morphosyntactic properties of the words. Since we are working with Universal Dependencies' CoNLL-U file format, these are stored in the `UPOS` column for Universal part-of-speech (POS) tags and the `FEATS` column for Universal features. The measure of lexical density combines selected POS tags, while the type-token ratio focuses on the lemmas of the tokens.

Part-of-speech Tags

The part-of-speech tags correspond to the Universal POS tags[4] used in UD's CoNLL-U file format. Seventeen part-of-speech categories are distinguished and they differ from the traditional ten Czech POS tags, for example, by distinguishing proper nouns from nouns, subordinating and coordinating conjunctions, or including the additional `PUNCT` tag to mark punctuation. The number of occurrences

[1] mentioned in the previous chapter

[2] `https://github.com/udapi/udapi-python`

[3] `https://universaldependencies.org/format.html`

[4] `https://universaldependencies.org/u/pos/index.html`

within a unit is calculated for each of them.

Lexical Density

Lexical density aims to measure lexical complexity and the informativeness of the text. It is calculated as the ratio of lexical (or auto-semantic) words to all the words in a segment, which is then multiplied by 100 to get percentages. Lexical words, in this case, are the ones that carry one of the following POS tags: ADV, ADJ, NOUN, PROPN, VERB (Johansson, 2008).

The motivation behind its use comes from simplification being seen as a translation characteristic (Baker et al., 1993) on lexical, as well as on syntactic and semantic levels. Therefore, texts with higher lexical density could be closer to the natural language or possibly could play a role in distinguishing the translation types. It could also reflect the limited vocabulary of the human translators or their intention to provide easier readability and interpretation of the translations.

Features

For the feature-based metrics, we examine the Universal Features[5] and their Values, which are contained in the `FEATS` column of the CoNLL-U file format. They serve to distinguish lexical and inflectional properties of words beyond the part-of-speech category, which is addressed in a separate column discussed previously. The format of this column consists of Feature=Value pairs separated by the '|' character. For each Featur=Value pair, we compute its number of occurrences in a given unit and divide it by the total number of occurrences of the Feature (with any Value) within the unit.

Some Features are Boolean, only being listed when their value is Yes, for example, the Reflex feature, which denotes reflexive pronouns or determiners, or Poss for marking possessive pronouns, determiners or adjectives. When the ratio of the number of this Value to the number of corresponding Feature is computed, we always obtain 1; therefore, if a given Boolean Feature is present in the whole collection of articles for two alternate translations, the difference will be $1 - 1 = 0$ and thus such features will omitted in our results. This raises a topic for future work, where the ratio of these Features to the total number of words can be examined.

Other Features have multiple predefined values, e.g. Gender with possible values Com, Fem, Masc, and Neut, or Tense with values Fut, Imp, Past, Pqp, and Pres. This thesis will focus on examining these.

Originally, the focus was only on the Voice Feature, however, to broaden the extent of the experiments and to utilize the Udapi tool, the decision to examine all the Features was made.

Nominalizations

Nominalization is the process of creating nouns from words that are not nouns, usually verbs and adjectives (Comrie and Thompson, 2007). For this metric, we calculate the ratio between the words that carry the NOUN POS tag and result

[5] `https://universaldependencies.org/u/feat/index.html`

from nominalization and all words in the unit. The detection of nominalization is a language-specific task. Therefore, we need to look at how these nouns are formed in each of the languages we work with.

In English, nominalization can be achieved, among others, by adding suffixes to verbs and adjectives, e.g. *develop* → *develop**ment***, *major* → *major**ity***, or *read* → *read**ing*** (as in '*Reading* is fundamental.'). Some English verbs can function as nouns even without morphological changes; for example, the verb *sleep* ('*Sleep* is for the weak.').

To detect whether a noun has been created by nominalization, it is checked whether the noun's lemma ends with any of the following suffixes: -ing, -ion, -ity, -ment, or -ness. This approach is not completely foolproof since there are nouns with these endings, which are not created by nominalization, e.g. thing, university, or business.

In Czech, nominalization is also achieved by suffixation or using the infinite form of the verb as a noun, e.g. "*Číst* je důležité." ('*Reading is important*'). The possible suffixes are -ní/-tí for a verbal noun (*povolit* → *povolen**ní*** ('*permission*'), *pobídnout* → *pobídnu**tí*** ('*encouragement*')) or -ot/-ba/-ka/-ce/etc. for an event noun (*vázat* → *vaz**ba*** ('*binding*'), *jásat* → *jas**ot*** '*cheering*') (Kralík, 2017).

When detecting the nominalizations in Czech, the focus is only on the verbal nouns; the latter-mentioned event nouns are not considered due to the fact that they are less naturally substituted for verbs. The list of the verbal nouns was obtained from the DeriNet word formation network (Vidra et al., 2021) via the DeriSearch[6] tool by inputting the following query: `[pos="VERB"][pos="NOUN" and lemma=".*[tn]í"]`. This list is used to avoid false positives, i.e. nouns that end with the suffix -ní or -tí, but are not created from verbs (e.g. příjmení ('*surname*'), Pobaltí ('*the Baltic*')).

It is important to note that the nominalizations identified by the different approaches for each language may not correspond one-to-one. However, since no comparison with the source text is performed, this will remain the case for the duration of this thesis. There is a possibility of devising a more complex metric in the future that would match the nominalization in one language to the corresponding form used in the other, thereby allowing a comparison to determine whether it was kept.

Example 1: Nominalization
I **forgive** you.
'***Odpouštím** ti.*'

Accept my **forgiveness**.
'*Přijmi mé **odpuštění**.*'

The use of nominalization contributes to a more formal text tone. Example 1 captures this with English and Czech sentence pairs, where, in the first pair, a verb is used, and in the second, a noun that was obtained by nominalization. If this metric yields significantly lower results in one of the alternate translations than the other, there is a possibility that the translation will be less formal and, therefore, show signs of simplification.

[6] `https://ufal.mff.cuni.cz/~vidra/derinet/search/index.xhtml`

Type-token Ratio

Type-token ratio (TTR), as Kyle (2020) defines it and as its name suggests, is computed as $\frac{number\ of\ types}{number\ of\ tokens}$, that is the total number of different words divided by the total number of words. In this thesis, we additionally multiply it to work with percentages. Different types can be considered unique word forms or lemmas in the text. Both versions — computed on lemmas and on forms — of the type-token ratio are be implemented. This metric computes the type-token ratio of the unit by performing a union of all sets of types from the segments of the unit and dividing its size by the total number of tokens.

If one wanted to compute a paired t-test for the TTR metric, one would struggle to find the appropriate corresponding sentence level result, as the TTRs obtained for the unit's sentences capture a wider range of types and have a different nature. Consequently, no t-tests will be performed for this metric.

3.2 Syntactic Metrics

The following metrics evaluate the syntactic structure and complexity of the segments. Only a few syntactic metrics were implemented for this thesis, so the list is not extensive, but it covers basic aspects. They were designed to be computed on sentences since the initial assumption was that the available data was sentence-aligned. However, even though it turned out to be segment-aligned, we kept these metrics and let the UDPipe parser struggle with creating the sentence-like structure.

The syntactic metrics examine basic concepts, like tree depth, tree width, and syntactic complexity, and also inspect the `DEPREL` column of the CoNLL-U file format.

Tree Depth and Tree Width

The tree depth metric measures the depth of the segment's dependency tree. To obtain this value, Recursive Depth-First Search is used.

The tree width metric measures the width of the dependency tree, which corresponds to the maximum number of nodes at any level of the tree. It is obtained via Breadth-First Search.

Dependency Relations

The dependency relations correspond to the Universal Dependency Relations.[7] The root relation points to the root of the segment from an artificially created `ROOT` node, which has an index of 0. There is exactly one root relation for each segment. When performing a computation for a given unit, the number of occurrences of each relation is obtained.

[7] `https://universaldependencies.org/u/dep/index.html`

Syntactic Complexity

Liu and Afzaal (2021) define the overall sentence complexity of a sentence as the number of clauses within that sentence. In this thesis, it will be called syntactic complexity, it will be computed for segments, and a clause will be identified by one of the following dependency relations: root, csubj, ccomp, xcomp, advcl, acl and their subcategories (e.g. acl:relcl, csubj:pass).

Similarly to lexical density, a lower syntactic complexity is a sign of a simpler sentence. Liu and Afzaal (2021) found that the syntactic complexity is higher for original English texts than for English texts translated from Chinese; furthermore, it plays a main role in distinguishing these two types of texts. It can be examined whether it has a similar effect also on the different translation types examined in this thesis.

3.3 Language Model-based Metrics

For the language model-based metrics, a KenLM[8] 2-, 3-, and 4-gram model for Czech was trained. KenLM applies the modified Kneser-Ney smoothing (Ney et al., 1995) to obtain an n-gram language model, not requiring large amounts of RAM. Its speed, low memory requirements, and built-in probability and perplexity calculations for querying were the main reasons behind choosing this toolkit.

All models were trained on the same input, the 2019 version of the News crawl corpus.[9] This dataset consists of news articles originally written in Czech, so querying a language model trained on these data will also hint at the similarity, or its lack, between authentic and translated texts. Before training, the corpus was tokenized using UDPipe 2.4, model `czech-pdt-ud-2.4-190531` to be exact. It was lowercased and space characters were added to separate each token from its neighbours.

Language Model Score

Language model score measures the probability of a sentence in a language model. The KenLM toolkit is used to query the score of a sentence. This returns the logarithm base ten of the probability of the sentence, which is easily converted to probability.

The final value of the metric for a given unit is obtained by calculating the mean language model score (i.e. probability) of the sentences within that unit.

Perplexity

The perplexity metric calculates the perplexity per token (perplexity normalized by the number of tokens in the corpus the language model was trained on) for a given sentence. Using the KenLM tool, a query requesting the perplexity per token is created for each sentence.

The mean perplexity per token of each sentence within a unit is used to obtain the final value of the metric.

[8] `https://kheafield.com/code/kenlm/`

[9] `https://data.statmt.org/news-crawl/cs/news.2019.cs.shuffled.deduped.gz`

3.4 Other Metrics

Length

When comparing the lengths of documents or smaller units, we primarily look at the number of words within the unit, i.e. the number of words in a segment, an article, or the whole document. We can also examine the mean segment length for each segment of the unit. As we know, the data is segment-aligned; therefore, comparing the number of segments or the number of articles is the same in all translations.

Untranslated

The untranslated metric measures the number of words that were not translated, i.e., kept in their original form. It is only measured for the Czech texts. For each segment, its token frequency and its English counterpart's token frequency are computed. From these, the ones belonging to the list of stopwords and those carrying the PROPN, PUNCT, SYM or NUM POS tags are excluded. The list of stopwords consists of words that appear in both English and Czech but have different meanings or belong to different POS categories, e.g. *a*, *my*, *i*, etc. This list was obtained by examining the faulty outputs of a previous version of the metric without this list.

The Foreign Boolean Universal Feature could also be applied to measure this metric.

4 Experiments and results

In this chapter the ORTs, SRTs and MTs will be compared. All the experiments were performed using the MCC (Metric Calculation and Comparison) tool (Attachment A.1.2) developed for this thesis. First, the metrics at the level of the whole document are computed. When they show a significant difference in the alternate translation, these metrics are analyzed more in-depth on the segment level. There are extracted example segments where these metrics exhibit high differences to examine further how the difference manifests. On some occasions, the correlation between the differences in the metric and differences in another metric, usually observed as used instead of the first one in the examples, is explored.

When selecting the examples to include in the thesis, the ones that contribute the most to the significance of a given metric were primarily selected. These are the ones that produced the highest (or lowest) difference on segment level in units corresponding to the metric (e.g. number of occurrences for length or percentages for lexical density). The chosen segment is not always the one with the most significant difference due to its size. On some occasions, segments with less significant difference are included, capturing the same phenomenon as the ones with a more substantial difference. In some cases, only the part of the segment displaying the difference is included.

4.1 Comparing Optimal Reference Translations and Standard Reference Translations

4.1.1 Length and Type-Token Ratio

As Table 4.1 shows in all pairs consisting of an ORT (N1, N2) and an SRT (P1, P2, P3), the ORT is longer considering the number of words. On average, the ORT is 993.8 words longer (1.7 words on segment level). The MCC tool (Attachment A.1.2) was used to compute the lengths using:[1]

```
./MCC.py english -i [inp1] [inp2] -m length
```

Similarly, the TTR on forms and lemmas in percentages is computed for each pair, which is reflected in Table 4.2. It shows that N1 and N2 have similar TTRs in both versions, N2 being slightly higher. This might be caused by N1 being closer to the original English texts (Kloudová, Mraček, et al., 2023a) and English having a lower TTR than Czech (Kettunen, 2014). Interestingly, standard reference translations P1 and P3 TTR (both on forms and lemmas) higher than any of the ORTs; meanwhile, P2 has lower. Therefore, TTR cannot be considered a metric that distinguishes ORTs from SRTs.[2]

[1] Here `english` is an obligatory argument specifying the path to the file containing the source text, `inp1` and `inp2` are paths to the files containing the translations that are compared. See the user documentation for more details.

[2] No t-test was performed for this metric since investigating any lower-level unit would reflect different data.

Translation	Length
N2	18035
N1	17900
P1	17269
P2	17181
P3	16471

Compared	Difference
N1 - P1	631
N1 - P2	719
N2 - P1	766
N2 - P2	854
N1 - P3	1429
N2 - P3	1564

Table 4.1 Translation lengths and their differences (words)

To obtain the values in Table (4.2), the following command was used for each ORT-SRT pair.[3]

```
./MCC.py english -i [inp1] [inp2] -m ttr --compare difference
```

Translation	TTR on forms	TTR on lemmas
N2	34.49	23.52
N1	34.39	23.50
P2	33.87	23.01
P1	34.85	23.71
P3	37.02	25.79

Compared	Difference (forms)	Difference (lemmas)
N1 - N2	-0.10	-0.02
N1 - P1	-0.46	-0.21
N1 - P2	0.52	0.49
N1 - P3	-2.63	-2.29
N2 - P1	-0.36	-0.19
N2 - P2	0.62	0.51
N2 - P3	-2.53	-2.27

Table 4.2 Translation TTRs and their differences

From this point onward, within this section, we will use terms SRT and ORT as defined in Sections 1.1 and 1.2, respectively, that is: SRT is a concatenation of P1, P2 and P3; ORT is N1 repeated three times, so it has the same number of segments as SRT.

When comparing the number of words in each segment, the mean length difference between ORT and SRT is 1.81, and the mode difference is 0. Figure 4.1 shows the distribution of the differences. It shows that a higher number of segments are longer in the ORT. When a paired upper-tailed t-test on the segment lengths is performed, the obtained t-value = 16.67 % and the p-value < 0.001; thus, we can reject the null hypothesis. This leads to the conclusion that an ORT segment is typically longer than an SRT one, leading to a higher overall length.

To further examine this phenomenon, two example segments will be examined: one where the length difference is close to the mean and the other where the length difference is close to the maximum. The words that cause the difference are highlighted. A simpler word aligner was implemented to identify these words. It creates all possible pairs of token alignments and sorts them according to the

[3]Whether the metric is calculate on forms or lemmas is determined by the boolean parameter `lemmatized` defaulted to true. No handler was developed for changing it, so if one wants to obtain the TTR on forms, it needs to be changed manually.

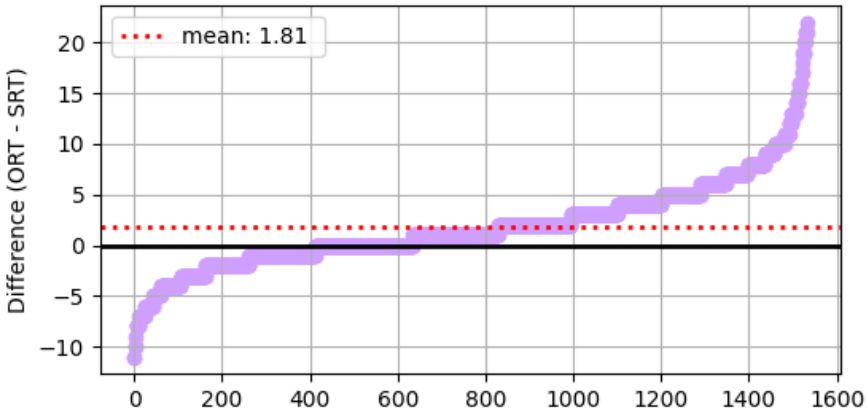


Figure 4.1 Segment length differences (ORT vs. SRT)

similarity of their forms multiplied by the similarity of their POS tags using Python's `difflib.SequenceMatcher()`. Then, it traverses the sorted pairs and hungrily selects the alignments for tokens that are not yet aligned. The unaligned tokens and those that got aligned with a token with a different POS tag are then marked as responsible for the difference by highlighting them in red and their background in yellow.

In later uses, when focusing on other metrics, all the words corresponding to the metric will be highlighted in red, and those that cause the difference will also have a yellow background.[4]

Example 1: Segment with length difference = 2 (`sent_id=61`)

English	The tool has mapped 22.5 million entities and people in Britain, and can identify in three minutes a network of connections that a staff member would have taken three hours to map out manually, he said.
ORT	Nástroj dosud ve Velké Británii zmapoval 22,5 milionu subjektů a jednotlivců a dokáže během tří minut identifikovat síť propojení, což by běžnému zaměstnanci při manuálním mapování trvalo tři hodiny, dodal.
SRT	Nástroj dosud zmapoval 22,5 milionu subjektů a lidí v Británii a dokáže během tří minut identifikovat síť propojení, což by zaměstnancům zabralo tři hodiny při manuálnímu mapování, dodal.

In Example 1, the word *Velké* ('*Great*') was added, creating a more formal tone by using the whole name of the country despite the source does not contain this word. Similarly, the word *bežnému* ('*ordinary*') was added; however, its purpose is to keep the singular form and highlight the indefiniteness used in the source (*a staff member*), meanwhile SRT uses the plural form *zaměstnancům* ('*employees*').

[4]The automatic highlighting was implemented only for length and POS tags, the examples for the rest of the metrics are highlighted manually.

Example 2: Segment with length difference = 19 (`sent_id=542`)

English	Christie Hefner, businesswoman and daughter of Playboy Founder, Hugh Hefner, says "The Girlfriend's Guide to Closing The Deal" is both highly entertaining and very helpful. Melody weaves together personal anecdotes, triumphs, and setbacks with useful lessons learned. From how to own your ideas in a meeting to when its time to "friend up," everyone can benefit from Melody's story."
ORT	Christie Hefnerová, podnikatelka a dcera zakladatele Playboye Hugha Hefnera, říká, že „ The Girlfriend ' s Guide To Closing the Deal " je velice zábavný i velmi užitečný. Melody splétá dohromady osobní příhody, vítězství i prohry a přidává užitečné poznatky, které si z toho všeho odnesla. Od toho, jak mít na schůzce své vlastní myšlenky, až po to, kdy je čas „skamarádit se". Z Melodyina příběhu mohou těžit všichni. "
SRT	Christie Hefnerová, podnikatelka a dcera zakladatele Playboy Hugha Hefnera, řekla, že kniha je velmi zábavná a velmi užitečná. Melody v ní spojuje osobní anekdoty, triumfy a neúspěchy s užitečnými ponaučeními. Z příběhu Melody může těžit každý, od toho, jak mít nápady na schůzce, až do chvíle, kdy je čas „se spřátelit".

The segment in Example 2 consists of multiple sentences, not corresponding one-to-one. This leads to an extended segment size, naturally leaving more possibilities to differ. Overall, the ORT is closer to the source — it contains the book title, *„The Girlfirend's Guide To Closing the Deal"*, and uses similar syntactic structures —, while the SRT only refers to the book as *kniha* ('*book*'). The ORT also adds the part saying *které si z toho všeho odnesla* ('*which she took away from all of it*'), which is not present in the original, probably to provide clarification. It keeps the last quotation mark also present in the original, even though it doesn't have its matching pair across the whole article (`chicago_defender.80`).

From these examples, we can observe that when the ORT is longer than the SRT, it is either because it attempts to be closer to the original or adds something to clarify or formalize.

Further on, it is interesting to take a look at the segment pairs where the opposite is observable, i.e. the ORT's segment is shorter. An example where the segment length difference is close to the minimum:

Example 3: Segment with length difference = −11 (`sent_id=320`)

English	However, Landmark Theaters has now announced that it will not allow customers to wear costumes during screenings of the **R-rated** "Joker."
ORT	Nyní ale Landmark Theaters oznámily, že zákazníkům během promítání nepřístupného filmu „Joker" neumožní mít na sobě kostým.
SRT	Nicméně kina Landmark nyní oznámila, že nedovolí zákazníkům nosit kostýmy během promítání filmu „Joker" , na který není povolen vstup osobám mladším 17 let bez doprovodu dospělé osoby .

In this Example 3, the ORT uses a single-word translation of the word *R-rated*, implementing the adjective *nepřístupný* with a broader meaning, which could be translated as '*unavailable*' or '*restricted*', however, providing an easier text processing for the reader. It maintains the property of being closer to the original. Meanwhile, SRT1 uses a clausal modifier to capture a narrower meaning and

closely explain the culture-specific word *R-rated* ('*not permitted for persons under the age of 17 without an accompanying adult*').

In the following subsections, we will abstract away from the texts' lengths and examine each metric's percentual representations.

4.1.2 Word-based Metrics

Word-based metrics are the ones that consider each word and assign it a True value if the word meets the criteria of the metric and a False value otherwise. The criteria are clearly distinguishable for each word. The concrete metrics that fall in this category are Nominalization (Section 3.1), Lexical Density (Section 3.1), Part-of-speech (Section 3.1), Dependency relations (Section 3.2) and Untranslated (Section 3.4).

To obtain Table 4.3, we computed the percentages of each metric in each document, i.e. summed the number of Trues for a given metric, divided it by the total number of words in the document, and multiplied by 100. None of the metrics take into consideration the English column; it is displayed just for the record. Then we computed the difference (ORT − SRT) between the percentages, creating the 'Comparison' column.

After that, another comparison was performed via a paired two-tailed t-test, and the resulting p-values are reported in the 'Comparison (p-val)' column. The input samples for the t-test were two arrays, one for ORT and one for SRT, containing for each segment in the document the number of Trues for the metric divided by the number of words in the segment and multiplied by 100. The t-test provides a view of how the metrics act on the segment level (low p-values suggest that it is highly improbable to observe such values of the metric for the segment if their averages are identical). The rows with low p-value (≤ 0.005) were highlighted.

Lastly, the rows where the difference was rounded to 0 or the p-value was 1 or `nan` were removed, and the table was sorted by the 'Comparison' column. This way, the metrics higher in the standard reference translation land at the top of the table, and those higher in the optimal reference translation land at the bottom.

All this can be reproduced by running the following command in the `source` directory of Attachment A.1.2.

```
./MCC.py English -i ORT SRT -m nominalization lex_density pos deprel
untranslated --compare difference_W --add_comparison ttest_perc --latex
```

The metrics highlighted in Table 4.3 are inspected more deeply. The root dependency relation is present in each segment exactly once, so the absolute number of occurrences is equal in ORT and SRT. The difference is produced by calculating the proportion of this tag to the number of all tokens, leading to a smaller percentage in the longer translation. The significant p-value of this metric reflects the difference in lengths discussed in Section 4.1.1.

The Pearson correlation coefficient (ρ) is computed for the differences between each pair of the rest of the highlighted metrics using Python's `numpy.corrcoef`. Where the ρ is positive, we can talk about correlation; where it is negative, it is about competition. Only the pairs where the the absolute value of the ρ coefficient is high enough ($|\rho| > 0.3$) will be listed in Table 4.4. This way, we find the metrics

Metric	English	ORT	SRT	Comparison	Comparison (p-val)
punct	12.099	14.693	15.351	-0.658	<0.001
PUNCT	12.105	14.715	15.373	-0.658	<0.001
ADJ	6.044	10.810	11.363	-0.553	<0.001
AUX	5.031	2.989	3.356	-0.367	<0.001
nmod	4.106	9.458	9.776	-0.318	0.007
NOUN	19.755	23.469	23.756	-0.287	0.015
amod	5.255	9.056	9.328	-0.272	<0.001
SCONJ	1.759	3.140	3.368	-0.228	<0.001
cop	1.438	1.307	1.514	-0.207	<0.001
mark	3.371	3.173	3.368	-0.195	<0.001
Nominalization	1.900	1.737	1.923	-0.185	0.013
appos	0.741	0.391	0.573	-0.182	<0.001
aux:pass	1.056	0.413	0.595	-0.182	<0.001
advcl	2.047	1.112	1.269	-0.157	0.002
root	2.782	2.855	3.011	-0.156	<0.001
NUM	2.233	2.291	2.425	-0.135	<0.001
nsubj:pass	0.882	0.453	0.579	-0.127	<0.001
ccomp	2.091	1.732	1.834	-0.102	0.244
nummod	1.552	1.196	1.275	-0.079	0.072
det	8.288	2.017	2.086	-0.069	0.135
Untranslated	0.000	0.564	0.624	-0.060	0.311
obj	4.852	4.754	4.802	-0.047	0.362
iobj	0.038	0.514	0.544	-0.030	0.737
obl:agent	0.000	0.022	0.051	-0.029	0.047
SYM	0.163	0.011	0.035	-0.024	0.004
nummod:gov	0.000	0.525	0.530	-0.005	0.753
compound	5.766	0.045	0.049	-0.004	0.594
discourse	0.044	0.000	0.004	-0.004	0.179
det:numgov	0.000	0.084	0.084	-0.001	0.251
INTJ	0.038	0.017	0.016	0.001	0.549
nsubj	7.160	6.480	6.477	0.004	0.636
PROPN	8.462	5.804	5.797	0.007	0.185
csubj	0.060	0.296	0.285	0.011	0.316
parataxis	0.757	0.246	0.230	0.016	0.547
det:nummod	0.000	0.050	0.024	0.027	0.004
orphan	0.000	0.173	0.143	0.030	0.024
flat	1.040	1.374	1.339	0.035	0.037
conj	3.289	4.497	4.462	0.035	0.794
aux	2.516	1.257	1.218	0.039	0.035
DET	8.402	4.095	4.051	0.044	0.957
PART	2.728	0.240	0.183	0.058	0.012
fixed	0.245	0.302	0.242	0.060	0.026
dep	0.000	0.402	0.342	0.061	0.785
xcomp	1.198	1.391	1.316	0.075	0.592
cc	2.793	3.341	3.254	0.087	0.126
expl:pv	0.000	1.363	1.253	0.110	0.042
expl:pass	0.000	0.296	0.183	0.113	<0.001
flat:foreign	0.000	0.458	0.344	0.114	0.029
Lexical density	49.110	55.313	55.154	0.159	0.416
CCONJ	2.782	3.503	3.331	0.172	0.005
acl	0.904	2.078	1.830	0.248	<0.001
obl:arg	0.000	1.425	1.149	0.276	<0.001
obl	5.304	5.743	5.444	0.299	0.002
VERB	11.233	10.542	10.182	0.359	<0.001
advmod:emph	0.000	0.916	0.550	0.366	<0.001
advmod	3.790	3.838	3.441	0.397	<0.001
case	10.515	10.235	9.815	0.419	<0.001
PRON	5.608	3.352	2.867	0.485	<0.001
ADP	10.019	10.335	9.841	0.494	<0.001
ADV	3.616	4.687	4.055	0.632	<0.001

Table 4.3 Results of word-based metrics (ORT vs. SRT)

Metric1	Metric2	Correlation
PUNCT	punct	0.99
ADJ	amod	0.84
AUX	cop	0.59
AUX	aux:pass	0.39
SCONJ	advcl	0.36
case	obl	0.31
ADV	advmod	0.83
ADP	case	0.97

Metric1 (ORT)	Metric2 (SRT)	Correlation
VERB	AUX	-0.47
	cop	-0.40
	aux:pass	-0.31
ADP	punct	-0.34
	PUNCT	-0.34
case	punct	-0.34
	PUNCT	-0.34

Table 4.4 Positive (left) and negative (right) correlations of significant metrics

that are either responsible for the same phenomenon ($\rho > 0.3$) or the ones that are used interchangeably in the alternative translations ($\rho < -0.3$).

The left part of Table 4.4 shows positive correlations, metrics with higher proportion in SRT at the top and with higher proportions in ORT at the bottom. The right part of Table 4.4 shows negative correlations, where Metric1 is preferred in ORT and Metric2 in SRT.

The high correlations between ADP and case are very natural since these dependency relations are, in general, assigned to these POS tags. Similarly with PUNCT and punct, ADJ and amod, and ADV and advmod.

The correlation of AUX and cop is also reasonably expected since an auxiliary verb is used to link the subject to a nonverbal root: ***Je*** *mi to líto.* ('*I* ***am*** *sorry.*') The passive auxiliary verb (aux:pass) dependency relation should be only attributed to auxiliary verbs (when they are in the passive voice), for example, *Cílem* ***bylo*** *poukázat na...* ('*The goal* ***was*** *to point out...*').

Adverbial clauses are a subtype of subordinate clauses; therefore, their use of subordinating conjunctions is quite typical. Oblique arguments or adjuncts are often accompanied by prepositions, marked by the case dependency relation. However, it is not necessary that obl comes with a preposition: *pokračovala na* ***začátku*** *21. století* ***založením*** *organizace* ('*continued at the beginning of the 21st century by founding the organization*'). The translation to English does not reflect it, but the two emphasized words in the Czech text are both obl, the first being preceded by the preposition *na*, and the second not needing any.

Now, we examine the negative correlations shown in Table 4.4.

The higher negative correlation between verbs and auxiliary verbs, as well as between verbs and copula and verbs and passive auxiliaries, can be connected with the high positive correlation with auxiliary verbs and copula and auxiliary verbs and passive auxiliaries. It suggests that in various segments, where the optimal reference translation uses a verb, the standard reference translation substitutes it with an auxiliary verb with a dependency relation of one of the two correlated ones.

Example 4: Use of AUX marked as cop in SRT instead of VERB in

ORT (`sent_id=53`)

English	It 's one of the first commercial uses of investment...
ORT	Jedná se o jedno z prvních komerčních využití investic...
SRT	Dodal: Je to jedna z prvních komerčních investic...

Example 4 shows that ORT uses a verb despite the source using an auxiliary, which SRT maintains. The ORT's version, which uses the VERB *jedná (se)* instead of the AUX *je* could be translated as '*It concerns one of the first commercial use of investments...*', creating this way a more formal tone.

Example 5: Use of auxiliary marked as aux:pass in SRT instead of verb in ORT (`sent_id=211`)

English	Williams was also fined a total of $17,000 last year...
ORT	Loni Williamsová dostala pokutu v celkové výši 17 000 dolarů...
SRT	Dodal: V loňském roce byla Williamsové uložena pokuta v celkové výši 17 000 dolarů...

The underlying phenomenon causing the difference in Example 5 is the use of active voice in the ORT — *dostala pokutu* ('*received a fine*') — instead of passive used in SRT — *byla uložena pokuta* ('*a fine was imposed*') —, even though the source uses a passive formulation. Whether it can be attributed as a differentiating element will be examined in the following section.

Example 6: Use of auxiliary marked as aux:pass in SRT instead of aux + verb in ORT (`sent_id=237`)

English	The as yet untitled film will be released on July 16, 2021.
ORT	Film, který dosud nemá název, bude mít premiéru 16. července 2021.
SRT	Dosud nepojmenovaný film bude uveden na trh 16. července 2021.

In Example 6, SRT uses a passive auxiliary, which is accompanied by an adjective, creating a passive construction. Here, the assignment of POS tags differs between English and Czech, where in both languages, the words *uvedený* and *released* are participles. Still, while English assigns it the VERB part-of-speech, Czech chooses ADJ. ORT uses an active form to express the future tense, using thus a verb. It also increases the verb count by using an adnominal clause *který dosud nemé název* ('*which does not yet have a name*') instead of the adjectival modifier *nepojmenovaný* ('*untitled*'). The use of active and passive voice will be further examined in Section 4.1.3.

ADP and case are strongly correlated, and both have fairly strong negative correlations with punctuation; therefore, they will be investigated in an attempt to find a single underlying phenomenon.

Example 7: Use of PUNCT in SRT instead of ADP in ORT (`sent_id=235`)

English	Looks like there will be a homecoming of sorts for Spider-Man, now that Disney and Sony have settled their differences.
ORT	Spider-Man se podle všeho vrací domů, neboť Disney a Sony vyřešily své spory.
SRT	Vypadá to, že se Spider-Man vrátí domů, když Disney a Sony vyřešily své rozpory.

Example 7 shows how the part *podle všeho* ('*by all signs*') in ORT is substituted by *Vypadá to,* ('*It looks like*') in SRT. The part in ORT functions as an obl, while the part in SRT forms the main clause, and the part *, že se Spider-Man vrátí domů* functions as its clausal complement. There are also other differences in the syntactic structure and POS tags, so finding some underlying structure is fairly difficult. This example, therefore, serves only to illustrate a segment that contributes to the negative correlation between ADP and punctuation and case and punctuation.

4.1.3 Features

In this section, we examine the Universal Features, as described in Section 3.1. For columns English, ORT and SRT of Table 4.5, the ratios of Values and their corresponding Feature were computed on the level of the whole document and multiplied by 100 to obtain percentages, i.e. as $100 \times \frac{\#Values\ in\ doc}{\#corresponding\ Feats\ in\ doc}$. The Comparison column was then obtained by calculating the difference between ORT and SRT. For column Comparison (p-val), a two-tailed paired t-test was performed for each Value of a Feature found in the segment-level results, which were also multiplied by 100. Some segments may not contain a given Feature at all, leading to a possible zero-division — in these cases, the number 0 was automatically used as the resulting percentage.[5] The rows with significant p-values (≤ 0.005) were highlighted in Table 4.5.

There are rows where the results in the Comparison column are high in absolute value, but the p-value is not significant (e.g. Style.Rare or Gender[psor].Fem). This can be attributed to the fact that the overall occurrence of these Values is relatively low — they appear in $< 1\,\%$ of all tokens.

To obtain Table (4.5),the following command is to be used:

```
./MCC.py English -i ORT SRT -m feats --compare difference_W --add_comparison
ttest_perc --latex
```

When examining the significant Values, we investigate the Pearson correlation coefficients between the segment-level differences between the values more frequent in SRT and those more frequent in ORT. We mention the ones with the absolute value of the coefficient being higher than 0.3, just like in Section 4.1.2.

Table 4.6 shows the significant positive correlations of Feature Values, all of which can be found at the bottom of Table 4.5, meaning their proportion is higher in ORT.

The differences between VerbForm.Fin (finite verb) and Mood.Ind (indicative mood) have a high correlation $\rho = 0.66$, meaning they often appear together in segments. This can be attributed to how finite verbs are detected in UD — if a verb's Mood feature is specified, it is automatically finite.[6]

[5]The same approach is used on any unit level if a Feature is not present in the unit.

[6]`https://universaldependencies.org/cs/feat/VerbForm.html#Fin`

Metric	English	ORT	SRT	Comparison	Comparison (p-val)
PrepCase.Pre	0.000	76.119	93.939	-17.820	<0.001
Style.Rare	0.000	22.222	32.653	-10.431	0.034
Style.Arch	0.000	33.333	36.735	-3.401	0.835
VerbForm.Part	20.020	46.160	49.423	-3.264	<0.001
Number[psor].Sing	0.000	63.303	66.168	-2.865	0.345
NumValue.1	0.000	0.000	2.500	-2.500	0.018
Gender[psor].Masc	0.000	22.353	24.632	-2.279	0.200
PronType.Int,Rel	0.000	17.773	19.848	-2.075	0.097
Style.Expr	0.000	0.000	2.041	-2.041	0.317
Tense.Past	61.640	50.791	52.472	-1.681	0.014
Voice.Pass	100.000	5.602	7.169	-1.567	<0.001
Gender.Masc	38.298	45.736	47.103	-1.367	<0.001
Case.Nom	79.505	23.625	24.866	-1.241	<0.001
NumForm.Digit	0.000	74.447	75.632	-1.184	0.041
Mood.Cnd	0.000	12.236	13.355	-1.120	0.393
Gender.Neut	36.809	14.565	15.401	-0.837	0.256
NameType.Geo	0.000	25.089	25.902	-0.813	0.836
Aspect.Perf	0.000	48.485	49.214	-0.729	0.438
Person.1	13.616	14.784	15.501	-0.717	0.685
NameType.Com	0.000	14.947	15.609	-0.663	0.713
Gender[psor].Masc,Neut	0.000	36.471	37.132	-0.662	0.157
Degree.Pos	90.917	92.351	92.947	-0.596	0.474
NameType.Pro	0.000	2.402	2.985	-0.583	<0.001
NameType.Geo,Sur	0.000	0.356	0.840	-0.484	0.021
PronType.Ind	0.000	3.555	3.980	-0.425	0.428
NumType.Sets	0.000	0.000	0.423	-0.423	0.023
Number.Sing	79.384	68.691	69.062	-0.371	0.596
PronType.Prs	31.795	45.875	46.185	-0.310	0.400
NameType.Nat	0.000	1.068	1.368	-0.301	0.079
Case.Acc	20.495	22.446	22.737	-0.291	0.392
Number.Plur,Sing	0.000	2.950	3.139	-0.189	0.018
Case.Gen	0.000	23.313	23.495	-0.183	0.391
Gender.Fem,Neut	0.000	3.486	3.668	-0.182	0.012
AdpType.Voc	0.000	8.811	8.980	-0.169	0.607
Polarity.Pos	0.000	97.985	98.124	-0.139	0.067
NumType.Mult,Sets	0.000	0.000	0.071	-0.071	0.317
Animacy.Anim	0.000	39.671	39.739	-0.068	0.316
NameType.Geo,Giv	0.000	0.000	0.062	-0.062	0.158
VerbForm.Conv	0.000	0.000	0.040	-0.040	0.090
NameType.Pro,Sur	0.000	0.000	0.031	-0.031	0.317
Number.Dual	0.000	0.000	0.014	-0.014	0.109
NameType.Giv,Pro	0.000	0.178	0.187	-0.009	0.166
Case.Ins	0.000	7.788	7.794	-0.006	0.837
NumType.Card	94.037	90.756	90.762	-0.005	0.390
NameType.Giv,Sur	0.000	0.089	0.093	-0.004	0.433
Case.Voc	0.000	0.010	0.007	0.003	0.566
PronType.Neg	0.000	1.543	1.521	0.022	0.997
AdpType.Comprep	0.000	0.162	0.140	0.022	0.374
Mood.Imp	0.863	0.736	0.698	0.038	0.027
Gender.Fem,Masc	0.000	1.399	1.355	0.044	0.972
NameType.Com,Geo	0.000	0.089	0.031	0.058	0.153
NumType.Frac	0.000	0.630	0.564	0.066	0.760
Animacy.Inan	0.000	60.329	60.261	0.068	0.863
PronType.Emp	0.000	0.537	0.456	0.080	0.300
NumType.Ord	5.963	7.773	7.687	0.086	0.853
NameType.Oth	0.000	1.779	1.679	0.100	0.604
PronType.Tot	0.000	3.555	3.447	0.107	0.339
Polarity.Neg	0.000	2.015	1.876	0.139	0.067
AdpType.Prep	0.000	91.027	90.880	0.147	0.097
Degree.Cmp	4.931	4.335	4.129	0.206	0.538
NumForm.Roman	0.000	0.246	0.000	0.246	0.083
Tense.Fut	0.000	2.918	2.659	0.259	0.043
Person.2	2.723	3.184	2.863	0.321	0.051
NumType.Mult	0.000	0.840	0.494	0.347	0.072
Degree.Sup	4.152	3.315	2.924	0.391	0.596
Person.3	83.661	82.032	81.635	0.396	0.037
Number.Plur	20.616	28.359	27.785	0.574	0.175
Case.Dat	0.000	6.145	5.496	0.649	<0.001
NameType.Giv	0.000	14.680	13.993	0.687	0.558
Aspect.Imp	0.000	51.515	50.786	0.729	0.156
Gender.Masc,Neut	0.000	2.527	1.791	0.736	<0.001
PronType.Dem	5.253	22.669	21.825	0.844	0.292
NumForm.Word	0.000	25.307	24.368	0.939	0.086
Case.Loc	0.000	16.673	15.605	1.068	<0.001
Mood.Ind	99.137	87.029	85.947	1.082	0.002
VerbForm.Inf	17.409	11.887	10.643	1.244	0.002
Tense.Pres	38.360	46.291	44.869	1.422	0.119
Voice.Act	0.000	94.398	92.831	1.567	<0.001
Gender.Fem	24.894	32.288	30.681	1.606	<0.001
PronType.Rel	5.531	4.494	2.738	1.756	<0.001
VerbForm.Fin	52.226	41.953	39.894	2.059	0.001
NameType.Sur	0.000	39.324	37.220	2.104	0.009
NumValue.1,2,3	0.000	100.000	97.500	2.500	0.167
Number[psor].Plur	0.000	36.697	33.832	2.865	0.227
Gender[psor].Fem	0.000	41.176	38.235	2.941	0.395
Style.Coll	0.000	44.444	28.571	15.873	0.050
PrepCase.Npr	0.000	23.881	6.061	17.820	<0.001

Table 4.5 Features (ORT vs. SRT)

Metric1	Metric2	Correlation
VerbForm.Fin	Mood.Ind	0.66
PrepCase.Npr	PronType.Rel	0.40

Table 4.6 Significant positive correlations Feature Values

Example 8: PrepCase.Npr and PronType.Rel (`sent_id=235`)

English	...co-defendant, Serge Nkorina, who 's being extradited from Tenerife, Spain.
ORT	...spoluobžalovaný Serge Nkorina, jehož v těchto dnech k trestnímu stíhání vydává španělské Tenerife.
SRT	...spoluobžalovaný muž Serge Nkorin, kterého právě vydává k trestnímu stíhání španělské Tenerife.

PrepCase.Npr (non-prepositional case) is only indicated for pronouns. The only lemma for which it appears together with PronType.Rel is *jenž* ('*which*'). Upon looking at the number of occurrences of this lemma in the translations, it is observed that the number of occurrences in ORT is 48, while in SRT only 6. Example 8 shows the use of the determiner *kterého* ('*who*') instead of the word *jehož* ('*who*'). The use of the pronoun *jenž* is associated with a more formal, more official language, as opposed to the use of *který*, thus contributing to a higher formality in ORT.

Metric1	Metric2	Correlation
Voice.Act	Voice.Pass	-0.85
VerbForm.Fin	VerbForm.Part	-0.82
Mood.Ind	VerbForm.Part	-0.63
Gender.Fem	Gender.Masc	-0.60

Table 4.7 Significant negative correlations Feature Values

Table 4.7 contains the Feature Values with a significant negative correlation. The metric in column Metric1 has a higher proportion in ORT and the one in column Metric2 in SRT.

Example 9: Voice.Act in ORT instead of Voice.Pass in SRT (`sent_id=235`)

English	Three Scottish students named among Europe's best
ORT	Tři skotští studenti patří mezi nejlepší v Evropě
SRT	Tři skotští studenti byli zařazeni mezi nejlepší v Evropě

The high negative correlation between active and passive voice suggests that Voice.Act is used in ORT where Voice.Pass is used in SRT on many occasions. This was first observed in Section 4.1.2 in Example 5 when examining the negative correlation between VERB and aux:pass. This also happens in Example 9 where the ORT's *patří* ('*belong*') is substituted by the passive construction *byli zřazeny* ('*were ranked*'). This happens again despite the fact that the English uses a passive construction.

The high negative correlation of differences in VerbForm.Fin and VerbForm.Part (participle), as well as of Mood.Ind and VerbForm.Part, can be again attributed to a higher use of the active voice in ORT than in SRT. This is due to how the passive form is constructed: it usually contains an auxiliary verb (with VerbForm=Part and Voice=Act) followed by an adjective which has the form of a participle (it has Voice=Pass and VerbForm=Part). In the active construction, a single verb is used, with Features Voice=Act, and according to these negative correlations, often also with VerbForm=Fin and Mood=Ind (the high correlation of these was already discussed).

Example 10: Gender.Fem in ORT instead of Gender.Masc in SRT (`sent_id=99`)

English	He has struggled with vocal issues and apparently is under strict instructions from his surgeon.
ORT	V poslední době má problémy s hlasivkami a zjevně se musí řídit přísnými pokyny svého lékaře.
SRT	Má problémy s hlasem a zjevně se musí řídit přísnými pokyny svého lékaře.

Example 11: Gender.Fem in ORT instead of Gender.Masc in SRT (`sent_id=296`)

English	The IAAF, the global governing body of athletics, said
ORT	Mezinárodní asociace atletických federací (IAAF) uvedla,...
SRT	IAAF, celosvětový dozorčí orgán pro atletiku, řekl,...

Two examples are included to show how feminine gender is used in ORT instead of masculine in SRT. Example 10 shows that what contributes to the difference is adding the part *v poslední době* ('*lately*'), not contained in the original, where both *poslední* and *době* appear with Gender=Fem. Instead of the masculine noun *hlas* ('*voice*'), the feminine *hlasivky* ('*vocal chords*') is used. Example 11 captures how a change of a single noun from one gender to another, like here the use of *asociace* ('*association*') instead of *orgán* ('*body*') requires all of its modifiers to change their gender, due the fact that Czech requires the noun and its adjectival modifiers to match in gender, number and case. When this noun is a subject, the gender of the predicate also needs to change, as seen here with *uvedla* ('*[she] stated*') being feminine and *řekl* ('*[he] said*') being masculine. Neither of these examples hints at a reason behind selecting a noun with Gender=Fem in ORT instead of a noun with Gender=Masc.

4.1.4 Syntactic Metrics

In this section, the syntactic metrics described in Section 3.2 are measured, except for the dependency relations, since these are discussed with the other word-based metrics. Syntactic complexity, tree depth and tree width, however, are segment-based metrics, meaning the smallest unit they can be evaluated for are segments. To obtain their value for the whole unit, they are averaged across the segments of the unit, i.e. the values for each segment are summed and divided by the total number of segments in the unit. This number is in the Comparison column of Table 4.8. To compute the values in the Comparison (p-val) column, an upper-tailed paired t-test is performed, and the resulting p-value is indicated there.

To obtain Table (4.8),the following command is to be used:

```
./MCC.py English -i ORT SRT -m tree_depth tree_width syntactic_comp --compare
difference_W --add_comparison ttest_abs --latex
```

Metric	English	ORT	SRT	Comparison	Comparison (p-val)
Syntactic complexity	3.620	3.329	3.183	0.145	<0.001
Tree depth	7.366	8.286	8.112	0.174	<0.001
Tree width	10.728	9.583	9.142	0.442	<0.001

Table 4.8 Syntactic metrics (ORT vs. SRT)

As Table 4.8 shows, all of the metrics it contains are higher for ORT than for SRT. The obtained p-values, all < 0.001, suggest that the metrics also differ on the segment level, being higher for ORT. To examine this further, we look at the distribution of the differences between the segments.

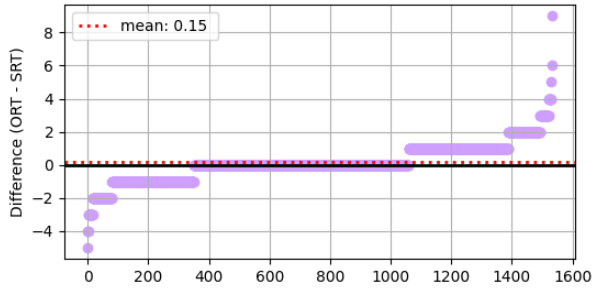


Figure 4.2 Segment syntactic complexity differences (ORT vs. SRT)

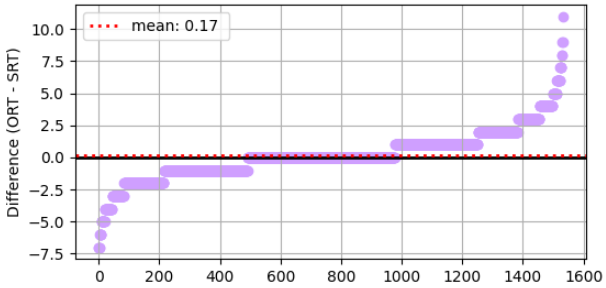


Figure 4.3 Segment tree depth differences (ORT vs. SRT)

For all of the metrics, more than half of the segments have a difference of zero or higher, and all have a mode of 0 but with various sizes of representation. Tree width has the highest mean difference, 0.44, but it also has the broadest range of values. Both syntactic complexity and tree depth have lower ranges and lower means, the first one ranging from -5 to 9 with an average difference of 0.15 and the second one with a range from -7 to 11 with its mean being 0.17.

Since syntactic complexity is computed from the clause heads, as defined in Section 3.2, they are the ones that contribute to its higher value. Table 4.9 shows

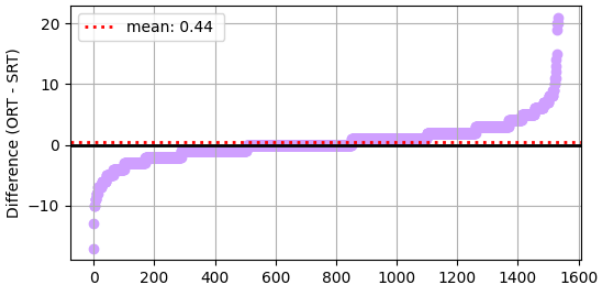


Figure 4.4 Segment tree width differences (ORT vs. SRT)

Metric	English	ORT	SRT	Comparison	Correlation[a] (ρ))
advcl	2.047	1.112	1.269	-0.157	0.29
root	2.782	2.855	3.011	-0.156	-0.18
ccomp	2.091	1.732	1.834	-0.102	0.26
csubj	0.060	0.296	0.285	0.011	0.14
xcomp	1.198	1.391	1.316	0.075	0.38
acl	0.904	2.078	1.830	0.248	0.41

Note: [a] between the clause head proportion difference and syntactic complexity difference

Table 4.9 Dependency relations contributing to syntactic complexity

their proportion in the whole token count and their differences, as well as how these differences on the segment level correlate with the differences in syntactic complexity on the segment level. The root dependency relation contributes to the syntactic complexity of each segment exactly by 1 since each segment has precisely one root. If it were not included, the average syntactic complexities would be smaller by 1.

The strongest obtained correlation is with adnominal clauses (acl), with $\rho = 0.41$. These are encountered, as Table 4.9 shows, with a higher proportion in ORT (and the difference with SRT is significant, seen in Table 4.4). They also present the highest difference between the proportions on the level of whole documents. Therefore, it seems to have the most significant contribution to a higher syntactic complexity ORT.

Example 12: Higher syntactic complexity and proportion of acl in ORT (`sent_id=838`)

English	Oregon is a powder keg of militancy right now, and its fuse is now burning bright and hot.
ORT	Oregon v tuto chvíli představuje sud se střelným prachem, jehož knot je pořádně rozžhavený.
SRT	Oregon se teď stal výbušným soudkem střelného prachu a jeho rozbuška teď silně hoří.

Example 12 shows, that instead of the adnominal clause *jehož knot je pořádne rozžhavený* (‘*whose wick is red-hot*’) used in ORT, SRT uses a coordinated clause *a jeho rozbuška teď silně hoří* (‘*and its fuse is now burning strong*’). Computing the proportion of coordinated clauses is more challenging; it does not depend only on the conj dependency relation since this relation is also assigned to the coordination of other constituents (nouns, adjectives, etc.). The implementation

of this metric and the comparison of the proportion of subordinate and coordinate clauses might lead to interesting observations of the relations between these, too. However, it will not be in the scope of this thesis.

All the dependency tree visualizations used in the following examples were obtained from UD's CoNLL-U Viewer,[7] where the file with the parsed segment was provided.

When investigating the tree depths, the segments with `sent_id=31`, `sent_id = 38` and others were examined, where the difference was high. Since their dependency trees are fairly large, they won't be included in this thesis. Refer to files `31ort_td.png`, `31srt_td.png`, `38ort_td.png` and `38srt_td.png` in Attachment A.1.1 for the tree for ORT and SRT of these segments, respectively.

After close observations, we notice that a parser error in ORT probably causes a high difference in both cases. In the segment with `sent_id=31` a wrong identification of the parent of the adverbial modifier *obviněných* is to blame. In segment `sent_id = 38`, the parser struggles with parsing the multiple sentences present — it doesn't attach the head of one sentence to the root, but to the head of a subordinate clause, thus contributing to the tree depth.

Therefore, we examine segments with lower differences, where we hope for a less distorted tree. The segment with `sent_id=1165` has a difference of 4. File `1165srt_td.png` in Attachment A.1.1 shows that there is an error in assigning the head of a subordinate clause in SRT, where it is connected to a parent at a higher level than in ORT. Other than this, what contributes to the difference in depth by 1 is the use of a nominal modifier with a preposition in ORT, *náklady na kapitál* ('*cost of capital*'), instead of the adjectival modifier *kapitálové příjmy* ('*capital revenue*') in SRT.[8] This phenomenon, however, contributes to the depth only if it is located at a low level.

To explore tree width in more detail, we look at a segment where the difference between ORT and SRT is higher but not extreme. A shorter segment is selected to make the difference easier to find; therefore, the treewidth difference (2) is low. In the segment with `sent_id=1619` ORT uses a clausal complement to express the object, while SRT uses a noun — the tree visualizations can be found in files `1619srt_tw.png` and `1619ort_tw.png` of Attachment A.1.1. This causes the level under the object to be widened by the punctuation, subordinate conjunction, noun functioning as subject, and also by an adverb and a noun. This level then became wide enough to cause the difference. What plays a role is the rest of the width of this level and also the structure of the subordinate clause. This segment is only used to show an example of ORT having a larger tree width than SRT; the various factors behind it make it hard to find a generalization.

The manifestations of the differences in the syntactic metrics are fairly complex to observe; therefore, it is hard to pinpoint any concrete phenomena and the reason behind them in either case.

[7] `https://universaldependencies.org/conllu_viewer.html`

[8] The difference between the nouns that are modified (*kapitál* and *příjmy*) is caused by the different word order — *kapitálové náklady* as well as *příjmy z kapitálu* would both be valid alternatives.

4.1.5 Language Model Score and Perplexity

In this section, the Language model-based metrics defined in Section 3.3 will be discussed.

First, the LM scores for the translations are computed (in percentages), and their difference is calculated. Then, a two-tailed paired t-test on the input arrays containing the LM score for each segment is performed. Since the segment probabilities are really low, as well as their differences, and the t-test doesn't provide enough significance in either of the 2-, 3- and 4-gram models, this table won't be included here.

Similarly, the average perplexities for the whole documents are computed, and their difference is obtained for each language model. As Table 4.10 shows, this difference is higher for the ORT for all three models. After performing an upper-tailed paired t-test on the segment-level perplexities, the p-values are obtained. None of these is significant. Further, the t-test is performed with the inputs containing the average perplexity for each article. This doesn't provide a significant p-value either.

To obtain Table 4.10, the following command is to be used for each language model:

```
./MCC.py English -i ORT SRT -m perplexity --model [path_to_model] --compare
difference_W --add_comparison ttest_abs --latex
```

Model	English	ORT	SRT	Comparison	Comparison (p-val) (on segments)	Comparison (p-val) (on articles)
2-gram LM	—	2623.140	2460.311	162.829	0.271	0.335
4-gram LM	—	2157.196	2070.748	86.448	0.369	0.406
3-gram LM	—	2191.060	2105.649	85.411	0.366	0.403

Table 4.10 ORT vs. SRT (Perplexity)

The perplexity, as well as the LM score, is distorted since the models are trained on a sentence-segmented corpus, while when querying the model, a whole segment is presented, on occasions consisting of multiple segments.

4.2 Comparing Optimal Reference Translations and Machine Translations

This section will compare the N1 optimal reference translation with machine translations MT1 and MT2. Only the comparison of their lengths is included in this thesis. The comparison of other metrics for these translations can be performed using the MCC tool (Attachment A.1.2) in the same way as for ORT and SRT but is left as a topic of future work.

4.2.1 Length

In both cases, as Table 4.11 shows, the ORT is longer than the MT. To obtain these lengths, the following command was used, with N1 as `inp1` and MT1 and MT2 as `inp2`:

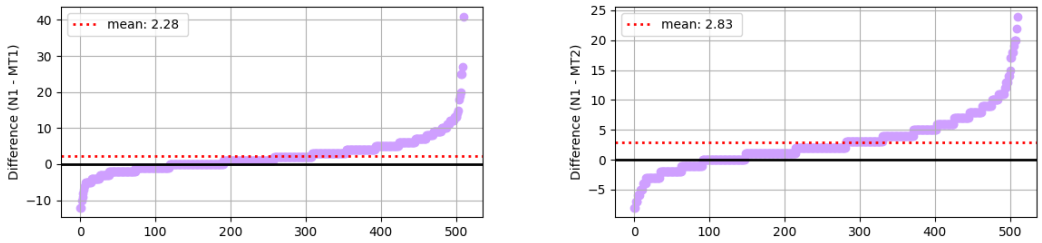
```
./MCC.py english -i [inp1] [inp2] -m length
```

Translation	Length
N1	17900
MT1	16737
MT2	16455

Compared	Difference
N1 - MT1	1163
N1 - MT2	1445

Table 4.11 Translation lengths and their differences (words)

Focusing on N1 and MT1, when examining the segment length differences, the mean of their distribution is 2.28, and the mode is 0. Upon performing a paired upper-tailed t-test on the segment lengths of the two translations, we obtain a t-value of 10.78 % and a p-value < 0.001. Thus, we can concur that N1 is typically longer than MT1 when it comes to the number of words.



(a) N1 vs. MT1 **(b)** N1 vs. MT2

Figure 4.5 Segment length differences

When examining the distribution in Figure 4.5a, there is an outlier, where the segment length difference is 40. For further examination, we look at the exact sounding of the texts in Example 1. We see that MT1 has significantly simplified the translation, leaving out entire sentences. The sentence it has translated, however, reflects the corresponding sentence in the source more exactly. N1 uses two alternatives for the translation of the word *dependent* — *rodinný příslušník* (‘*family member*’) and *závislý* (‘*dependent*’). This is not the first time machine translations have left out whole sentences; it can also be seen in the everyday use of commercial translators.

Example 1: Segment with length difference = 40 ((`sent_id=61`))

English	The vast majority - nearly 70 percent - were female spouses under the age of 40, while 70 percent of the dependent suicides were males. About half of the dependents who died by suicide were at least 18 years old and, for those younger than that, most of the deaths were youth between 15 and 17.
N1	Velkou většinu - téměř 70 procent - tvořily manželky mladší 40 let, zatímco 70 procent dalších rodinných příslušníků , kteří spáchali sebevraždu , bylo mužského pohlaví . Přibližně polovině závislých osob , které dobrovolně odešly ze života , bylo alespoň 18 let . Co se týče zesnulých mladších 18 let , bylo obětem sebevraždy nejčastěji mezi 15 a 17 lety .
MT1	Drtivá většina – téměř 70 procent – byly manželky mladší 40 let, zatímco 70 procent závislých sebevražd byli muži.

When comparing N1 and MT2 in more depth, Figure 4.5b shows that the mean difference between segment lengths is 2.83. The mode, computed separately, is 1. In this case, a paired upper-tailed t-test is also performed. The resulting t-value is 14.4 % and the p-value is less than 0.001. This leads to the conclusion that N1 typically comprises longer segments, resulting in a larger overall length.

Example 2: Segment with length difference = 18 (`sent_id=167`)

English	Kit Malthouse, minister for crime, policing and fire, said he welcomed the chief constable's "swift action to address these failings and will be monitoring the position closely," adding: "Transparency and accountability are vital for public trust in policing.
N1	Ministr pro zločinnost, policejní a hasičské sbory Kit Malthouse řekl, že vítá ředitelovy „rychlé kroky vedoucí k řešení tohoto selhání " a že bude situaci " bedlivě sledovat“ , a dodal : „ Transparentnost a odpovědnost jsou zásadní pro důvěru veřejnosti v činnost policie . “
MT1	Kit Malthouse , ministr pro zločin, policii a požár , uvedl, že přivítal „rychlé kroky vedoucího konstituce k řešení těchto selhání a bude tuto situaci pečlivě sledovat“.

To further examine the segment length difference, Example 2 is selected from the data, with N1 having 18 more tokens than MT2. As in the previous translation pair, here we can also attribute the difference to the fact that the machine translation omitted an entire sentence.

Other segments with high differences in length were examined, too, for both translation pairs and the same reason was observed there. Therefore, what contributes most to differences between the segment, and thus the overall lengths, is that on some occasions MTs don't include a translation of one or more sentences without replacing them with any form of shorter expressions of their content.

Conclusion

In this thesis, the analysis of alternative translations (ORTs, SRTs and MTs) was executed, with the main focus on the first two.

When comparing the lengths, it was observed that each of the ORTs were longer than the SRTs as well as as the MTs, on document level, and also significantly on sentence level. With the ORT-SRT pair, this was contributed to the ORT being closer to the source and providing additional clarification or formalization. With the ORT-MT pair, the largest differences were produced by the machine translation occasionally omitting entire sentences.

The rest of the metrics were only computed for the ORT-SRT pair.

Upon computing the word-based metrics, many gave significant results after they were compared. To select a subset to be used for further manual analysis on example segments, the correlations of their segment-level differences were computed. Pairs with the absolute value of Pearson correlation higher than 0.3 were selected. Most importantly, the negative correlation of the tokens with the POS tag VERB and AUX led to observing a more formal tone in ORT when a VERB was used instead of an AUX with the cop dependency relation present in the SRT, and the use of active sentence formulation in ORT instead of a passive one in SRT.

The analysis of features applied a similar approach to the word-based metrics. The findings here contributed to the frequent use of active voice in ORT in segments where SRT uses passive voice. It was also observed that ORT frequently uses the feminine gender in segments, whereas SRT uses masculine in the corresponding segments. What increases the difference in the proportion of words with a given gender is that the Czech grammar requires agreement in gender, number and case between a noun and selected dependency relations. The reason behind ORT preferring nouns of masculine gender was not detected.

When analyzing the syntactic metrics, all three gave results that supported their higher value in ORT. The higher proportion of use of the subordinate clause acl (adnominal clause) contributed the most to the higher syntactic complexity. No concrete phenomena were identified that would be behind a higher depth or width of a dependency tree.

Lastly, the computation of the LM score and perplexity did not yield any significant results. However, the language models used were all trained on sentence-segmented data, while when computing these metrics the whole segment was used.

Future Work

Future work may consist of a more detailed comparison of optimal reference translations and MTs or a more in-depth analysis of each of the metrics. New metrics can also be developed, for example, ones that take the source text into consideration, ones that detect coordinate clauses, or combine the metrics explored here. If more significant differences were found a classifier could be developed for determining the type of a translation.

Bibliography

Baker, Mona, Gill Francis, and Elena Tognini-Bonelli (1993). *Text and technology: in honour of John Sinclair*. John Benjamins Publishing.

Barrault, Loïc et al. (Nov. 2020). "Findings of the 2020 Conference on Machine Translation (WMT20)". In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 1–55. URL: https://www.aclweb.org/anthology/2020.wmt-1.1.

Bestgen, Yves (2021). "Using CollGram to compare formulaic language in human and machine translation". In: *Proceedings of the translation and interpreting technology online conference*, pp. 174–180. URL: https://aclanthology.org/2021.triton-1.20.

Comrie, Bernard and Sandra Thompson (2007). "Lexical nominalization". In: *Language Typology and Syntactic Description*, pp. 334–381. DOI: 10.1017/CBO9780511618437.006.

Evert, Stefan and Stella Neumann (2017). "The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German". In: *Empirical translation studies: New methodological and theoretical traditions* 300, p. 47. URL: https://www.degruyter.com/document/doi/10.1515/9783110459586-003.

Johansson, Victoria (2008). "Lexical diversity and lexical density in speech and writing: A developmental perspective". In: *Working papers/Lund University, Department of Linguistics and Phonetics* 53, pp. 61–79. URL: https://journals.lub.lu.se/LWPL/article/download/2273/1848.

Kettunen, Kimmo (2014). "Can Type-Token Ratio be Used to Show Morphological Complexity of Languages?" In: *Journal of Quantitative Linguistics* 21.3, pp. 223–245. DOI: 10.1080/09296174.2014.911506. URL: https://doi.org/10.1080/09296174.2014.911506.

Kloudová, Věra, Ondřej Bojar, and Martin Popel (2021). "Detecting Post-edited References and Their Effect on Human Evaluation". In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Stroudsburg, USA: Association for Computational Linguistics, pp. 114–119. ISBN: 978-1-954085-10-7. URL: https://aclanthology.org/2021.humeval-1.13.pdf.

Kloudová, Věra, David Mraček, et al. (2023a). "Možnosti a meze tvorby tzv. optimálních referenčních překladů: po stopách „překladatelštiny" v profesionálních překladech zpravodajských textů". cze. In: *Slovo a slovesnost* 84.2, pp. 122–156. ISSN: 0037-7031. URL: https://doi.org/10.58756/s2228425.

— (2023b). *Optimal reference translation of English-Czech WMT2020*. URL: http://hdl.handle.net/11234/1-5141.

Kralík, Petr (2017). *NOMINALIZACE*. https://www.czechency.org/slovnik/NOMINALIZACE. Accessed: 2024-7-13.

Kyle, Kristopher (2020). "Measuring Lexical Richness". eng. In: *The Routledge Handbook of Vocabulary Studies*. 1st ed. Routledge, pp. 454–476. URL: https://doi.org/10.4324/9780429291586-29.

Liu, Kanglong and Muhammad Afzaal (2021). "Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification". In: *Plos one*

16.6, e0253454. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0253454.

Ney, H., U. Essen, and R. Kneser (1995). "On the estimation of 'small' probabilities by leaving-one-out". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17.12, pp. 1202–1212. DOI: 10.1109/34.476512. URL: https://doi.org/10.1109/34.476512.

Popel, Martin (Nov. 2020). "CUNI English-Czech and English-Polish Systems in WMT20: Robust Document-Level Training". In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 269–273. URL: https://aclanthology.org/2020.wmt-1.28.pdf.

Straka, Milan and Jana Straková (Aug. 2017). "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe". In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, pp. 88–99. URL: http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Vidra, Jonáš et al. (2021). *DeriNet 2.1*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. URL: http://hdl.handle.net/11234/1-3765.

Zouhar, Vilém et al. (2024). "Evaluating optimal reference translations". In: *Natural Language Processing*, pp. 1–24. DOI: 10.1017/nlp.2024.3. URL: https://doi.org/10.1017/nlp.2024.3.

List of Figures

List of Tables

List of Abbreviations

LM language model. 32, 35

MT machine translation. 6, 8, 9, 17, 32, 34, 35

ORT optimal reference translation. 6, 7, 9, 17, 21, 23, 32, 35

POS part-of-speech. 10, 11, 12, 16, 19, 23, 24, 25, 35

SRT standard reference translation. 6, 7, 8, 9, 17, 21, 23, 35

TTR type-token ratio. 10, 14, 17

UD Universal Dependencies. 6, 10, 31

A Attachments

A.1 attachments.zip

All of the attachments are included in the attachments.zip compressed file.

A.1.1 img

The attachments.zip compressed file contains the image folder, which contains images that were too large to be included in the thesis (e.g., dependency trees).

A.1.2 tool

The tool folder in the attachments.zip compressed file contains the tool's source code in the source folder and the user and developer documentation. It also contains a Makefile, a requirements.txt file, a compressed file consisting of part of the data used in the experiments and the scripts folder, containing scripts used by the Makefile.